Synthese (2024) 204:89
https://doi.org/10.1007/s11229-024-04657-1

ORIGINAL RESEARCH

A non-causalist account of the explanatory autonomy in the psychological sciences

José Díez[1] · **David Pineda**[2]

Received: 1 November 2023 / Accepted: 28 May 2024 / Published online: 27 August 2024

Abstract
It has been often claimed that physicalism challenges the explanatory autonomy of psychological sciences. Most who advocate for such explanatory autonomy and do not want to renounce to physicalism, presuppose a causalist account of explanatoriness and try to demonstrate that, adequately construed, (causal) psychological explanations are compatible with (some sufficient version of) physicalism. In Sect. 1 we summarize the different theses and assumptions involved in the seeming conflict between explanatory autonomy and physicalism. In Sect. 2 we review the main attempts to make them compatible assuming a causalist account of explanation and argue that none succeeds. In Sect. 3 we introduce a recent, non-causalist account of scientific explanation as ampliative, specialized embedding (ASE) that has been successfully applied to other fields. In Sect. 4 we apply ASE to elucidate two paradigmatic cognitive explanations of psychological phenomena: déjà vu and action production. We conclude that ASE elucidates well the autonomy of the cognitive explanations of these phenomena independently of what finally happens with the causal exclusion problem and that it may be generalized to other psychological explanations.

Keywords Explanatory autonomy · Psychological explanations · Causal exclusion problem · Explanatory embedding

Since Fodor (1974) and the general acceptance of physicalism, a sort of ontoepistemic problem has taken the philosophical community on its grip. It is a metaphysical problem with deep and (for many) unacceptable epistemic consequences. On the one hand, we just take for granted the epistemic legitimacy of explanations in what Fodor called special sciences, including, of course, the explanations in the psychological sciences. On the other hand, a commitment to physicalism which is (again, according

✉ José Díez
jose.diez@ub.edu

1 LOGOS, BIAP, Universitat de Barcelona, Barcelona, Spain

2 LOGOS, BIAP, Universitat de Girona, Girona, Spain

to many) also inalienable seems to just compromise such epistemic legitimacy. In the last four decades or so, a significant number of metaphysicians, epistemologists and philosophers of science, have been struggling with this problem. None of the variegated solutions proposed to the problem, however, has been regarded by the majority of philosophers as convincing enough.

After so many years of inconclusive discussions, we think it is time to confront the problem in a different way, by looking for an entirely different solution to it. If this new approach succeeds, we can achieve the long yearned result of reconciling the legitimacy of cognitive science explanations with a metaphysical commitment to physicalism.

We start in Sect. 1 by laying out the problem in its bare bones. Although the problem can be raised for all special sciences, we present it with a focus on psychological sciences, emphasizing, whenever relevant, some issues which are specific of this case. The solution we will propose is as general as the problem, thus if our proposal for psychological sciences works one should make sure that explanations of other special sciences also meet the conditions for explanatoriness we will specify below. In the second section, we will offer a survey of the main causalist solutions to the problem and their shortcomings. In the third section, we will present our account for scientific explanations as ampliative, specialized embeddings (ASE). In the fourth section, we will show how two paradigmatic, cognitive explanations of two different psychological phenomena meet the demands posed by ASE, thus giving plausibility to our claim that the explanatory import of cognitive explanations comes from their ASE features regardless of what finally happens with the causal problem. We conclude that our account escapes the alleged explanatory exclusion problem by denying the crucial causalist premise necessary to generate the problem while preserving the explanatory value of cognitive explanations.

1 Explanation in cognitive sciences: the problem

We begin simply by acknowledging the epistemic legitimacy of explanations in psychological sciences. Assuming that explaining is a crucial scientific activity, this simply amounts to an acknowledgment of the activity in psychological sciences as a genuine scientific activity. Such acknowledgment, however, also tends to come along with an acknowledgment of psychological explanations as being autonomous with respect to physical explanations, meaning that they cannot be reduced to physical explanations, nor dispensed with without epistemic loss.[1] We can encapsulate both ideas in one single first thesis:

[1] By "epistemic legitimacy" we simply mean that cognitive psychology is considered a bona fide explanatory scientific practice, that it provides prima facie (that is, setting aside the problem of explanatory exclusion) bona fide explanations of human behavior. By "epistemic loss" we mean that dispensing of cognitive explanations would in principle imply a loss in the understanding of human behavior, that is cognitive explanations prima facie provide a specific genuine understanding of human behavior not provided by neuroscientific explanations.

T1 Explanatory autonomy (EA): (At least some) Explanations in the psychological sciences are bona fide explanations and they cannot be reduced to physical explanations, in the sense of being dispensed without epistemic loss.

The next thesis simply captures the widespread idea that explanations in the psychological sciences, and special sciences in general, are causal explanations:

T2 Explanatory causalism (EC): Bona fide explanations in the psychological sciences are causal explanations, that is to say, explananda in the psychological sciences are explained by mentioning in their explanans mental events that feature in their causal ancestry.

We finally need to lay out our general commitment to a physicalist metaphysics. The old days of physical reductivism are of course long gone, and physicalists nowadays claim to be non-type-reductivists. The idea is that everything empirical is metaphysically dependent, though not necessarily identical to, something physical. At the same time, this relation of dependence is metaphysically robust, in the sense that physical things metaphysically determine all the rest of empirical things. We can summarize this idea in the following third thesis:

T3 Non-reductive physicalism (NRP): All mental events and processes are metaphysically determined by physical events and processes.[2]

Of course, this relation of metaphysical dependence and determination needs to be substantiated. As it is well-known, supervenience, realization and grounding have been the traditional candidates. We do not want to dwell into this, however, at this preliminary stage. We defer the discussion of possible options in this regard to the next section, where we will review the different solutions proposed to the general problem.

Now, the problem as it stands does not look like a real problem. It would seem that one can, for instance, hold that bona fide psychological explanations are causal explanations where the mental events featured in the explanans and causing the explanandum are determined by physical events. What's the problem with this? This actually amounts to Fodor's position. But the three theses alone do not add up to a solid ontoepistemic position. In fact, we start with a purely epistemic thesis, T1, and end with a purely metaphysical thesis, T3. In between, we have a thesis T2, that mingles epistemic with metaphysical issues. And here it lurks the problem. In order to have a coherent epistemic and metaphysical position we need to make some auxiliary assumptions that help us to bridge the gap between the epistemic and the metaphysical domains, and it is here, as we will see, that the problem shines in all its glory.

[2] We will not dwell in this paper with issues having to do with a correct formulation of physicalism. So we will not discuss how best to define the physical, whether physicalism should be cashed out only in terms of microphysical entities rather than in terms which include also macrophysical things, nor will we discuss the delicate issue about how best to understand the modality of the physicalist thesis (although in the next section, when discussing alleged solutions to the general problem, we will consider different ways of understanding how the physical is supposed to metaphysically determine the mental). Of course, dealing with these issues goes beyond the scope of this paper. We simply assume that physicalists will be content with assuming T3, which suffices for raising our problem, and we will set aside here such qualifications and technicalities. We hope that nothing relevant is left out by proceeding in this way.

Let us then substantiate the problem with some (generally also accepted) further auxiliary assumptions.

A1 The principle of the Nomological Character of Causality (PNCC): Causal relations hold between token events but in virtue of these tokens instantiating types or properties featuring in causal laws.

Almost everyone agrees with A1.[3] Notice that A1entails that bona fide causal explanations in science are general. Essentially the same explanation, this is the idea, can be given for different token events provided they belong to the same relevant types. We assume also, as it is commonly done, that events belong to types in virtue of instantiating properties. Finally, A1 makes clear that the generality of explanation is achieved because each causal explanation is backed by a causal law, which can be the same for different token cases. The assumption hence includes the metaphysical view that causal laws are constituted by properties both featuring in their antecedent and consequent parts. Thus, a causal law covers an explanation of a token event if, and only if, the token event instantiates properties featuring in the consequent part of the causal law and the explanation resorts to tokens in the explanans instantiating properties in the antecedent part of the law.

We need also an auxiliary assumption making clear the nonreductive character of physicalism which does fit with the clarification of T2 offered by A1. Here it is:

A2 Identity without metaphysical property-reduction (NRD): Each coarsely-grained token mental event is identical to a coarsely-grained token physical event, but different tokens belonging to the same mental type may be identical to different physical tokens belonging to different physical types.

Given the assumptions in A1, auxiliary assumption A2 entails that we cannot simply identify mental properties with physical properties. There is strong dependence, but not identity, between mental properties and physical properties as it should be, given our commitment to nonreductive physicalism. This is why we couched A2 in terms of token events coarsely-grained, that is to say, as eventually involving two different properties, one mental, another one, physical. Also, T3 and A2 jointly entail that a finely-grained mental token (i.e., consisting merely in the instantiation of a mental property) is necessitated by a finely-grained physical token (individuated as merely consisting in the instantiation of a physical property).[4]

The next assumption is, as we will see in the next section, by far the most controverted in discussions of our general problem. Actually, many consider it the culprit. We will defer its discussion until we reach the next section; let us here, for the moment, just lay it out:

[3] A classical exception is allegedly Anscombe (1971).

[4] A finely-grained individuation of events assumes that a token event consists in the instantiation of a (relational) property by an object (or objects); a coarsely-grained individuation assumes that a token event may consist in the instantiation of several properties by some object or objects. For instance, Trump's headache yesterday at 4 pm may be coarsely individuated as a token event consisting in Trump's instantiating at 4 pm both a phenomenological property and a neurological one. If we assume instead a finely-grained individuation of events, then we should say that at 4 pm there were two token events: one exhausted by Trump instantiating the phenomenological property and another one exhausted by Trump instantiating the neurological property, assuming, of course, that these properties are distinct (see Kim 1976).

A3 No causal redundancy (NCR): There cannot be two complete but different causal explanations of the same fine-grained effect (i.e., involving the instantiation of a property), unless it is a case of standard causal overdetermination.

Some clarificatory comments about (NCR) are in order. First, A3 bears some resemblance with the principle of explanatory exclusion proposed long ago by Jaegwon Kim. Yet there are two important differences to note. The first difference is that Kim's principle involves explanations in general, not just causal explanations. This makes it dubious. It is certainly open to question that we cannot explain an event in both a causal and a non-causal way (for instance, by making a description of the constituents that make up the event in question). The other difference is that Kim's principle forbids two complete and *independent* explanations of the same event. Yet, as many commentators have noted, given that it is standardly assumed that a putative (fine-grained) mental cause of an event is dependent on its (fine-grained) physical cause (given T3 and A4 below), then the psychological explanation of such event will not be independent from its physical explanation, and hence the postulation of such explanations will be perfectly compatible with Kim's principle and, a fortiori, will leave us with no general problem to deal with. To avoid this, we dropped the requirement that the explanations are independent from each other. Second, as events can be individuated finely or coarsely, we have made clear in A3 that, when considering causal explanations, it is fine-grained individuations that matter. This is of course in accordance with A1, the nomological character of causality. Finally, in cases of (standard) causal overdetermination, we seem to have two complete yet different causal explanations of the same fine-grained event, each one involving one of the two causes overdetermining the said effect. Take the classical example of two shots fired by two different shooters reaching the heart of a victim at exactly the same time. So, we need to put these cases to one side sor that A3 can be accepted as true (again, we leave discussion of the parenthetical 'standard' for the next section).

Our final auxiliary assumption follows simply, according to most theorists (with one single exception to be discussed briefly in the next section), from our physicalist commitment when applied to causal explanations:

A4 The Causal Closure of the Physical Domain (PCCP): For every caused fine-grained physical event, E, involving the instantiation of a physical property, there is a complete physical causal explanation of E involving only the instantiation of physical properties.

The rationale for A4 is that, given our assumptions about causal explanations, its rejection would amount to a rejection of physicalism. To see this, suppose there was no complete physical causal explanation of a caused fine-grained physical event E. This means that a complete causal explanation of E would include some fine-grained non-physical event (involving the instantiation of a non-physical property C). Therefore, there would be a fundamental empirical fact, the causation of physical event E, which would not be a purely physical fact (since it would include C). Hence, not all fundamental empirical facts would be physical and, therefore, physicalism would be false.

We are now in a position to formulate our general problem as applied to psychological explanations. Suppose we want to explain a finely-grained mental event,

consisting in the instantiation of mental property M2, by mentioning a finely-grained mental event consisting in the instantiation of a different mental property M1, in the explanans. Given T2, we are dealing here with a causal explanation and (a token of) M1 is supposed to belong to the causal ancestry of (the token of) M2 we are trying to explain. Given T3 and A2, however, there are finely-grained physical events, P1 and P2, which metaphysically determine, and on which depend, respectively, M1 and M2.[5] This is so because A2 tells us that there must be coarse-grained physical events identical to M1 and M2 and also that mental properties are not identical to physical properties. Given A4, there is a complete physical causal explanation of the instantiation of M2 which will involve the instantiation of P1. This is so, because the instantiation of P1, we can assume, causally explains the instantiation of P2[6] and a fortiori also the instantiation of M2 given that, again by T3 and A2, the instantiation of P2 metaphysically determines the instantiation of M2.[7] Given A3 and given that this is no standard case of causal overdetermination because the putative causes of M2 (P1 and M1) are not independent from each other (more on this in the next section), it follows that there is no causal explanation of M2 involving the instantiation of M1. Since this would apply to any putative psychological causal explanation involving the instantiation of any mental properties M1 and M2, then given T2 we should conclude that there are no bona fide psychological explanations, thus contradicting T1.[8]

2 Causalist solutions to the problem

Different solutions to the above problem have been propounded, generally by rejecting some of the theses or of the auxiliary assumptions we have highlighted. The most radical solution is to defend an eliminativist position in relation to cognitive science (e.g. Churchland, 1981). That would amount to rejecting T1. Of course, after such rejection, the remaining claims form a consistent set. This is to place oneself at the opposite side where we want to be since we want to vindicate psychological explanations as bona fide explanations.

Something similar occurs with other suggestions. At some point, Jaegwon Kim, who did much to argue for the general problem (Kim, 1993), suggested as a possible solution that each mental token finely-grained should be identified with a physical

[5] To avoid an inconvenient profusion of variables, we will use the same variable to refer to a finely-grained event and to the instantiation of the property in which such finely-grained event consists. Thus, a finely-grained event consisting in the instantiation of a mental property M1 will be simply referred to as M1.

[6] If not P1, other finely-grained physical event must be the cause. So, as it is common in discussions of our general problem, it simply makes the matter simpler to assume that the cause is P1, the event on which M1 depends.

[7] One causalist trial to avoid the general problem used to be the dual explananda view (see Vicente 2002, for a review), according to which the explanandum of a psychological explanation would be a finely-grained mental event, while the explanandum of a physical explanation would be a finely-grained physical event. Since the explananda are different, so would be the explanations, even though they consist of finely-grained events which hold different relations of metaphysical dependence between them. Yet, as our reasoning makes clear, these relations of metaphysical dependence entail that whatever causally explains a finely-grained physical event E also explains whatever finely-grained mental event is metaphysically necessitated by E, thus cancelling this dual explananda strategy as a way of avoiding the problem.

[8] A similar reasoning would preclude a psychological explanation of a physical event by a mental cause.

token finely-grained (Kim, 1998). Given A2, however, this entails that different tokens of the same mental property will have different causal powers, thus rendering mental properties as causally heterogeneous. Given T2 and A1, however, this again would amount to rejecting T1.[9]

The subset account of realization of properties can be seen as an improvement on the previous proposal. According to this account, a physical property P realizes a mental property M if, and only if, the (forward-looking) causal powers of M are a (proper) subset of the (forward-looking) causal powers of P (Shoemaker, 2007; Wilson, 2011). Forward-looking causal powers of a property A are individuated by those properties the instantiation of which A can cause, usually together with the instantiation of other properties. In standard cases of multiple realizability of the mental, the powers individuating a mental property will be a proper subset of the powers individuating its physical realizers. Now, in this case mental properties will not be causally heterogenous and can feature in causal laws as required by A1 and T2. Yet, since physical realizers cause whatever is caused by the realized mental property, then the view would contradict A3, unless one can argue it is a standard case of causal overdetermination, which seems not to be the case, since in standard overdetermination the two causes are metaphysically independent of each other. This also raises the critical issue of whether the subset account gets right the sort of metaphysical dependence between the mental and the physical that physicalism requires (Pineda & Vicente, 2017; see also Morris, 2011). Notice that in order to satisfy T3, we need that mental events are metaphysically necessitated by their physical realizers. This can only happen, given the subset account, if we individuate properties, in general, by their forward-looking causal powers. But then, mental properties appear to be proper parts of their physical realizers and hence they seem to be more fundamental, against T3, which requires also that the mental metaphysically depends on the physical and not the other way around.

By far the most widespread causalist reaction to our general problem has been to reject A3. Notice that, as we just said, cases of psychological and physical explanations of the same finely-grained mental event cannot be taken as involving standard cases of causal overdetermination, because in such cases the two putative causes are metaphysically independent one from the other, but this is not the case with the mental and the physical cause given T3 and A2. Many theorists, however, have tried in different ways to argue that the causal redundancy involved in the mental and physical case should be accepted as well. There are basically two strategies in this regard. One strategy is to argue for an analysis of causation according to which such causal redundancy is only to be expected. The other strategy is to argue that the sort of metaphysical relation of dependence between the mental and the physical is such as to render acceptable causal redundancy.

To start with the first strategy, it has been argued that causal redundancies of the sort that our general problem problematizes simply follow from a difference-making analysis of causation. One brand of this analysis is the counterfactual analysis of causation (Lewis, 1973; List & Menzies, 2009), according to which C causes E if,

[9] A similar problem arises for classical disjunctivism, i.e., the view that a mental property is a disjunction of its physical realizers. Again, the resulting mental properties turn out to be causally heterogeneous: two tokens of the same mental property M which involve different physical realizers will have different causal powers.

and only if, had C not occurred, then E would not have occurred and had C occurred, then E would have occurred as well (with some refinements on which we need not enter). Using a possible world semantics for counterfactuals, and assuming that a finely-grained physical event P1 both determines a finely-grained mental event M1 and causes a finely-grained physical event P2 which, in its turn, determines a finely-grained mental event M2, it turns out that M1 causes M2.[10] However such analysis of causation has been justly criticized because it delivers wrong results in certain cases like causal preemption or double effects (McDonnell, 2017).

Another version of this first strategy is to use an interventionist account of causation. The interventionist account is also a difference-making account of causation which tries to overcome the limitations of the counterfactual account. The main idea is that (the instantiation of) a property P causes an effect E if an intervention on P (making it absent or present, depending on the cases) alters E (Woodward, 2003). The intervention should be suitable, meaning that potential confounders (other properties that may have a causal influence on E) are kept at bay when intervening on P. The notion of suitable intervention allows the interventionist to avoid the problems of the counterfactual analysis. Thus, an intervention on a double effect E1 making it present may make also present the other double effect E2 but only by making present the common cause C. Since C is a potential confounder here, such an intervention is however not suitable and one can argue that on the interventionist account double effects are not causes one from the other since there are no suitable interventions on one which changes accordingly the other.[11] Again, however, serious doubts arise about the interventionist analysis when it has to deal with properties linked by metaphysical relations of dependence, which is precisely the case with mental and physical causes if one assumes T3. It has been argued that since physical realizers are potential confounders of putative mental causes, either the analysis delivers wrong results in these cases or it simply cannot be applied to them (Baumgartner, 2009). Interventionists have replied by refining the notions of potential confounders and suitable interventions to avoid these bad consequences by arguing that realizers are not in fact confounders (Woodward, 2017; Polger et al., 2018), but it remains controversial whether they are successful.

The other strategy is to develop a theory about the metaphysical relation of dependence between finely-grained mental events, or mental properties, for short, and physical properties which is both consistent with nonreductive physicalism and allows for causal redundancies between mental and physical causes. Long ago it became clear that no relation of supervenience could accomplish this task. Every relation of supervenience states a relation of modal covariance between families of properties. But modal covariance falls short of metaphysical dependence (Horgan, 1993). Some physicalist-minded theorists proposed then stronger relations of dependence and called them

[10] The argument goes roughly as follows. In the closest possible worlds in which M1 does not occur, no physical realizer of M1 occurs and, a fortiori, nothing causes a realizer of M2. Hence, in such closest possible worlds, M2 does not occur. In the closest possible worlds in which M1 occurs, some physical realizer of M1 occurs which causes a realizer of M2. Hence, in such closest worlds M2 also occurs.

[11] As one can see from this brief discussion, difference-making analyses of causation are not reductive analyses, since one needs to use an intuitive notion of cause in order to test their suitability. This casts a serious shadow on their usefulness when dealing with metaphysical problems involving causation, like our general problem. We will need however to put this line of argument to one side, since we lack the space here to develop it properly.

realization relations.[12] The two best worked out proposals are the functional analysis and the subset analysis of realization. We have already laid out our misgivings about the subset view. As regards the functional analysis, the idea is that a mental (functional) property M is a property instantiated by something X if, and only, if X instantiates a physical property having the causal role CR individuating M (Melnyk, 2003). The problem here is that the analysis itself seems to make clear that all the causal work is done by the physical realizer and not by the mental functional property M, thus against T1&T2.

Finally, and more recently, a further notion of metaphysical dependence is gaining wide currency: grounding. Grounding is supposed to relate facts rather than properties, for instance, the fact that something X instantiates a mental property M is supposed to be grounded on the fact that X instantiates physical property P. Grounding is understood as a relation of metaphysical dependence which is actually a strict order (hence, asymmetric and irreflexive). Recently, some writers have argued that if mental facts are grounded on physical facts, then it follows that both mental causes and their physical grounders cause the same facts (Kroedel & Schulz, 2016; Stenwall, 2021). The trouble with this suggestion is that there is no elucidation of the notion of grounding other than it denotes a strict partial order. Such opacity makes it impossible to test claims like this one about mental and physical causal overdetermination (Wilson, 2014). Moreover, the notion of grounding brings along problems not had, for instance, by a notion of realization. For instance, that something grounds something else is supposed to be a fact and therefore it falls under the scope of the grounding relation. This raises the annoying issue of what grounds grounding facts, an important problem that brings us to either an infinite regress or to the postulation of ungrounded (partially) mental facts, thus contravening T3.

Irrespective of whether some analysis of causation or of the physicalist relation of dependence of the mental on the physical succeeds in making a case against A3, it is worth to close this section motivating A3, given that, as we said, it is the main target of causalist solutions to our general problem.

The main rationale for A3 is that if one drops it then one is left with massive cases of causal overdetermination in nature. Whatever is caused by a finely-grained mental event M will be as well caused by finely-grained physical events distinct from M.[13] One could perhaps reply that there is nothing wrong with massive causal overdetermination in nature. Well, the intuitive idea is that coincidences in nature are rare, and a good principle of epistemic prudence invites us to accept them only when we have good (independent) evidence for them and there is no alternative explanation of the event to be explained. This is why, in fact, A3 is compatible with standard cases of causal overdetermination. In standard cases of causal overdetermination, like the two bullets shot by different shooters causing the death of a victim at exactly the same time, we have good independent evidence for holding that we are dealing with two causes here, and not just one. This is so because even though the two bullets causally overdetermine the death of the victim there are many other effects that are only had by one of the bullets (e. g., the perturbations of the air caused by each bullet; maybe one bullet

[12] Notice that the etymology of 'realize' is to make real.

[13] This is mainly due to A4, as explained in the main text.

hits a leaf of a by-standing tree and the other not, or one bullet hits the head of the victim, other the heart, etc.) or are the conjoint effect of the two bullets (both hit the heart at approximately the same location making a hole bigger than if only one bullet had reached the heart, etc.) (Braddon-Mitchell & Jackson, 1996). The problem with massive cases of causal overdetermination in the mental and physical case is that its postulation lacks any of the features that epistemic prudence counsels. First, we do not have any independent evidence for the existence of the two independent causes, the mental one along with the physical one. All the evidence is purely ad hoc: we need to accept it in order to make compatible nonreductive physicalism with the idea that causal explanations in the psychological sciences are bona fide explanations. Second, we can explain mental effects without postulating such coincidences. For every effect E allegedly caused by the instantiation of a mental property we can explain E by mentioning physical causes and the fact that E is metaphysically necessitated by a physical effect. So, we end up postulating massive coincidences in nature for which we lack any independent evidence and such that can be epistemically dispensed. This is, to say the least, a shaky assumption to make. So, we think that A3 is on firm ground.

In view of the fact that decades of debates trying to find a solution to our general problem assuming a causalist notion of explanation have not proved fruitful enough,[14] our suggestion is to confront the problem from a completely different perspective. We propose to solve it by rejecting T2. We will defend that psychological explanations are bona fide scientific explanations, but not in virtue of being causal explanations.[15] In the next two sections we will substantiate this claim. In Sect. 3 we present an account of scientific explanation as ampliative, specialized embedding (ASE) that is non-causal and that has already demonstrated fruitful in other cases in physics and biology. In Sect. 4 we show that the same account can be successfully applied to psychological explanations in general, and cognitive science in particular, drawing on a particular cognitive explanation of the psychological déjà-vu phenomenon and the explanation of the mechanisms for action. The moral will be that psychological explanations, whose

[14] An anonymous reviewer objects that our formulation and discussion of the problem in these two sections engage too much, and unnecessarily, in the metaphysics of mental causation, while scientific explanation, and mental explanations in particular, are an epistemic matter. We agree on the later, as our ASE account below suggests, but the whole problem precisely consists in that standard causalist accounts of mental explanations make the epistemic explanatory import dependent on the metaphysics of the mind by demanding mental explanations being both causal and compatible with physicalism (another metaphysical thesis). Our proposal is precisely a way of breaking this dependence by proposing a non-causal elucidation of mental explanations preserving their explanatory value.

[15] Davidson'a anomalous monism may be thought to defend a similar view. Davidson held that the attribution of mental states is governed by a rationality constraint which is alien to the attribution of physical states. So, explanations in mental terms, by contrast to physical explanations, won't be causal, but, rather, so to speak'rational'. As it is well-known, however, such a view has been discredited by the passage of time. One objection had it that the view, contrary to what Davidson claimed, is dubiously compatible with physicalism. If a fine-grained physical event P metaphysically determines a fine-grained mental event M, then we would have a physical criterion for the attribution of M based on a physical condition, namely, P, which avoids the rationality constraint, something which according to Davidson does not make sense (Kim 1985). Later on, Davidson argued that all he was claiming was that there cannot be strict psychophysical laws, but his anomalous monism would entail that there are psychophysical ceteris paribus laws. As many commentators suggested, however, this view actually amounts to a physicalist non-reductivist position, not different from the one generating the general problem we have just discussed. Also, it remains unclear why the rationality constraint forbids strict psychophysical laws yet it allows cp laws.

causal elucidation suffers from the above difficulties, are better elucidated as (perhaps non-causal) ASE explanations.

3 Scientific explanation as ampliative, specialized embedding (ASE)

The account of scientific explanation as ampliative, specialized embedding (ASE) can be considered a neo-Hempelian analysis in that it departs from the standard Hempelian account of explanations as "nomological expectabilities" and modifies it as little as possible to solve its well-known difficulties while preserving Hempel's empiricist strictures. ASE is non-causalist in that it does not conceptually require that the explanans must provide causal antecedents of the explanandum.

Before starting with ASE, it is worth clarifying that what is at stake is the notion of "possible explanations", that is, what kind of things explanations are. We are not concerned here with whether psychological explanations are "materially correct", i.e. have true explanans, or other material epistemic virtues. An "overall" correct explanation is an explanation that is both (i) "conceptually correct/possible" and (ii) "materially correct". We are here concerned just with (i), for (ii) is not a matter of philosophical analysis but of theoretical-and-empirical acceptability (although some philosophical concerns may affect also the content of the explanans). This said, we focus now in (i), what kind of thing scientific explanations are, what kind of things explanans and explanandum are and what the relation must be for the former to (possibly) explain the later. This is important, for in the next section we are going to take as case studies two typical cognitive explanations of psychological phenomena, regardless the (material) problems they may have for some psychologists that do not endorse them.

According to the standard Hempelian account from which ASE departs, to explain a phenomenon consists, roughly, in (deductively-DN- or inductively-IS) inferring it from antecedent conditions and nomological regularities that connect such conditions with the explanandum. That is, to explain consists in making the explanandum *nomologically* (DN-certainly or IS-probabilistically) *expectable* from antecedent conditions. This Hempelian model suffered from a series of deficiencies that made it soon criticized, due to different kinds of counterexamples that proliferated in the literature (see Salmon, 1989 for a summary). On the one side, indeterministic explanations with low probability (the explanans makes the explanandum not highly probable but just *more* probable, e.g. major's paresis case) makes it doubtful that explanations, at least indeterministic ones, are valid inferences. On the other side, and as a general problem for both deterministic and indeterministic explanations, nomological expectability seems not sufficient for explanatoriness, for there are cases of nomological expectability that seem clearly non-explanatory (symmetries, forks, irrelevances, merely phenomenological expectabilities-e.g. Galilean kinematics, Keplerian astronomy, more on this below-, and others). These problems motivated the development of alternative proposals.

With regard the non-sufficiency of nomological expectability, the obvious move seems to be to look for additional conditions X that added to nomological expectability correctly distinguish between explanatory and non-explanatory expectabilities

depending on whether X is or is not also satisfied. The two main alternatives, causalism (Lewis, 1986; Salmon, 1984; 1998; Woodward, 2003, including the recent new mechanicism Glenan 1996, Machamer et al., 2000) and unificationism (Kitcher, 1981, 1993) can be read as two alternative proposals for the missing X. Causalists demand that the particular antecedent conditions in the explanans must be part of the causal history of the explanandum. Unificationists, in turn, demand that the inference belongs to the best inferential system, i.e. the system that best balances simplicity and strength. Causalist and unificationist alternatives solve most of Hempel's problems, but not all, and generate some new ones.

With regard causalism, causation seems neither conceptually necessary nor sufficient for explanation. On the one hand, as many have argued, causation does not seem to be conceptually necessary for explanatoriness, for one may find bona fide explanations in different scientific fields whose causal nature is far from clear, among the most mentioned (leaving for now psychology aside): several social sciences, some parts of biology, quantum mechanics, relativistic gravitation, reductive explanations, network biological explanations, among others; see Díez, 2014 for a summary). On the other hand, it is at least unclear that causation suffices for explanatoriness. Nomological expectabilities based on merely descriptive or phenomenological generalizations do not seem explanatory, regardless of whether one refers to causal antecedents of the explanandum For instance, one can nomologically predict a posterior position of a planet at a time from a position and velocity at a previous time, and Kepler's laws, nevertheless such expectability, though nomological, can hardly count as explanatory given the merely descriptive nature of Kepler's laws: these laws describe systematically planets' movements, but they do not tell *why* they move as they move. And, arguably, prior positions are part of the causal history of posterior positions (what tells us not only how but also *why* planets move so is Newton's celestial mechanics, that introduces masses and forces). Likewise for Galilean kinematics and other descriptive theories such as merely descriptive, Mendel's statistical, non-accidental phenotypic regularities.

As for unification, it is neither necessary nor sufficient either. Unification does not always have explanatory import. Galilean kinematics or Keplerian astronomy are again cases in point, for they use merely descriptive or phenomenological laws which are unificatory (at least at their times) although as we have seen the expectations/predictions based on them can hardly qualify as explanatory (not even at their times). On the other side, some bona fide explanations can hardly qualify as unifying; for instance, Newton's gravitational explanation of free fall cannot (*taken alone*) qualify as unifying but this does not undermine its explanatory value (which may, of course, *increase* within the whole unifying Newtonian theory).

ASE is proposed by Díez (2014), elaborating on some ideas from Sneedean structuralism (cf. Balzer et al., 1987; Bartelborth, 2002; Forge, 2002), as a new manner to overcome Hempel's difficulties without falling into either causalism or unificationism, and sticking close to Hempel's empiricists strictures. According to Díez, ASE qualifies for a minimal, yet substantive, general theory of explanation applicable across scientific practice. He claims that his minimal account is compatible with acknowledging additional (causal manipulativist, causal mechanistic, unifying, …) features in specific

fields, that *add* supplemental explanatory values, though none is conceptually *necessary* for (minimal) explanatoriness. Thus, every scientific explanation is (minimally) explanatory because is an ampliative, specialized embedding; and some explanations are in addition causal, some others are in addition unifying, still others are in addition reductive, etc., and maybe some are just ASE. So, although ASE makes room for a variety of explanations, it is not simple pluralism for it requires a monistic conceptual core that suffices for minimal, yet sufficient, explanatoriness. The account, though minimal, is not minimalist, or deflationary, for the ASE conditions are not platitudinal but substantive.

ASE preserves the core of the Hempelian nomological expectability, though formulated within a model-theoretic framework with the notion of *nomological embedding*. The basic idea is that explaining a phenomenon consists of in (at least) embedding it into a nomic pattern within a theory-net (see Díez, 2014 for details). Now explanandum and explanans are certain kinds of models or structures: the data model DM = < D1,…,Dn, f1,…, fi > one wants to explain, and the theoretical model, TM = < D1,…,Dm, g1,…,gj > , which must involve at least the same kinds of objects and functions (but can introduce new ones, more on this crucial point soon), and that is defined by the satisfaction of certain laws. In the Classical Mechanics Earth-Moon case, for instance, the explanandum is the data model that represents the Moon's spatio-temporal trajectory around the Earth actually measured, and the explanans model is the dynamical structure including masses and forces and satisfying Newton's Second Law and the Law of Gravitation. To explain the Moon's trajectory consists in embedding it in the mechanical system, i.e. to obtain/predict the Moon's- kinematic trajectory from the dynamical model. Embedding here means (if we simplify and leave idealizations aside now), that the data model is (or is isomorphic to a) part of the theoretical model. Likewise in other cases. In Mendelian Genetics, for instance, the explanandum model describes the patterns of transmission of certain phenotypes, and the explanans model includes also genes and satisfies certain genetic laws. The transmission of traits is explained when one embeds traits transmission into the theoretical model, that is, when the observed phenotype statistics sequence coincides with what is predicted by the full genetic model.

ASE's basic idea is that if the explanation succeeds, then we find the, previously measured, data to be explained as part of the theoretical model defined by certain laws (as the structure that satisfies such laws). That is, if things behave as such laws say, then one should find some results at the data level; and when the actual data coincide with the expected results, the embedding obtains and the explanation succeeds. This account preserves the nomic expectabilty idea. The embedding provides the expectability part, for if the embedding succeeds one may "expect" to find the explanandum data as part of the theoretical model. This expectability, though, is weaker than Hempel's inferentialism for it does not demand that explanations must be logical inferences *stricto* sensu (then it is not subject to the "explanations are not inferences" criticisms), but nevertheless makes also room for both deterministic and probabilistic (including low probability cases, if needed) explanations depending on whether the regularities that define the explanans models are deterministic or probabilistic. The nomic component comes from the fact that the explanans theoretical model that embeds the explanandum model is defined by the system satisfying certain laws, understood merely as

non-accidental, counterfactual-supporting generalizations (cf also Mitchell, 1997). As Díez emphasizes, this sense of nomological explanation is quite modest, meaning just that the explanans model satisfies certain non-accidental generalizations, no matter how ceteris paribus, local, or domain restricted they are. On the other hand, the explanandum data model is measured without using the laws that define the theoretical model, i.e. independently of such laws, what guarantees that the intended embedding is not trivial and may fail.

All this is a model-theoretic weakening of Hempel nomological inferentialism, but says nothing yet as how to face Hempel's non-sufficiency difficulties, that is, what X should one add to nomological expectability for distinguishing explanatory and non-explanatory embeddings. The missing X that ASE proposes consists in two new conditions: for the nomological embedding to be explanatory, it must be both *ampliative* and *specialized*.

With regard to ampliativeness, as our mechanical and genetic examples illustrate, the explanans model must include additional ontological (in metaphysical terms) or conceptual (if one prefers a more epistemic formulation) components with respect to the explanandum model. In the mechanical case, the explanans includes, together with kinematic properties, new dynamic ones, namely masses and forces, behaving with the former as the mechanical laws establish. In the genetic case, the explanans model includes, together with the phenotypic properties, new genetic ones, genes or factors that behave with the former as the genetic laws establish. This ampliative character of the embeddings is what explicates their explanatory nature compared to other embeddings that lack explanatory import. In the Keplerian case, for instance, a nomological embedding is also found, but this embedding does not qualify as explanatory since the explanans model (defined by Kepler's laws) does not introduce additional conceptual/ontological apparatus with respect to the explanans; both the explanandum and the explanans model include only kinematical properties. Likewise for Galilean kinematics. And the same applies to the genetics example, if one takes purely phenotypic statistical regularities as defining models that embed certain phenotypic data (Mendel's purely phenotypic laws). Again, this would be a case of a nomological embedding with no explanatory import. Nomological embedding without ontological/conceptual ampliation is not explanatory (this of course breaks the alleged Hempelian symmetry between -nomological- prediction and explanation).

The second additional condition is that the ampliative laws used for defining the explanans model must be "special" laws, and not merely schematic or programmatic principles. This distinction originates in Kuhn's (1974, p. 465) difference between "generalization-sketches" and "detailed symbolic expressions",[16] further elaborated by structuralist metatheory as the distinction between guiding-principles and their specialized laws in a *theory-net* (see e.g. Balzer et al., 1987, for several examples).

[16] "[…] generalizations [like f = ma…] are not so much generalizations as generalization-sketches, schematic forms whose detailed symbolic expression varies from one application to the next. For the problem of free fall, f = ma becomes mg = md2s/dt2. For the simple pendulum, it becomes mgSinθ = – md2s/dt2. For coupled harmonic oscillators it becomes two equations, the first of which may be written m1d2s1/dt2 + k1s1 = k2(d + s2 – s1). More interesting mechanical problems, for example the motion of a gyroscope, would display still greater disparity f = ma and the actual symbolic generalization to which logic and mathematics are applied". (Kuhn 1974, p. 465).

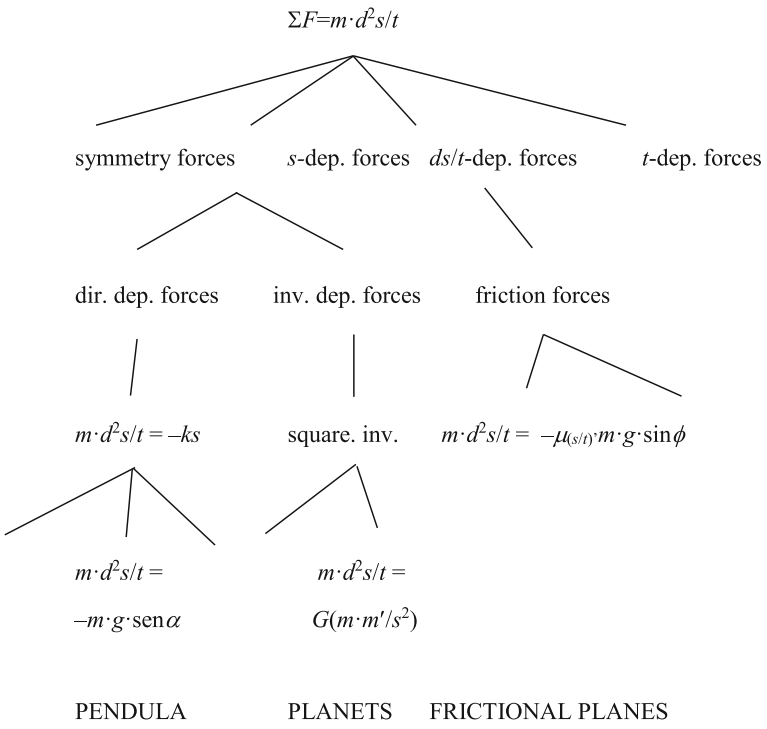
Most theories are hierarchical net-like systems with laws of very different degrees of generality within the same conceptual framework. Often there is a single fundamental law or guiding principle "at the top" of the hierarchy and a variety of more special laws that apply to different phenomena. Fundamental laws/guiding principles are kind of "programmatic", in the sense that they establish the kind of things we should look for when we want to explain a specific phenomenon, and the general lawful scheme that specific laws must develop. It is worth emphasizing that general guiding principles taken in isolation, without their specializations, are not very empirically informative for they are too unspecific to be tested in isolation. To be tested/applied, fundamental laws/guiding principles have to be specialized ("concretized" or "specified") by specific forms that, in the above referred Kuhn's sense, specify some functional dependences that are left partially open in the laws above in the branch.

The resulting structure of a theory may be represented as a net, where the nodes are given by the different theory-elements, and the links represent different relations of specialization. For instance, the theory-net of Classical Mechanics (**CM**) has Newton's Second Law as the top unifying nomological component, i.e. as its Fundamental Law or Guiding Principle (Balzer & Moulines, 1981; Moulines, 1978/1984; Balzer et al., 1987; Díez & Moulines, 2022), that can be read as follows:

CMGP: For a mechanical trajectory of a particle with mass m, the change in quantity of movement, i.e. $m{\cdot}a$, is due to the combination of the forces acting on the particle.

This **CMGP** at the top specializes down opening different branches for different phenomena or *explananda*. This branching is reconstructed in different steps: first, symmetry forces, space-dependent forces, velocity-dependent forces, and time-dependent ones; then, e.g., the space-dependent branch specializes into direct and indirect space-dependent; direct space-dependent branch specializes in turn into linear negative space-dependent and…; inverse space-dependent branch specializes into square inverse and…; at the bottom of every branch we have a completely specified law that is the version of the guiding-principle for the specific phenomenon in question: pendula, planets, inclined planes, etc. (Kuhn's "detailed symbolic expressions").

The theory-net of **CM** looks (at a certain historical moment) as follows (only some, simplified, terminal nodes are shown here, which suffices to our present goals; at bottom in capitals examples of phenomena explained by the branch):



Now we can spell out the second additional condition for an embedding to be explanatory, namely, the (ampliative) embedding must be *specialized*. The basic idea is that among the (non-accidental) generalizations that define the explanans model, at least one must be a specialization, i.e. the explanans model cannot be defined exclusively using general guiding principles because otherwise the embedding becomes trivial, empirically void. Take for instance Newton's second law alone, $\Sigma F = m \cdot d^2\, s/t$ without any specific systematic constraint on the kind of functions f that we can make use no matter how crazy these functions were, then, as Diez points out, "with just some mathematical skill we could embed any trajectory" (2014, p. 1425); even for weird trajectories, such as the pen in my hand moved at will, with enough purely mathematical smartness one could find out $f1, f2, \ldots$ whose combination embeds it. Or take the general guiding principle of Ptolemaic astronomy "For every planetary trajectory there is a deferent and a finite series of nested epicycles (with angular velocities) whose combination fits the trajectory". As it has been proved (Hanson, 1960), any continuous, bounded periodical trajectory may be so embedded. The moral, then, is that for a nomological embedding, even ampliative, to be properly explanatory and not merely an ad-hoc trick, the explanans model must be specified using some special law in the referred sense.

One could object that, given this presentation of the notion of special law, one then could find bona fide explanations only in highly developed unified theories of the net-like structure, which seems counterintuitive for, as we ourselves noted above, there are bona fide explanations quite isolated, e.g. Newton*'s gravitational explanation of free

fall previously to developing the rest of his (unifying) mechanics. This is a reasonable concern due the way we have introduced the notion, but it is actually not the case. Although the notion of a special law is particularly clear by contrast to that of a general principle in the framework of a theory-net, it is not true that special laws can exist, and be applied, exclusively within an already highly developed theory-net. The law for harmonic oscillators, for instance, is a special law, as the Law of Gravitation also is, and they are so no matter when they were discovered or whether there were integrated in a bigger, unified theory-net with a top guiding principled already formulated (in several theories, some special laws are formulated even before that a general guiding principle is *explicitly* formulated, cf. e.g. Lorenzano, 2006 for the case of Classical Genetics).

It is worth emphasizing that, though minimal, ASE is not minimalist or platitudinal. Hempel's nomological expectability was not platitudinal to start with, and the two new conditions added, ampliativeness and specialization, though minimal additions and keeping empiricist strictures, are not platitudinal either, they are substantive ones that some embeddings, e.g. Keplerian astronomy or Mendelian Genetics, do not satisfy. We want to conclude our presentation of ASE by insisting that it is not incompatible to acknowledging that there may be *additional* explanatory virtues, such as causation, mechanisms, unification or reduction (when they are explanatorily virtuous). ASE's point is that, no matter how explanatorily valuable these other features are, none is *conceptually necessary* for explanatoriness. All nomological embeddings that are intuitively explanatory comply with ASE, and the intuitively non-explanatory fail in some ASE conditions.[17] It is our claim that this account be used to explicate the explanatory autonomy of psychological explanations without falling prey to the above "causalist problem".

4 Psychological explanations as ampliative, specialized embeddings

We are going to defend our claim not by general considerations but by exemplification, i.e. by taking two cases that, according to the scientific community, are bona fide candidates of psychological explanations and showing that, problems with causalism notwithstanding, they fit well in the ASE analytical framework. What matters is that they are considered bona fide psychological possible explanations, even by opponents who object to their theoretical-or-empirical adequacy. If they satisfy ASE conditions,

[17] An anonymous reviewer objects that "the key" of ASE account is embedding, and since as we acknowledge this is basically (a mode-theoretic version of) prediction/expectability, ASE faces the traditional insufficiency problems for successful predictions "often have nothing to do with explanations", and then refers to some additional examples of non-explanatory nomological predictions. It is worth emphasizing that prediction/embedding is *not* "the key" of ASE. ASE has no single key but three: embedding, ampliativeness and specialization; the three are necessary, and jointly sufficient, but also "individually insufficient". On the other side, we disagree that the existence of non-explanatory embeddings means that embedding has "nothing to do" with explanation: one thing is that prediction is not individually sufficient, quite other thing is that it has nothing to do with explanation for, though insufficient, it may be necessary, as ASE, and any other (with slight differences) account, claim. The explanatory work is not completely done by prediction, but it is partially done by it.

this suffices to us for vindicating their explanatory autonomy independently of causalist considerations.

Standard psychological explanations explain different (non-verbal as well as verbal) behaviors calling for "mental/psychological states" responsible for the behavior in point. Folk psychology explains behavior in terms of beliefs and desires, or, in general, doxastic and conative, mental states that relate to behavior in a special, non-trivial nomological manner, e.g. "if a subject S desires to meet a person P, and believes that P is going to be in a certain place L at a given moment T, and that performing action A she will get to location L at T, then (cp) she intends to perform action A".

Different scientific psychologies develop this prescientific explanatory practice in different manners. Behaviorism substitutes mental states and processes by behavioral dispositions and processes such as conditioning, inhibition, recovering, and others, and explains behavior by postulating nomological, non-accidental regularities that relate behavior to specific dispositions and processes: forward (delay/trace) conditioning, simultaneous conditioning, second order conditioning, backward conditioning, etc. Neuroscience substitutes ("reduce"?) mental states by brain entities and states, postulating nomological connections relating brain states to behavior, for instance the ones that neuroscientists postulate for vision (e.g. Farah, 2000). Both behaviorism and ("pure") neuroscience depart from folk psychology in that both get rid of (eliminate, or reduce) standard mental states such as beliefs and desires in favor of "different kind" of entities and processes. Cognitive psychology, on the contrary, preserves belief, desires, emotions, and other mental states close to their folk interpretation but develops a much more sophisticated theoretical network of nomological connections with behavior, "scientifying", so to say, folk psychological explanations and the corresponding explanatory notions: exogenous/endogenous control attention, short-term/long-term memory, perception, metacognition, etc. This cognitive explanatory machinery is applied to a whole variety of psychological explananda, specifying particular mental states, processes and processings for particular phenomena. In order to test whether ASE elucidates well the explanatory practice in cognitive psychology, we are going to take two specific explanations of two particular psychological phenomena, enough simple to be tractable here but also enough representative of cognitive explanations as for serving as keystones of how ASE fares with respect to explanatoriness in cognitive psychology: Cleary and collaborators' cognitive explanation of déjà vu, and goal-directed vs stimulus-driven explanations of instrumental behavior.

4.1 Déjà vu

Déjà vu is the psychological phenomenon that happens when the subject has (i) the feeling that she has already lived the same "scene" or "experience" that she is living at the moment, (ii) the inability to recall what and when that happened, and (or given) (iii) the belief/awareness that she actually has not lived that scene before (Nepe, 1983; Brown, 2004; Kusumi, 1998, 2006; Cleary et al., 2009, 2012, 2019). One scientific explanation of some cases of déjà vu has to do with neurological conditions, such as epilepsy: in temporal-lobe-epileptic cases the brain triggers déjà vu experiences by spontaneous neural activity in absence of external stimuli (O'Connor and Moulin,

2008). Kusumi (1994, 1996, 1998, 2006) and Cleary, alone (2008, 2014) and with collaborators (C&C hereafter) (2009, 2012, 2018, 2019); Clearly and Claxton (2018) independently propose that at least some other cases are a memory phenomenon that have a cognitive stimulus-driven explanation related to non-recall familiarity-based recognition, and conducted a series of experiments to test it.

Familiarity is a memory phenomenon in which the subject has "a feeling of familiarity": the current place, face or, in general, scene "feels familiar" to her. On some occasions (recall-familiarity), the subject is able to recall a past experience that is the source of the familiarity, for instance, a prior occasion in which she saw the same face or place she is now seeing accompanied by the feeling of familiarity. On many occasions, though, the feeling of familiarity is not followed by the recall of the previous familiar scene. On some of these occasions there is no recall simply because there is no previous experience of the same scene, one feels familiarity despite one has never previously experienced it. When one is aware that it is the first time one confronts the scene but nevertheless has the feeling of familiarity, one has a déjà vu experience. In some déjà vu occasions, although the subject does not recall the previous experience of the same scene since there simply is no one and the subject is aware of that, she nevertheless recalls some previous experience of a scene similar in some specific respects to the current scene. This may suggest that déjà vu is based on similarities between the current scene and a previously experienced different but related scene, even when, as often, one is unable to recall any such past experience. This is the source of the familiarity-based recognition (in Cleary, 2008 terms, FBM hereafter) explanation of (some cases/types of) déjà vu. The explanation may seem natural once it is put forward after the empirical data, but it is far from that; actually, the previous most common non-neurological explanation in psychiatry was that déjà vu is a mental disorder in the family of psychotic hallucinations, schizophrenia or extreme fatigue; and according to other previous explanations (e.g. Brown, 2004), déjà vu is unconnected to the empirical world, contrary to what FMB claims.

According to C&C, FBM déjà vu obtains when the current scene, e.g. a place, has relevant similarity with a previously experienced (and often non-recalled) one. This relevant similarity is mainly of two kinds:

> "a strong familiarity signal can stem from a high degree of overlap between the elements of the current situation and those of one particular prior situation, or it can stem from more global familiarity resulting from a moderate degree of overlap between the current situation and each of multiple prior situations that have been stored in memory." (Cleary, 2008, p. 4).

When the familiarity is accompanied by the awareness that it is the first time that one experiences the scene, the familiarity recognition phenomenon is of the déjà vu kind:

> "Because déjà vu occurs when one experiences a sense of having experienced something before despite evidence to the contrary, déjà vu experiences may be limited to situations in which there is a strong global match producing a feeling of familiarity, an inability to identify the source of the familiarity, and evidence suggesting that the event could not have been experienced before. When

a situation meets the first two criteria but not the third, it may simply be labeled as a feeling of familiarity (and not a déjà vu experience). However, in both cases, the underlying process may be the same: it may be familiarity operating in the absence of identification of its source." (Ibd. 4).

The explanation receives support from the observed positive correlation between the increase in the number and variety of previous experiences and the number of déjà vu experiences:

"there is a positive relationship between frequency of reported déjà vu experiences and frequency of travel (Brown, 2003), frequency of reported dreams (Brown, 2003; Wallisch, 2007), and frequency of movie watching (Wallisch, 2007). Such relationships would be expected if déjà vu reflects familiarity-based recognition, as people who travel more often, dream more often, and watch movies more often should have more potential sources of familiarity stored in memory than people who experience these activities less frequently." (ibd. 3).

C&C conducted a series of experiments (2009, 2012, 2019) to test their explanatory hypothesis, according to them with positive results. They acknowledge, though, that the explanation is not free from difficulties, for instance, the frequency of déjà vu experiences declines with age despite the accepted fact that familiarity-based recognition remains impervious to aging (Cleary, 2008).

Kusumi (2006) proposes a similar explanation. According to him, déjà vu is not a memory disorder but a normal adaptive metamemory phenomenon:

"[a] normal metacognitive mechanism … that occurs during an analogical reminding process in which a present experience automatically reminds an individual of similar past experiences. Therefore, the déjà vu experience is generated by similarities between a present experience and corresponding past experiences." (p. 303). "When individuals feel a sense of familiarity with a present experience or problem, they retrieve past experiences or problems by matching cues. They evaluate the similarity and dissimilarity between the two experiences and then transfer useful information from past experiences to the present one. This process performs the same function as analogical problem solving (Holyoak & Thagard, 1995), in which a similar old problem provides a solution to a new one. This metacognitive mechanism appears to be adaptive in humans" (p. 312).

Kusumi also makes empirical experiments to test his theory, in his case a battery of questionaries that according to him confirm it (1994, 1996, 1998).

It is our claim that the ASE model of scientific explanation outlined in Sect. 3 elucidates well this case of cognitive explanation. First, it is a case of nomological embedding/expectability. The data and the experiments conducted by both C&C and Kusumi are intended to demonstrate that given their explanans conditions, the explanandum déjà vu phenomenon is expectable. In this case we have probabilistic expectability for the experiments do not grant that given the explanans conditions then we can deterministically expect a déjà vu, but only that the explanans are positively statistically relevant for the explanandum (C&C, 2009, 2012; Kusumi, 2006). And this expectability is claimed not to be accidental but nomological: the connection between

past similar experiences and déjà vu is claimed to be due to a non-accidental, counterfactual supporting cognitive regularity, as it is implicit in C&C and almost explicit in Kusumi's summary of the cognitive process in the last quote. It is clear that in both cases the regularities are counterfactual-supporting, as also their experimental work assumes.

But FBM also satisfies the two additional ASE conditions for explanatoriness, namely ampliativeness and specialization. With regard ampliativeness, the explanandum is déjà vu phenomena, that is a feeling of familiarity accompanied by the awareness that the experience is experienced for the first time. And the explanans introduces new theoretical machinery, roughly a previous perceptual experience, relations of partial overlap and of overall similarity between perceptions, cognitive storage, and a second-order metacognitive process of (implicit) comparison and transfer of information by the subject.

As for specialization, it is also apparent that the non-accidental cognitive regularities mentioned in the explanans are not just general schematic guiding principles but lower-level generalizations with specific empirical content. We do not claim that there are not such general principles involved in cognitive explanations in general and in FBM in particular. It seems plausible that cognitive psychology (implicitly) has such general guiding-principle, something of the kind: "if one wants to account for a cognitively based psychological event (behavior, feeling or emotion), look for antecedent mental events and for first and second order mental processes that give (sometimes together with other motor processes) rise to the psychological event in point". And different psychological phenomena would be explained by different specializations of this general principle, as different kinematic phenomena are explained by different specializations of the general mechanical principle, or different adaptive phenomena are explained by different specializations of a general Natural Selection guiding principle. The complete details, though, require a detailed reconstruction of the theory, which goes beyond our present needs. It suffices for now to witness that the nomological regularities involved in the present case relating previous perceptions, perceptual similarities, cognitive storage, and metacognitive comparison and information transfer are *concrete specifications* of mental events and first and second order mental processes postulated by the theory for our specific explanandum in point, thus squaring with ASE's specialization condition.

We conclude, then, that the FBM account of déjà vu satisfies ASE conditions and that it is in virtue of this that it has its explanatory power, regardless of whether mental states are causal or not. It is worth emphasizing that all this is compatible with the eventual disregarding of FBM as a materially (theoretically-or-empirically) incorrect explanation of déjà vu. Actually, we have already seen that it is not free from empirical difficulties. And it is not free of a need of a better theoretical development either, for instance, specifying better the origin of the difference between FBM déjà vu and other kinds. But what matters here is not material correctness but conceptual adequacy, that is, that the practice is taken seriously in the field as a bona fide possible/candidate explanation of the phenomenon, as FBM actually is. Our claim is that this bona fide explanatory practice is well explicated by the ASE account of scientific explanation, regardless of what finally happens with the causal powers of the involved mental events when the above causal problem is solved (if ever). The fact that psychologists

take FBM as a bona fide possible explanation without waiting for the causal diagnosis suggests that they do not make their claim conceptually dependent on what finally happens with causation.

4.2 Mechanisms of action production

Psychological literature specialized on action postulates two processes or mechanisms subserving the production of action: the stimulus-driven and the goal-directed mechanism (Dickinson & Balleine, 1994).[18] A stimulus-driven mechanism is one in which detection of a stimulus activates a response (which is understood as an action tendency) by way of an existing associative link between stimulus and response. An action-tendency is understood as a mental state of readiness to perform a certain type of behavior, for instance, to avoid or to attack an object. Sometimes the response may be a tendency to suspend interaction with an object, which is why sometimes in these theories one speaks of (in)action tendencies as responses (Frijda, 2007). The activation of this associative link requires that the agents regard the present stimulus as similar enough, but not necessarily a perfect match, to the triggering stimulus. It is very important to highlight that on this mechanism the response, the (in)action tendency, is selected without the intermediation of any representation of any goal. Associative links between stimuli and (in)action tendencies may in some cases be innate (for instance, a startle response when hearing a loud noise), while many of them occur as a result of learning, notably associative learning, or of mere re-exposure to the same stimuli being responded with the same (in)action tendency. By contrast, on the goal-directed mechanism the response or (in)action tendency selected is that with the highest expected utility. Hence, this mechanism requires from the agent to (perhaps implicitly) consider and value the outcomes of each relevant possible response and to calculate the expectancy of obtaining each outcome in the contextual circumstances. So, this mechanism requires, among other things, representations of values or goals and comparison between them.

Actions produced by a goal-directed mechanism are flexible and can easily adapt to changes in the outcomes of actions and their values. For instance, if an agent notices that her extremely kind behavior with others, which used to make her popular and well regarded, now elicits the impression of pretense, she will easily adapt and change her behavior accordingly. If, by contrast, her behavior is under the control of a stimulus-driven mechanism, since the response is selected without consideration of goals, changes in the outcomes and their values will not alter such behavior, thus becoming maladaptive (one speaks of bad habits, in these cases). Based on this, a standard experimental technique to diagnose whether an action is produced by a stimulus-driven or a goal-directed mechanism is to degrade or devalue the relevant

[18] The reader may have noticed that in these, and other examples, cognitive psychologists eventually refer to "psychological mechanisms". We acknowledge that on one standard usage, the word 'mechanism' strongly suggests a causal process. Yet, in many scientific contexts the word is used more loosely referring broadly "to what provides the explanation", i.e. to the explanans, leaving open whether there is a causal process strictly understood. We then do not believe that the wording of some explanations using the expression "mechanism" is uncontroversial evidence for the causal nature of the explanation provided and thereby an objection to our non-causalist account.

outcome. It is the devaluation method (Adams & Dickinson, 1981). If the behavior remains the same, it is stimulus-driven; if it adapts and changes accordingly, it is goal-directed. Thus, a stimulus-driven action is very rigid and can easily become maladaptive, but due to its computational simplicity it is optimal for extremely rapid and precise actions (for instance, in skilled behavior) and can operate under very poor conditions.

Psychologists differ as to which of the two mechanisms is prevalent. According to some, the default system is the stimulus-driven mechanism (Wood & Neal, 2007). According to others, the goal-directed mechanism is the typical one, especially in the human case (Moors et al., 2017). This and many other issues of importance should be addressed in a proper discussion of these models of actions, but we think what we have just said is enough to see that they fit the ASE account. Let us then show this.

Regarding embeddedness, notice how the theory makes different predictions in experiments using the devaluation method, depending on whether the action results from a stimulus-driven or a goal-directed mechanism. Hence, if the action is produced by a stimulus-driven mechanism, since the response is selected without the representation of any goal, we expect that a change in the goals attained by the action will not change it in the least. These predictions/expectabilities are obtained by using regularities that are taken to be non-accidental, that is, counterfactual-supporting, thus nomologically in the weak sense of law that ASE demands.

Regarding ampliativeness, we can see how the theory introduces notions and entities which are not among the explananda, which merely consist of data about plain behavior. Thus, some of these additions include (in)action tendencies, associative links, representations of goals, or computation of expectancies. Also, these models assume that an action occurs when the selected (in)action tendency is translated into overt behavior by way of some regulative mechanisms. The job of such mechanisms is basically to ensure that the action tendency translates into an adequate behavior given the specific contextual circumstances by defining sub-goals and monitoring whether they are satisfied and, if they are not, defining new subgoals accordingly in a process which typically involves feed-back loops. For instance, assume that the selected response is leaving the scene. Then, given the contextual circumstances, some subgoals will need to be defined, like opening the entrance door, putting the right hand on the knob, etc. Hence, some regulative mechanisms are required and they are also additions to the explananda. Therefore, the non-accidental generalizations used in the predictions nomologically connect, as ASE claims, the explanandum machinery with the new machinery introduced by the explanans.

Finally, these models also satisfy ASE's specialization condition. The easiest way to see this is to realize that the theory assumes a sort of guiding abstract principle to the effect that an action occurs when, given a stimulus situation (which may be real, remembered or imagined), an (in)action tendency is selected which finally translates into behavior by way of the operation of some regulative mechanisms. Thus, both the stimulus-driven and the goal-directed mechanism can be seen as special cases of how the (in)action tendency is selected, which specialize in turn in more specific stimulus-driven or goal-directed mechanisms, respectively, for explaining specific behaviors.

In sum, we can see that this psychological explanation of the production of actions is well elucidated by ASE in terms of ampliative, specialized embeddings, regardless

of the diagnosis about its causal nature. It is true that many proponents of these models think of them as revealing causal mechanisms of action. But, as we have argued, if they are correct, this will simply add an additional contextual value to their status as scientific explanation, a status which will in fact survive an eventual demonstration that, due to the general problem above or to other more specific reasons, these cognitive mechanisms are not after all causal. This is, in sum, the point we have been trying to bring home in this paper.

5 Concluding remarks

We have shown that two paradigmatic, bona fide explanatory practices in cognitive psychology fit well the ASE conditions for explanatoriness, independently of the final diagnosis with regard causal powers of mental events. This, according to us, sufficiently explicates the explanatory import of these practices, and their epistemic autonomy. We believe that the same applies to other cases in cognitive psychology, such as a related explanation of cryptomnesia (Macrae et al., 1999, McCarrol & Sant'Anna, 2023), or Beck's cognitive explanation of depression (Beck, 2008, 2019; Beck & Steer, 1984, Beck et al., 2005), or attentional explanations of the cocktail effect (Haykin & Chen, 2005; Getzman et al., 2017), and others. Yet, a substantive defense of these cases would require a detailed reconstruction that goes beyond the limits of the present paper. We believe that the two cases analyzed here suffice at least for giving plausibility to our claim that the explanatory autonomy of cognitive psychology is better substantiated in ASE non-causal terms than making it dependent on the final solution of the "causalist problem".

Our proposal of explicating cognitive explanations in terms of the ASE account rather than in causalist terms, solves then "the causalist problem" simply by sidestepping causalism, i.e. by denying the crucial T2 in the argument of Sect. 1: if one provides a non-causalist elucidation of the explanatory import of cognitive explanations, as ASE does, then the whole concern derived from physicalist causal exclusion simply vanishes. We admit that there are other ways of solving the problem by denying other theses or assumptions above, but we think they are far more costly than ours.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work was supported by Ministerio de Ciencia e Innovación (Grant Nos. PID2020-115114GB-I00, CEX2021-001169-M, PID2021-127046NA-I00).

Declarations

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

References

Adams, C., & Dickinson, A. (1981). instrumental responding following reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B, 33*(2b), 109–121.

Anscombe, G. E. M. (1971). *Causality and determination: An inaugural lecture*. Cambridge University Press.

Balzer, W., & Moulines, C. (1981). Die Grundstruktur der klassischen Partikelmechanik und ihre Spezialfsierungen. *Zeitschrift Für Naturforschung A, 36*(6), 600–608.

Balzer, W., Moulines, C. U., & Sneed, J. D. (1987). *An architectonic for science*. Reidel.

Bartelborth, T. (2002). Explanatory unification. *Synthese, 130–1*, 91–107.

Baumgartner, M. (2009). Interventionist causal exclusion and non-reductive physicalism. *International Studies in the Philosophy of Science, 23*(2), 161–178.

Beck, A. T. (2008). The evolution of the cognitive model of depression and its neurobiological correlates. *The American Journal of Psychiatry, 165*(8), 969–977. https://doi.org/10.1176/appi.ajp.2008.08050721

Beck, A. T. (2019). A 60-year evolution of cognitive theory and therapy. *Perspectives on Psychological Science, 14*(1), 16–20. https://doi.org/10.1177/1745691618804187

Beck, A. T., Emery, G., & Greenberg, R. L. (2005). *Anxiety disorders and phobias: A cognitive perspective*. Basic Books/Hachette Book Group.

Beck, A. T., & Steer, R. A. (1984). Internal consistencies of the original and revised Beck depression inventory. *Journal of Clinical Psychology, 40*(6), 1365–1367. https://doi.org/10.1016/j.jpsychires.2021.09.018

Braddon-Mitchell, D., & Jackson, F. (1996). *Philosophy of mind and cognition*. Blackwell.

Brown, A. S. (2003). A review of the déjà vu experience. *Psychological Bulletin*, *129*, 394–413.

Brown, A. S. (2004). *The déjà vu experience*. Psychology Press.

Churchland, P. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy, 78*, 67–90.

Cleary, A. M. (2008). Recognition memory, familiarity, and déjà vu experiences. *Current Directions in Psychological Science, 17*, 353–357.

Cleary, A. M. (2014). On the empirical study of déjà vu: Borrowing methodology from the study of the tip-of-the-tongue phenomenon. In B. L. Schwartz & A. S. Brown's (Eds.), *Tip-of-the-tongue states and related phenomena.* Cambridge University Press.

Cleary, A. M., Brown, A. S., Sawyer, B. D., Nomi, J. S., Ajoku, A. C., & Ryals, A. J. (2012). Familiarity from the configuration of objects in 3-dimensional space and its relation to déjà vu: A virtual reality investigation. *Consciousness and Cognition, 21*, 969.

Cleary, A. M., & Claxton, A. B. (2018). Déjà vu: An illusion of prediction. *Psychological Science, 29*, 635–644. https://doi.org/10.1177/0956797617743018

Cleary, A. M., Huebert, A. M., McNeely-White, K., & Spahr, K. S. (2019). A postdictive bias associated with déjà vu. *Psychonomic Bulletin & Review, 26*, 1433–1439. https://doi.org/10.3758/s13423-019-01578-w

Cleary, A. M., McNeely-White, K. L., Huebert, A. M., & Claxton, A. B. (2018). Déjà vu and the feeling of prediction: An association with familiarity strength. *Memory (special Issue on Déjà Vu)*. https://doi.org/10.1080/09658211.2018.1503686

Cleary, A. M., Ryals, A. J., & Nomi, J. S. (2009). Can déjà vu result from similarity to a prior experience? Support for the similarity hypothesis of déjà vu. *Psychonomic Bulletin & Review, 16*, 1082–1088.

Dickinson, A., & Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior, 22*, 1–18.

Díez, J. A. (2014). Scientific w-explanation as ampliative, specialized embedding: A neo-hempelian account. *Erkenntnis, 79*(8), 1413–1443.

Díez, J. A., & Moulines, C. U. (2022). Guiding principles and special laws. *Theoria (sweeden), 88*(4), 782–798. https://doi.org/10.1111/theo.12393

Farah, M. (2000). *The cognitive neuroscience of vision*. Blackwell.

Fodor, J. (1974). Special sciences or the disunity of science as a working hypothesis. *Synthese, 28*, 97–115.

Forge, J. (2002). Reflections on structuralism and scientific explanation. *Synthese, 130–1*, 109–121.
Frijda, N. H. (2007). *The laws of emotion*. Lawrence Erlbaum.
Getzmann, S., Jasny, J., & Falkenstein, M. (2017). Switching of auditory attention in "cocktail-party" listening: ERP evidence of cueing effects in younger and older adults. *Brain and Cognition, 111*, 1–12.
Glennan, S. (1996). Mechanisms and the nature of causation. *Erkenntnis, 44*(1), 49–71.
Hanson, N. R. (1960). The mathematical power of epicyclical astronomy. *Isis, 51*(2), 150–158.
Haykin, S., & Chen, Z. (2005). The cocktail party problem. *Neural Computation, 17*(9), 1875–1902.
Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. Cambridge: MIT Press.
Horgan, T. (1993). From supervenience to superdupervenience: Meeting the demands of a material world. *Mind, 102*, 555–586.
Kim, J. (1985): Psychophysical laws, en Lepore y McLaughlin (eds.) (1985), pp. 369–86.
Kim, J. (1976). Events as property exemplifications. In M. Brand & D. Walton (Eds.), *Action theory* (pp. 310–326). Reidel.
Kim, J. (1993). *Supervenience and mind*. Cambridge University Press.
Kim, J. (1998). *Mind in a physical world*. MIT Press.
Kitcher, P. (1981). Explanatory unification. *Philosophy of Science, 48*, 507–531.
Kitcher, P. (1993). *The advancement of science*. Oxford University Press.
Kroedel, T., & Schulz, M. (2016). Grounding mental causation. *Synthese, 193*, 1909–1923.
Kuhn, T. S. (1974). Second thoughts on paradigms. In F. Suppe (Ed.), *The structure of scientific theories* (pp. 459–482). University of Illinois Press.
Kusumi, T. (1994). Déjà vu phenomena by analogical reminding. Paper presented at the 11th annual meeting of the Japanese Cognitive Science Society, Tokyo, Japan.
Kusumi, T. (1996). Situational factors of déjà vu experiences: Representational similarities in autobiographical memory and dream. Paper presented at the 7th annual meeting of the Japan Society of Developmental Psychology, Tokyo, Japan.
Kusumi, T. (1998). Déjà vu experiences: An explanation based on similarities of experiences in analogical reminding. Paper presented at the 1st Tsukuba International Conference on Memory, Tsukuba, Japan.
Kusumi, T. (2006). Human metacognition and the déjà vu phenomenon. In K. Fujita & S. Itakura (Eds.), *Diversity of cognition: Evolution, development, domestication, and pathology* (pp. 302–314). Kyoto University Press.
Lewis, D. (1973). Causation. *The Journal of Philosophy, 70*(17), 556–567.
Lewis, D. (1986). Causal explanation. *Philosophical papers II* (pp. 214–240). Oxford University Press.
List, C., & Menzies, P. (2009). Non-reductive physicalism and the limits of the exclusion principle. *The Journal of Philosophy, 106*(9), 475–502.
Lorenzano, P. (2006). Fundamental laws and laws of biology. In N. Karl-Georg (Ed.), *Philosophie der Wissenschaft—Wissenschaft der Philosophie* (pp. 129–155). Mentis.
Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science, 67*, 1–25.
Macrae, C. N., Bodenhausen, G. V., & Calvini, G. (1999). Contexts of cryptomnesia: May the source be with you. *Social Cognition, 17*(3), 273–297.
McCarroll, C. J., & Sant'Anna, A. (2023). Cryptomnesia: A three-factor account. *Synthese, 201*, 23.
McDonnell, N. (2017). Causal exclusion and the limits of proportionality. *Philosophical Studies, 174*, 1459–1474.
Melnyk, A. (2003). *A physicalist manifesto*. Cambridge University Press.
Mitchell, S. D. (1997). Pragmatic laws. *Philosophy of Science, 64*, S468–S479.
Moors, A., Boddez, Y., & De Houwer, J. (2017). The power of goal-directed processes in the causation of emotional and other actions. *Emotion Review, 9*(4), 310–318.
Morris, K. (2011). Subset realization, parthood, and causal overdetermination. *Pacific Philosophical Quarterly, 92*(3), 363–379.
Moulines, C. U. (1978/1984). Cuantificadores existenciales y principios-guía en las teorías físicas. Crítica 10, 59–88. English translation: Existential quantifiers and guiding principles in physical theories. In: J.J.E. Gracia, E. Rabossi, E. Villanueva & M. Dascal (Eds.), *Philosophical analysis in Latin America* pp. (173–198). Dordrecht: Reidel.
Nepe, W. (1983). *The psychology of déjà vu*. Witwatersrand University Press.

O'Connor, A., & Moulin, C. (2008). The persistence of erroneous familiarity in an epileptic male: Challenging perceptual theories of déjà vu activation. *Brain and Cognition*, *68*(2), 144–7. https://doi.org/10.1016/j.bandc.2008.03.007.

Pineda, D., & Vicente, A. (2017). Shoemaker's analysis of realization: A review. *Philosophy and Phenomenological Research, 94*(1), 97–120.

Polger, T., Shapiro, L., & Stern, R. (2018). In defense of interventionist solutions to exclusion. *Studies in History and Philosophy of Science, 68*, 51–57.

Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.

Salmon, W. (1989). Four decades of scientific explanation. In P. Kitcher & W. Salmon (Eds.), *Scientific explanation* (pp. 3–219). University of Minnesota Press.

Salmon, W. (1998). *Causality and explanation*. Oxford University Press.

Shoemaker, S. (2007). *Physical realization*. Oxford: Oxford University Press.

Stenwall, R. (2021). A grounding physicalist solution to the causal exclusion problem. *Synthese, 198*, 11775–11795.

Vicente, A. (2002). The dual 'explanandum' strategy. *Critica, 34*(101), 73–96.

Wallisch, P. (2007, May). *The déjà vu experience as episodic source memory failure*. Paper presented at the Annual Meeting of the Midwestern Psychological Association, Chicago, IL.

Wilson, J. (2011). Non-reductive realization and the powers-based subset strategy. *The Monist, 94*(1), 121–154.

Wilson, J. (2014). No work for a theory of grounding. *Inquiry, 57*, 535–579.

Wood, W., & Neal, D. (2007). A new look at habits and the habit-goal interface. *Psychological Review, 114*, 843–863.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Woodward, J. (2017). Intervening in the exclusion argument. In H. Beebee, C. Hitchcock, & H. Price (Eds.), *Making a difference: Essays on the philosophy of causation* (pp. 251–268). Oxford Univrrsity Press.