Reduction, causation and natural selection

by

Andrés Rubio Krohne

Submitted to the department of Philosophy, Central European University

In partial fulfilment of the requirements for the degree of Master of Arts

Supervisor: Maria Kronfeldner

Vienna, Austria

2025

Copyright Notice

For bibliographic and reference purposes this thesis should be referred to as: Rubio Krohne, Andrés. June 2025. Reduction, causation and natural selection. MA thesis, Department of Philosophy, Central European University, Vienna.

Author's declaration

I, the undersigned, Andrés Rubio Krohne, candidate for the MA degree in philosophy declare herewith that the present thesis titled "Reduction, causation and natural selection" is exclusively my own work, based on my research and only such external information as properly credited in notes and bibliography. I declare that no unidentified and illegitimate use was made of the work of others, and no part of the thesis infringes on any person's or institution's copyright.

I also declare that no part of the thesis has been submitted in this form to any other institution of higher education for an academic degree.

Vienna, 16.06.2025

Andrés Rubio Krohne

Thus, in comparing the eyes of different kinds of animals, we see, in their resemblances and distinctions, one general plan laid down, and that plan varied with the varying exigences to which it is to be applied.

—William Paley, *Natural Theology* (1809, p. 214)

1. Introduction **1**

2. The assumption of physicalism **3**

3. Trouble with causal reductionism **6**

3.1. Defining causal reductionism 6

3.2. Irreducible properties 10

3.2.1. Some plausible examples 10

3.2.2. An implausible example 12

3.2.3. In defense of irreducibility 13

3.3. These properties are causally efficacious 16

3.3.1. Kim and Yablo on causal efficacy 16

3.3.2. An argument for causal efficacy 20

4. Trouble with causal functionalism **26**

4.1. Defining causal functionalism 26

4.2. Two objections 29

4.2.1. The problem of necessitation 29

4.2.2. The problem of over-generation 31

5. Natural selection: a possible way forward **36**

5.1. Two ways of integrating natural selection 37

5.2. The fate of Yablo's theory 39

5.3. Some implications 41

7. Concluding remarks **45**

References **46**

1. Introduction

According to a popular view, mental states are irreducible to the physical (e.g., Putnam, 1967; Davidson, 1970; Block, 1980). What this means, roughly, is that there isn't one single physical state that corresponds to, say, pain, since two beings could both feel pain while differing in important ways in their physical constitution. Even if every individual being that feels pain is physical, pain itself does not correspond to any one physical property or series of properties. This doctrine is very attractive, among other reasons, because it allows us to accept a physicalist world view, according to which every individual thing is physical, while attributing significant independence to the mental from the physical.

But this doctrine also raises some questions. Firstly, if one accepts that properties such as mental states have effects on the world (for example, that pain causes people to tend their wounds), does this entail that there are non-physical causes? If so, how can one reconcile this with a view that the world is nothing over and above the physical? This would amount to the seemingly inconsistent idea that everything is physical, but some things happen because of non-physical causes. A second and related question is how much this extends. Are mental states the only non-physical causes? If not, what are the other ones?

I will discuss two influential views regarding these topics. The first one, which one could call "causal reductionism", claims that, to have effects on the world, a property must be physical or reducible to the physical. Another view, which I call "causal functionalism", says that there are some properties which have effects on the world, but which are "functional" rather than physical. To be functional, in this context, means that these properties are defined by their causal roles. Causal functionalism allows for non-physical causes to exist in a variety of different fields. For example, properties of biology or meteorology could be defined by their causal roles. In this thesis, I discuss problems for both of these views.

I will begin by defining physicalism, the idea that the world is nothing over and above the physical, which I will assume from the start (section 2). Then, I will argue against reductionism. I argue that properties such as pain are not reducible to the physical but have effects on the world. Therefore, not every cause is physical (section 3). The view I call "causal functionalism" agrees with this and claims that these properties should be defined by their causal roles, and not their physical constitution. Nevertheless, I also discuss two potential problems with causal functionalism (section 4). My verdict for causal functionalism is more favorable than for reductionism: it can overcome these issues, but to do so it needs a way of demarcating the functional properties that can be causes and those that cannot.

In discussing these problems, a possible solution will become very apparent. In section 5, I suggest that the problems I mention can be overcome if one uses natural selection to distinguish between properties that have effects on the world and properties that don't. This would mean that some properties that are not reducible to the physical can be causally efficacious if they are, directly or indirectly, a result of natural selection. I suggest two ways in which this intuition could be developed. One of these ways involves reforming functionalism, whereas the other one involves abandoning it. I also mention some implications this approach would have as well as some questions for further research.

These arguments have implications for ontological and epistemological matters. Ontologically, they have implications for what kind of properties count as causes, which is a metaphysical claim. Epistemologically, they have implications for causal explanation, that is, for scientific knowledge that purports to explain a phenomenon by saying what caused it. It is possible that there are other kinds of explanations that are relevant for science, but I will not consider them here. Rather, I will concentrate on why things happen, in the sense of what caused them.

2. The assumption of physicalism

In this thesis, I will take the doctrine of physicalism for granted. Physicalism is, informally, the view that there is nothing over and above the physical. The more precise way of formulating this view is as the idea that every fact supervenes on physical facts. To make this concept clear, I will say, in turn, what I mean by "supervenes" and what I mean by "the physical".

Supervenience is a kind of necessitation: one set of facts supervenes on another set of facts iff the former entails the latter. An equivalent formulation says that every change in the former requires a change in the latter (see Lewis 1983). To say that everything supervenes on the physical, then, means that physical facts entail every other fact. If there was a world which was physically identical to ours, then it would be identical to ours in every respect. If you replicate every atom, every physical law, etc., then you will also have replicated everything, including facts about the mind, economics, meteorology, etc. This also means that, if there is a change in, say, a mental fact, there necessarily will also be a change in some physical fact.

Supervenience, in the way that the term will be used in this thesis, involves metaphysical modality, not only a nomological one (this distinction is to be found, inter alia, in the work of David Lewis; e.g., 1983, p. 36). Nomological necessity is that which could not have been otherwise *given the natural laws that exist*. Therefore, for a proposition[1] to be nomologically possible means that it is compatible with the laws of nature. For example, it is nomologically impossible for me to spontaneously start flying right now, since that would go against gravity. But the laws of nature are, in a sense, contingent: one can conceive of a world where gravity does not exist. One can also conceive of a world in which gravity does not always work and I do start to fly spontaneously. This suggests a different kind of modality,

[1] I talk here of propositions rather than facts because calling something a fact has the connotation of it being true, rather than merely hypothetical.

namely metaphysical modality. A metaphysically possible fact, or set of fact, is a one that does not involve a contradiction. That is, for me to fly spontaneously is nomologically impossible, but metaphysically possible.[2]

Physicalism, then, is the doctrine that physical facts metaphysically entail all other facts. According to physicalism, it is metaphysically impossible for a world to be physically identical to ours while being different in other ways. To claim that this entailment is only nomologically necessary would be too weak for a physicalist. For example, saying that physical facts nomologically entail every other fact does not rule out the possibility that there are immaterial minds, as long as there is a law of nature that always connects them to physical reality. That is, if one claims that there is a contingent law of nature which makes immaterial minds appear every time that there is a human physical body, one would still accept that physical facts nomologically necessitate mental facts. Physicalism is a stronger view. To be a physicalist, one needs to claim that it would be a contradiction for every physical fact to be in place and some mental facts to be different, not just that it would go against the laws of nature.

There is a complication in explaining what is meant by "physics", "physical facts", and similar expressions. Do we mean the current theories favored by physicists? Probably not, since it is very likely that these will be shown to be at least partially false, like many scientific theories favored in the past. Or do we mean some ideal, "finished" science? But this would make our statements about physics meaningless, since we don't know what such a finished science would look like. (This dilemma was introduced by Hempel, 1969). Fortunately, there is a way of circumventing this difficulty. Papineau (2001, pp. 12-13) tackles this issue from the perspective of the philosophy of mind and claims that, for his purposes, it is sufficient to define the physical negatively, as that which is not mental.

[2] There are controversies around the distinction between nomological and metaphysical modalities, an issue I discuss in section 4.2.1.

Physical properties, then, he defines as non-mental properties. Defined in this way, physicalism amounts to the claim that mental properties supervene on non-mental facts (that is, facts that don't make reference to mental properties of entities). Understood in this way, physicalism amounts to the claim that mental properties are not fundamental properties of the world. Since I am concerned with more than just the mental in this thesis, I will define the physical more broadly. By "physical", I will mean anything which is not paradigmatically high-level phenomena. That is, physical means non-biological, non-mental, non-social and non-meteorological.[3] Physicalism, then, is the claim that facts of biology, psychology, etc., are not fundamental features of the world, but rather supervenient on other facts. The fact that something has a biological, psychological and social property supervenes on other facts, which don't themselves involve biological, psychological or social properties.

Some deny physicalism, so conceived. Famously, Chalmers (1996) has claimed that conscious experience is a fundamental feature of the world, not supervenient on anything else (even if nomologically connected to the physical). Nevertheless, in this thesis I will assume it. I am interested in pursuing the conditional question of what would, and what would not, follow from this view, rather than in the truth of the antecedent.

[3] There are probably other paradigmatically high-level phenomena, but for the purposes of this thesis these four will be enough.

3. Trouble with causal reductionism

The view that I call "causal reductionism" claims that everything happens because of fundamental physics. A more precise definition, that I will develop shortly, defines it as the view that every causally efficacious property is reducible to fundamental physical properties. This section describes the view and argues against it. The argument is an old one: it claims that there are some properties which are not reducible to fundamental physics and which are causally efficacious. A classic example of such a property is pain. Pain, it is argued, takes different physical forms, and therefore it is not reducible to the physical. But pain causes things (for example, it causes people to tend their wounds). Therefore, some causes are not reducible to the physical. I will present the motivation for both of these two premises and I will defend them against some recent challenges. The conclusion is that properties like pain are non-physical causes (even if supervenient on the physical). That is, that not everything happens because of physics, fundamental or otherwise.

3.1. Defining causal reductionism

Causal reductionism, in the way in which I define it, is the view that every causally efficacious property is reducible to fundamental physics. This presupposes the idea that there is such a thing as a causally efficacious property, that is, that properties can be causes. Some might not agree with this. Famously, Davidson (1970) claimed that causation is a relation between events, and not properties. If this is true, then we could talk about causally efficacious events, but there would be no distinction between causally efficacious and inefficacious properties. (Here, and in the rest of this thesis, I am using "property" very liberally, to mean any meaningful predicate; that is, anything that can be said, truthfully or falsely, about a particular. In this, I follow Weiskopf's (2011) usage.)

Davidson's claim that causation is a relation among individual events has nevertheless fallen out of favor. As argued by Kim (2000, p. 10), Davidson's view does not say anything about what are the properties of the causing event *in virtue of which* the effect was caused. Suppose that I drink red wine and get drunk. We could say, in this case, that an event causes another event. But the causing event could be described in different ways: as *I drink an alcoholic beverage*, for example, or as *I drink a red beverage*. The first of these descriptions captures a fact about the beverage which made me drunk, namely the fact that it is alcoholic. Of course, a complete description should name a number of other factors that contributed to me getting drunk, including properties of my body. The point is that the alcoholicness of the beverage is one of these relevant properties, whereas its color is not. It would have made no difference if the wine was white, for example. Following a common usage (Kim, 2000; Menzies & List, 2010), I call properties "causally efficacious" when they are one of the properties by virtue of which the effect took place, and I call them "causally inefficacious" when they are not. I follow people like Yablo (1992), Kim (2000) and Menzies and List (2010), which in different ways have said that causation is not just a relation between events, but between events under descriptions. These descriptions will attribute properties to the particulars involved in the event. In the rest of this text, when I use the word "event", or the variable e_x, what I will have in mind is an event under a description. This description will involve some of the properties of the particulars involved in the event, but not all. Later, I will use the uppercase E_x to refer to types of events under descriptions (such that the description applies to all events of that type), rather than individual ones.

Causal reductionism ("reductionism", for short) claims that the descriptions of causing events can always be put in terms of fundamental properties. That is, for every event e_1 which causes an event e_2, e_1 could be put under a description that only includes fundamental properties, which would then capture all the properties which were causally

efficacious in its causing e_2. Since I am assuming that high-level properties, such as biological and psychological ones, are not fundamental, reductionism entails that every causing event could be put under a description that does not include any biological or psychological property. We could describe every cause in terms of things like atoms, charge, mass, etc., without making any reference to organisms or mental states, and that description would include all causally efficacious properties.

What about, say, explanations in astronomy, which make reference to entities like stars? Surely, the property of being a star does not seem to be fundamental. Nevertheless, a star is arguably something which can be defined in terms of fundamental properties. I will assume that defining a concept consists in giving a series of conditions that are metaphysically necessary and sufficient for it (similar conceptions are used by Strevens, 2012; Yablo 1992). That is, to define the property "being a star" in physical terms is to provide a series of physical properties that are metaphysically necessary and sufficient for something to be a star. Roughly, one could define a star as an object with self-gravity made out of plasma.[4] In a sense, one could say that the property of being a star is the same as the property of being an object with self-gravity made of plasma, since it would be contradictory for a being to have one without the other (since the modality involved is a metaphysical one). That is, a description that refers to the property of being a star and a description that instead refers to the property of being a plasmatic object with self-gravity are, in an important sense, the same description. One should not say that there are two causes of the sun shining: on the one hand, that it is a star and, on the other hand, that it is a plasmatic object with self-gravity. Rather, these are one and the same cause.[5]

[4] Note that definitions, in this sense, could be a posteriori (one might not know that stars are made out of plasma) and that there might be different and equivalent definitions of the same property.
[5] I will assume this conception of what it is to define a property. If one disagrees with that conception, one could potentially find ways of resisting my argument, although it is not straightforward how these would look like.

It seems that one could take the properties that figure in the definition of "star", such as "self-gravity" and "plasma", and define them in ways that are closer to fundamental physics. After doing this a number of times, one could then produce a definition of "star" that would only refer to fundamental properties. That is, in principle (although probably not in practice), one could produce an equivalent definition of "star" that only refers to fundamental properties. Some might disagree with this example, but so far I use it merely as an illustration. That fact is that, even if a property is not fundamental, it could still be in principle definable referring exclusively to fundamental physical properties. I call such a property "reducible".

Suppose, then, that to be a star is the same as having a long list of fundamental properties. In that case, an event under a description like *A star rotates* would be equivalent to an event under a much longer description using only fundamental properties. In this case, to say that an effect took place because a star rotated would be the same as saying that it took place because a series of entities had a series of fundamental properties. The property of being a star could be replaced by other properties that say nothing about starhood. That is, to say that the property of being a star was causally efficacious is the same as saying that a number of physical properties, together, were causally efficacious.

Reductionism, then, is the view that every causally efficacious property is reducible to fundamental properties. Every property that matters in terms of causation is fundamental or definable in fundamental terms. If an event is a cause, then it is under a description that could be put in terms of fundamental physics. Usually, descriptions of causes are not in terms of fundamental physics, but the physicalist sees these apparently non-physical explanations as shorthand ways of stating longer physical explanations. If this view is correct, then no property from high-level sciences is indispensable for describing causing events. They could

all be replaced by a series of fundamental properties. Examples of reductionists are Kim (2000) and Strevens (2012).

In the next two sections, I will present the motivation for claiming that, even if physicalism is true, reductionism is not. The argument attempts to present a counterexample: a property that is not reducible to the physical, but which is causally efficacious.

3.2. Irreducible properties

3.2.1. Some plausible examples

There have been various properties which have been claimed to be "irreducible" (my usage of this term follows Strevens, 2012). An irreducible property is one that is not fundamental and not definable in fundamental terms.[6] To say that a biological property, for example, is irreducible, means that it supervenes on the physical, but it cannot be defined in physical terms. Since reductionism claims that all causally efficacious properties are fundamental or reducible to fundamental properties, it is committed to the claim that no irreducible property is causally efficacious. In other words, the question of whether reductionism is correct is the question of whether there are any irreducible properties which are also causally efficacious.

The irreducibility of high-level properties could seem incompatible with physicalism, but in fact it is not. There is no contradiction in claiming that high-level properties supervene on the physical while claiming that they cannot be defined in physical terms. Every instance of the irreducible property is a physical entity, but what they have in common, by virtue of which they share the same property, is not a series of physical properties. This is closely related to the doctrine of "token-token" identity theory (heavily inspired by Davidson, 1970): the idea that every particular in the universe, every token, is a physical particular, but not every type is identical to a physical type.

[6] Given how I have defined the terms, there are three kinds of properties: fundamental, reducible and irreducible.

A way to establish that a property is irreducible is by showing that it is "multiply realized" (a concept introduced by Putnam, 1967). If all the instances of a property don't have any physical property in common which is exclusive of them, then the property is multiply realized.[7] This can be illustrated through Putnam's influential example of pain. Presumably, both humans and octopuses can experience pain. But their nervous systems are very different. One can then have the intuition that there is no set of physical properties which are unique to pain (i.e. sufficient) but also apply to every kind of pain, including human and octopus pain (i.e. necessary). Pain is the case I will concentrate on in this thesis.

If a property is multiply realized, then it is irreducible. If there is no set of physical properties that are shared by every instance of pain and nothing else, then pain cannot be given a physical definition. That is, if Putnam's intuition that pain is multiply realized is correct, then pain is irreducible. A similar claim was made by Weiskopf (2011), who claimed that visual perception cannot be defined in physical terms in such a way that it applies for both crabs and humans. Later, in section 3.2.3, I will defend the intuition that such properties are multiply realized.

Putative examples of multiple realization often involve different species. But there are also properties that, plausibly, take different physical forms even in the same species. A possible example of this are propositional attitudes like beliefs. Two people might share the belief that it will rain today while, seemingly, having very different neuronal states. The exact realization of this belief might differ depending on factors such as how they came to believe that it would rain, how much attention they paid to this fact, what their emotional reaction to this was, etc. The neural pathway through which this belief is formed, stored and retrieved need not be similar in both cases. It is, however, the same belief. Since the case of belief has

[7] I am formulating multiple realizability in a somewhat different way than Putnam, which is based on the previous discussion and on Strevens (2011).

many additional complications, such as the problem of content, I will not concentrate on it, but it is also a seemingly plausible example of multiple realization.

Since in this thesis I will suggest a relation between irreducible properties and natural selection, it is a good sign that plausible examples such as pain, perception and belief come from biology and psychology. There are, nevertheless, other putative examples that have been put forward.

3.2.2. An implausible example

Some philosophers have presented other putative examples of irreducible properties which have nothing to do with biology or psychology. Batterman (2000), for example, has used temperature. But, given the definition of irreducibility and multiple realization that I have presented here, there is no reason to think that this property is multiply realized.

Batterman's reasoning to claim that temperature is multiply realized is the fact that "most of the details of the microstructure (...) are (...) irrelevant" (2000, 119). The temperature of an object does not depend on the details of its microphysical structure. This is, of course, true: an object can have a temperature of, say, 15 degree Celsius, independently of its mass, its state, the number of electrons to protons in it, the exact position of every atom, etc. Two very different objects could have the same temperature. But this does not mean that temperature is not definable in physical terms. Temperature can roughly be defined as mean kinetic energy, which is a definition that refers to physical properties, and which arguably could be put in physical terms. To say that two objects have the same temperature is to say that they share a number of physical properties. Of course, there could be other physical properties that they do not share. Two objects with the same temperature will share some physical properties and not others, but the ones that they do share are necessary and sufficient for the property, i.e., they can constitute a definition of that property. This point can be seen with the case of mass. Batterman admits that mass is a physical property (2000, p. 120, p.

131); nevertheless, two objects can have the same mass while differing in almost all of their physical details. For example, a horse can have the same mass as a certain amount of water. Despite their differences, these beings share a property that can be defined in physical terms (a similar point is made in Strevens, 2012, 754-760).

Since we don't know for sure what the fundamental building blocks of the universe are, we cannot say for sure that properties such as temperature are not multiply realized. For all we know, a certain temperature may take completely different forms on the fundamental level. But, based on what we know, there is no reason to suppose that. The fact mentioned by Batterman, namely that there are differences on the fundamental level between things with the same temperature, does not justify the claim that this property is multiply realized in the sense that I am assuming.

3.2.3. In defense of irreducibility

So far, I have explained what it means for a property to be irreducible and presented some plausible examples. But one might doubt whether these properties really are irreducible, instead of mere plausible candidates for irreducibility. In this section, I will start by responding to some considerations that could suggest that these properties are reducible. After responding to these worries, I will present a positive argument for the conclusion that they are, in fact, irreducible.

Recently, there has been some skepticism about multiple realization. One main critic has been Cao (2022), who responds to philosophers that have adamantly claimed that our brains could be replaced by a silicon replica without affecting the way in which it functions. As she argues, not any physical substratum can perform any function. The more complex and integrated the distinct parts of a system are, the more constraints there are on their realizers. Neurons perform functions that are remarkably complex and integrated, which means that these functions can only be performed by very specific realizers. As far as we know, it is

possible that “the specific chemical and biological properties of brain tissue” are necessary for the realization of the mental states exhibited by human beings (Cao, 2022, p. 25).

Cao effectively puts pressure on some radical claims involving multiple realization. Claims that we could produce a silicon brain that is functionally identical to ours seems less plausible after Cao’s argument. Like she claims, it is to be expected that the various realizers of a mental predicate have some physical properties in common. But the fact that they have some physical properties in common does not entail that they are definable in physical terms. The constraints on the realizers only show that states such as pain have some necessary physical conditions (such as brain tissue), but those constraints might still not be unique to the mental predicate (brain tissue, for example, is not unique to any mental state). That is, Cao has shown that there are necessary physical properties, but this does not mean that they are necessary and sufficient, which they would have to be to constitute a definition.

For example, Cao (2022, p. 24) mentions five physical similarities between human and octopus pain: (1) peripheral touch receptors, (2) the use of chemical neurotransmitters, (3) sensory areas distinct from motor areas, (4) a neuron-based brain and (5) endogenous opioids. Even if we were to grant that all these properties are strictly necessary for any form of pain, this would not mean that pain can be defined only with them. All of these properties apply, for example, to perception-based pleasure (e.g., sexual pleasure). That is, these physical constraints are so generic that they do not allow us to characterize the difference between pleasure and pain. For these constraints to refute the irreducibility of pain, they would have to be sufficiently general to apply to both human and octopus pain, but also sufficiently specific to not apply to human or octopus pleasure. It is one thing to claim that there are physical constraints to the realization of pain, and a different one to claim that there is nothing to pain beyond those constraints.

Couldn't one define pain in terms of nociceptive receptors (that is, receptors that detect harm) firing? This is usually specific of pain and not pleasure. One problem with this idea is that nociceptive receptors might also be multiply realized. But, even if they are not, nociception is not sufficient for pain. For example, patients under general anesthesia have nociceptive responses but feel no pain (Subramanian et al., 2024). Pain cannot be defined just by nociception, without taking into account how those stimuli are processed.

I have, then, argued that physical constraints do not refute irreducibility, since they do not show that the properties are definable in physical terms. Constraints are necessary, but they might not be sufficient. This is only a negative argument: the existence of constraints does not entail that these properties are reducible. I have not yet shown that they are, in fact, irreducible. For all we know, properties like pain could be analogous to temperature: their instances might differ in significant ways, while retaining enough physical properties for a definition.

Some, like Putnam (1967), have accepted multiple realization merely by intuition. But I believe that there are two additional considerations that could motivate the claim that properties like pain are multiply realized and hence irreducible. The first reason is that a physical definition of pain would have to be simultaneously very fine-grained and very coarse-grained. For example, it cannot make reference to the thalamus, as octopuses don't have one. But it also has to be fine-grained enough to capture how pain is processed. In the case of human pain, the thalamus plays a significant role in discriminating, modulating and relaying nociceptive stimuli (Ab Aziz & Ahmad, 2006; Galdino et al., 2024). In fact, damage to the thalamus can make people cease feeling pain (Shantanna, 2018). So it would seem that, for the definition to capture the way in which pain is processed in humans, it has to mention the thalamus. But then it won't apply to octopuses. It seems, then, that we have a dilemma. If the physical definition of pain mentions the thalamus, it won't apply to octopuses. But if it

does not, it will not capture the processing that makes pain different from mere nociception in humans. Either it is too coarse-grained or too fine-grained.

Another reason to believe that pain is multiply realized arises from the idea of natural selection. Humans and octopuses both feel pain because they were selected to do so. That is, their ancestors developed behaviors that proved evolutionarily beneficial, and therefore proliferated. What matters about pain, what makes it appear in the evolutionary process, are its results: that it makes organisms behave in ways that avoid future harm, for example. Any neural state that achieves this will be favored by natural selection. If what matters are the results, namely the behavior, and not the means, namely the physical realizers, then there is no reason to expect the means to be always the same.

These are not knock-down arguments that can definitively prove that properties like pain are irreducible. It is possible that there is a way of defining pain that applies to all species that feel it and makes no reference to anything as specific as the thalamus. There is no apparent way of ruling this out completely. Nevertheless, I believe that it would be a much riskier bet to assume that these properties are definable in physical terms. For that reason, I will assume that pain is irreducible.

3.3. These properties are causally efficacious

Even if one accepts that properties like feeling pain are irreducible, this is not an automatic defeat for the reductionist. One still needs to show that these properties are causally efficacious.

3.3.1. Kim and Yablo on causal efficacy

Jaegwon Kim (2000) argues that irreducible properties are not causally efficacious. His argument could be paraphrased as follows. Suppose that I feel pain and tend my wound. The event of my being in pain could be described in a way that only refers to physical properties,

for example, by saying that I had such-and-such brain state (assuming that specific brain states are definable in physical terms). If pain is irreducible, these two descriptions are not equivalent. The description that mentions pain attributes a property to me that I could share with an octopus, whereas the physical description attributes properties that I could not. So these are two different events-under-descriptions. Kim argued that a causal explanation based on the physical description could explain why I tended my wound. Saying that I had such-and-such brain state is sufficient, according to Kim, to explain why I tended my wound. There is, then, no work left for the other event, the one that is described in terms of pain, to do. If one claims that pain is causally efficacious, one would have to say that there are two events-under-descriptions which cause my wound-tending: one of them mentioning pain and another one using only physical terms. That is, my wound-tending would be overdetermined: there would be various causes for it, each of them enough to make me tend my wound.[8] Since this is unlikely, Kim claims that only the event under the physical description is causally efficacious.

If one accepts Kim's argument, one can still claim that *human* pain is causally efficacious, as long as human pain, unlike pain simpliciter, is reducible to the physical. That is, one could say that the cause od wound/tending in octopuses is octopus pain and the cause of wound/tending in humans is human pain. My wound-tending and the octopus' wound-tending have, under this view, different causes. To say that I tended my wound because I was under human pain is equivalent to saying that I tended my wound because I had a set of physical properties. The reductionist can claim, based on Kim's argument, that pain simpliciter is causally inefficacious, whereas human pain or octopus pain are each reducible and therefore causally efficacious.

[8] Note that the problem here is not that there are two causes. Effects often have various causes (for example, drunkenness can be caused by dehydration combined with alcohol intake). The problem is that at least one of these causes is sufficient, without the other, to produce the effect. The problematic implication is that, in every case of pain-related wound-tending, we would have more causes than necessary.

There are ways of resisting Kim's argument. An influential one consists in accepting a counterfactual approach to causation. The spirit of counterfactual approaches to causation has been summarized by David Lewis: "We think of a cause as something that makes a difference, and the difference it makes must be a difference from what *would have happened* without it. Had it been absent, its effects – some of them, at least, and usually all – would have been absent as well" (1986, p. 161, my emphasis). Counterfactual approaches to causation define causes in terms of counterfactuals: had a cause been absent, the effect would not have taken place.

I will follow a particular counterfactual theory of causation, namely Yablo's (1992). Yablo's counterfactual theory is based on the metaphysical concept of determinates and determinables. An event e_1 is another event e_2's determinate iff e_1 is a way in which e_2 could take place, that is, a more specific version of e_2. This makes e_2 into e_1's determinable, that is, into a less specific version of e_1. Drinking wine, for example, is a determinate of drinking alcohol. Drinking wine is of course not the only way to drink alcohol, since ine can also drink beer. That is, drinking wine, the determinate, entails drinking alcohol, the determinable. Determinables are events under more general descriptions than determinates.[9]

Yablo uses these concepts to distinguish between causally efficacious and inefficacious properties. According to him, a causally efficacious property must fulfill two conditions: it must be *enough* and it must be *required* for the effect. Yablo uses these words as technical terms. To be *enough* for an effect means that any of its determinables would have resulted in the effect taking place. For example, my drinking alcohol is enough for my being drunk, because I will get drunk regardless of what kind of alcohol I drink. Any determinate of the event *I drink alcohol* would produce the same effect (provided the same background conditions). The property also has to be *required* for the effect: without it, and all other things

[9] Yablo puts it in terms of the essences of events, but we need not follow that metaphysically loaded language. I follow Woodward (2010) in representing Yablo's theory in a more metaphysically neutral way.

equal, the result *would not have taken place*. Had I drank juice rather than alcohol, I would not have gotten drunk. Since this approach defines causes in terms of what would have happened without them, it is a counterfactual account.

For Yablo, properties can fail to be causally efficacious by being not enough or not required for the effect. Drinking white wine, for example, is not required, as I could have gotten drunk with a different determinate of drinking alcohol. Saying that I got drunk because I was drinking a liquid, on the other hand, is false, because drinking liquid is not enough for the effect. By contrast, saying that I got drunk because I was drinking an alcoholic beverage is at the right level of specificity. If a property is enough and required for an effect, then it is proportional to it, and hence causally efficacious.[10]

With this counterfactual account of causation, Yablo is able to avoid the overdetermination problem raised by Kim. It is not the case that there are two distinct events that independently caused my wound-tending, one of them mentioning pain and another one mentioning only physical properties. The event that is described using pain causes of my wound-tending, since it is enough (any determinate of pain would have produced wound-tending) and required (had I not felt pain, I would not have tended my wound). There can be an event that is described through specifically human pain, which I am assuming to be reducible to the physical, but that event would not cause my wound-tending. This physical description is enough for my wound-tending, but it is not required: had I felt a different kind of pain, I would still have tended my wound. That is, if I had had a different, more octopus-like nervous system, I would not have felt human pain, but I would still have tended my wound. For this reason, the event that mentions only human pain is a determinate of the

[10] Yablo's views on causation should not be confused with my description of reducibility and definability, even if both are based, in a sense, on necessary and sufficient conditions. I am conceiving of definition as a kind of constitution: being a plasmatic object with self-gravity constitutes being a star. Causation is different: drinking causes me to be drunk, but it does not constitute my being drunk. This is because definition is a relation that is metaphysically necessary, whereas causation seemingly is not (although more on this in section 4.2.2). It is also worth noting that causation involves two events, such that one of them taking place is enough and required for the other to take place. Definition, by contrast, involves just one particular, such that one set of properties of this particular is necessary and sufficient for it having another property.

real cause, which is the event that mentions pain simpliciter. Yablo's account is not the only way of resisting Kim's argument, but I will proceed under the assumption that it is the best one.

If one accepts Yablo's views on causation, one can avoid the problem of overdetermination and reject reductionism. Nevertheless, if one does not accept Yablo's account, one can still hold on to reductionism. In the following section, I want to motivate accepting Yablo's views rather than Kim's.

3.3.2. An argument for causal efficacy

A problem one might have with reductionism is that some regularities in nature are described and explained through irreducible properties. For example, the similarity between the behavior of humans and octopuses can be explained by saying that they both feel pain. Here, b "explanation", I mean causal explanation: one could say that there is a common cause for wound-tending in humans and octopuses, namely pain simpliciter. But, if Kim's views are correct, then human and octopus wound-tending don't have a common cause. Under reductionism, these regularities seem to be merely coincidental (a similar intuition is expressed by Kitcher, 2011, pp. 70-71, although in very different terms). The argument can be stated as a modus tollens: if reductionism is correct, then these regularities are coincidental; these regularities are not coincidental; therefore, reductionism is not correct.

To see what I mean by a coincidence, take two types of events that are correlated, E_1 and E_2. For example, suppose 90% of people who read a particular book become rich in the next year. Not only that, but they become rich in a way that depends on them reading the book. That is to say, had they not read the book, they would not have become rich. Even in this case, it is possible that this regularity is a mere coincidence. This would happen if events of type E_1 caused events of type E_2 for independent and unrelated reasons that are not directly dependent on them belonging to those types. Imagine that people become rich after reading

that book, but they do so in different ways. One person buys a used copy and finds, on the last page, a check which the previous owner left by mistake. Another person decides to use the number of pages of that book as a lottery number and wins. A third one decides to go read it at a café where she happens to meet a billionaire, who then later invests in her company. And so on. In this case, it remains true that 90% of people who read the book became rich, and it remains true that, had they not bought the book, they would not have become rich. Nevertheless, the causes of them becoming rich are different. For the person who met a billionaire, for example, the fact that she read the book at that specific café is what caused the effect. But, in that case, it does not matter what copy of the book she read. For the person who found the check, it doesn't matter where she read the book, but it does matter what copy. That is, the details matter. These details are the causes of each person getting rich, and the fact that they all did that because of reading the same book is a mere coincidence. That is, all these events start in the same way (with book-reading) and end in the same way (with wealth), but what causes one of these things to lead to the other is different in each case.

Now consider a regularity involving pain. One such regularity is the fact that being with complex nervous systems (that is to say, beings with centralized nervous systems that have different types of neurons) often tend their wounds when they are hurt. Events of type E_1, namely these organisms getting hurt, are correlated with events of type E_1, namely the organisms tending their wounds. Unlike the correlation between reading the book and getting rich, this regularity does not seem to be a mere coincidence. To see why, consider two key differences between coincidental and non-coincidental regularities. The first difference is their likelihood. There is nothing intrinsically unlikely in the fact that different beings with complex nervous systems tend their wounds when hurt. By contrast, the book case is very unlikely, because it just happens to coincide. The second difference concerns inductive inferences. In the first case, one can justifiably infer that other beings with complex nervous

systems will tend their wounds when hurt. This inference might not be infallible, but it is reasonable. But, in the second case, one should not make an analogous inference. If the people who read the book became rich in different and unrelated ways, there is no reason why people who read the book in the future will continue to get rich.

The book case and the pain case are different. The issue is why. What makes one case coincidental but not the other? One way of characterising the difference would be the following. In the case of pain, different organisms react in the same way because of a common property, namely pain. The regularity is explained not by human or octopus pain, but by pain simpliciter, which is present in all cases and causes all these beings to tend their wounds. In the book case, there is no such common property. In the case of the book, we have one initial event, which is the book being read, and one end result, which is wealth, but diverse causes in the middle. In the case of pain, we have one initial event, namely bodily harm, and one end result, namely wound-tending, and the same cause in the middle, namely pain. That is seemingly the reason why one case is coincidental and the other is not.

Although this response seems plausible, the reductionist cannot accept it, since it would involve recognizing pain as a causally efficacious property. It would seem that the reductionist would have to regard the case of pain as analogous to the case of the book: the same result achieved by different causes. In one case, it was human pain that caused the organism to tend their wounds, in the other it was octopus pain.[11] The reductionist regards both cases as analogous, in the sense that they both involve different properties that happen to produce the same result. Because of this, it is not clear how the reductionist can distinguish these cases. If one adopts Yablo's account instead, the can be elegantly explained: human pain, octopus pain, etc., are mere determinates of the causally efficacious property of pain. In

[11] Remember that I am assuming that human pain, octopus pain, etc., are each reducible to the physical, but not pain simpliciter.

the case of pain, there is one property in common to all cases, whereas in the book case there is not.

One response that the reductionist could give is to appeal to the physical constraints mentioned by Cao. The nervous systems of humans and octopuses have various properties in common, such as peripheral touch receptors and chemical neurotransmitters. These properties are present for both humans and octopuses, so, the response goes, perhaps that can help explain why they all reacted in similar ways to bodily harm. The problem is that, as mentioned, these properties are too general. It is perfectly possible to have all of these properties and to not react to bodily harm with wound-tending. People under general anesthesia and some people with damage in their thalamuses, for example, also have peripheral touch receptors and chemical neurotransmitters, but they don't react to bodily harm with wound-tending. The fact that all these cases have some properties in common does not mean that the similar effects are not coincidental, because the common properties are not enough to cause the effect. Similarly, finding out that there were some similarities in all the cases of the people who got rich after reading the book does not mean that this case is not coincidental, because in every case there were some indispensable elements that are absent in other cases, such as the check. In the case of human pain, there are some indispensable elements that are missing in the case of the octopus, such as the thalamus.

A second strategy that the reductionist might use is to attribute a disjunctive property to the organisms. One could define pain as the property of having either human pain, or octopus pain, or dog pain, etc. In that way, one could attribute the same disjunctive property to all instances of pain and claim that that property was the cause of these organisms' wound-tending. The problem with this is that it is very unlikely that such disjunctive properties are causally efficacious. For example, in the book case, one could say that all people had the common property of either finding a check on the book, or meeting a

billionaire, etc., but this does not mean that there is a common cause in all cases. If one wants to explain the regularity through a common cause, it cannot be a disjunctive property (a very similar point is made by Fodor, 1974).

A third strategy that the reductionist could use is to appeal to natural selection. Humans and octopuses don't behave in the same ways because they have a property in common, the reductionist could say, but because behaving in these ways promotes survival. The organisms that did not behave in those ways are not around anymore, since they died and did not leave any offspring. There is some truth to this response: natural selection is indeed responsible for making humans and octopuses behave similarly despite having different nervous systems. But it's not clear that this can save the reductionist, as it only pushes the question back. Did natural selection *cause* humans and octopuses to behave in the same ways? It does not seem like the reductionist can make this move, as it would involve claiming that natural selection is a causally efficacious property.[12] There is hardly a more plausible example of a multiply realizable property than natural selection. To put it simplistically, any environment in which a self-replicating organism can appear, and in which there is evolutionary pressure, is apt for natural selection. That is, natural selection is defined in terms of non-fundamental properties, such as organisms and survival, and most likely not in physical terms. It would be strange to accept that pain is irreducible but to deny that natural selection is. Therefore, it seems that the reductionist cannot point to natural selection as a cause without committing to the causal efficacy of an irreducible property.

The reductionist could claim that natural selection provides a non-causal explanation of why humans and octopuses behave in similar ways. This would seemingly commit her to the claim that the regularities involving pain don't have a causal explanation, since focusing only on physical properties would lead to explaining the case of the human and of the octopus

[12] One might ask: what is natural selection a property of? It seems that it would be a property of the evolutionary history of these animals, that is, a property of events.

through different causes. If the reductionist takes this route and gives an account of what kind of non-causal explanation this is, I concede that she can successfully resist my argument. But I will assume that one would rather accept irreducible causes than accept that some empirical facts, like the fact that humans and octopuses non-coincidentally react in similar ways to bodily harm, are explained non-causally.

Please note that, although I have mentioned regularities and causation, I am not committing myself to a regularity-based theory of causation (such as that of Hume). I am not saying that causation is constituted by regularities. On the contrary: as I have claimed, it is possible for regularities to exist without causation. What I am saying is that, in some cases, regularities are explained, not constituted, by causation. Other regularities are coincidental, and therefore do not involve causation. But, in the case of pain, it is not plausible to assume that the regularities are coincidental. Instead, they should be explained by causation.

The problem for reductionism arises out of this combination of multiple realizability and causal efficacy. Properties like feeling pain cannot be defined in physical terms, since there is no conjunction of physical properties that is necessary and sufficient for them. They are irreducible. Additionally, the fact that they appear in non-coincidental causal generalizations means that they are also causally efficacious. That is to say, pain is an example of an irreducible and causally efficacious property. As Yablo's account of causation shows, this can be accepted without committing to overdetermination. Following this account, pain is the cause of my wound-tending because of two reasons: had I not felt pain, I would not have tended my wound, and had I felt a different kind of pain, I would still have tended my wound. I take these to be good reasons to accept Yablo's account and the causal efficacy of irreducible properties such as pain.

4. Trouble with causal functionalism

The conclusion of the previous chapter is that some causally efficacious properties cannot be defined in physical terms. This raises the question of in what terms they should be defined. In this chapter I will consider an influential answer to that question, called "functionalism", which defines properties in terms of their causal roles. In this section, I present two difficulties for this approach. I also mention what it would take for functionalism to overcome them.

4.1. Defining causal functionalism

I will start with a terminological issue. The word "function" is often used in a teleological way, as roughly synonymous with "purpose". For example, the *function* of having eyes is to allow the organism to gain information from the environment. Some of the authors that I will discuss in section 5 use the term in this way. But, when I talk about "function," "functionalism," and "functional properties," I will not have this teleological concept in mind. Rather, I will use the term in a way that it is often used in the philosophy of mind, that is, in connection with a tradition that attempts to account for the mental in terms of causal roles (exemplified by Putnam 1967 and Block 1980). In this tradition, the word "function" does not refer to a purpose, but rather to the idea of a *mathematical* function which connects inputs and outputs. In what follows, please do not interpret any usage of "function", "functional", etc., in the teleological sense.

"Function," then, is inspired by the idea of a mathematical function: an algorithm that expresses the algorithmic relationship between two values, an input and an output. This concept has been used by some philosophers of mind, who define mental states by the way in which they connect two values. The input value usually is taken to involve perceptual stimuli and the output value is taken to involve behavior. Both input and output can nevertheless

involve other things as well, depending on the mental state to be defined. Levin (2023, §1) illustrates this view thus:

> For (an avowedly simplistic) example, a functionalist theory might characterize pain as the state that tends to be caused by bodily injury, to produce the belief that something is wrong with the body and the desire to be out of that state, to produce anxiety, and, in the absence of any stronger, conflicting desires, to cause wincing or moaning.

In this example, pain is defined functionally, that is, by its inputs and outputs. Its inputs are what causes it and its outputs are what it causes. Any state that tends to be caused by the same things as pain and which tends to cause the same things counts as pain, independently of its physical constitution. As Levin (2023) mentions, there are different kinds of functionalism, such as analytic or machine-state functionalism, but we need not be concerned with these fine-grained distinctions here.

Two things should be noted about functionalism. Firstly, the operator "tends to be" could be interpreted as "in normal circumstances". In normal circumstances, there is no pain unless there is some kind of injury, that is, pain is normally caused by an injury. But it is of course possible for pain to exist without an injury, in abnormal circumstances. It is also possible, in abnormal circumstances, that pain does not have the effects that it usually has. Defining what these normal circumstances are could prove difficult for the functionalist, but for our purposes there is no reason to get into it. Secondly, the effects of pain are probably much more numerous than what Levin's quote could suggest, and a complete functional definition of pain would include things like wound tending, a future disposition to avoid the source of damage, etc. The effects of such states can, and usually do, involve other functional states as well as dispositions.

One main motivation for functionalism is the fact that it allows for multiple realization. If pain is whatever state that has certain causes and effects, then pain doesn't need to be definable in physical terms. Any organism whose body works in the appropriate way, responding in the right ways to the right stimuli, can count as feeling pain, independently of whether it resembles a human or an octopus or something else. Defining properties in terms of their functional properties — that is to say, in terms of their causes and effects — is different from defining them in physical terms. Therefore, functionalists can attribute the same mental state to humans and octopuses, and need not worry about the differences in their brains. Similarly, visual perception can be defined functionally, in a way that allows for the different kinds of eyes and visual systems that exist (Weiskopf 2011).

If functionalism is merely taken to be the view that predicates such as "feeling pain" *can* be defined functionally, then I don't necessarily reject functionalism. I am not sure that there is one correct way of defining these terms, since it is possible that both lay people and experts use them in different ways. What I will argue against is the claim that functional properties are causally efficacious, which I call "causal functionalism". According to causal functionalism, the functional properties of pain cause me to tend my wound. Put differently, causal functionalism claims that pain causes me to tend my wound by virtue of its functional properties. One can be a functionalist without being a causal functionalist, and in this chapter I discuss causal functionalism rather than functionalism in general.

Causal functionalism is very much compatible with Yablo's views on causation. This is because functional properties are proportional to their purported effects. Having functional properties is enough to make me tend my wound, since making me tend my wound is one of the definitional effects of pain. It is also required: if I hadn't had these functional properties, I would not have tended my wound. Therefore, Yablo's counterfactual theory can justify causal functionalism.

One apparent implication of causal functionalism is that all kinds of properties, belonging to all kinds of sciences, can in principle be causally efficacious. There is no apparent reason why one couldn't define chemical or meteorological properties in functional terms. If causal functionalism is correct, then any kind of phenomenon could involve non-physical, causally efficacious properties.

Causal functionalism is able to avoid the problems of reductionism by postulating non-physical causes. The functionalist can say that humans and octopuses behave in similar ways because their behaviors had the same cause, namely pain (functionally defined). It is also compatible with Yablo's insights into causation. Nevertheless, in this section I mention two complications it has.

4.2. Two objections

In this section, I will present two objections to causal functionalism. The first one, the problem of necessitation, has been put forward by some critics of causal functionalism. I will present it and argue that, although it is plausible, it is inconclusive. Then I will present a second objection inspired by, but different from, the problem of necessitation, which I call the "problem of over-generation". I will argue that, at best, this problem calls for a reform or development to causal functionalism and, at worst, for abandoning it.

4.2.1. The problem of necessitation

One reason to reject the idea that functional properties are causally efficacious is what Rupert (2006) has called the "problem of metaphysically necessitated effects" (for short, the "problem of necessitation"). This argument, which various philosophers have considered,[13] could be put as a syllogism: causes are metaphysically independent from their effects;

[13] E.g.: Shoemaker (1997), Antony and Levine (1997) and Kim (2000).

functional properties are not metaphysically independent from their purported causes; therefore, functional properties are not causes.

In paradigmatic cases of causation, it seems that it is metaphysically possible for the causing event to occur without the effect occurring. Drinking wine, for example, can cause me to be drunk, but there is no contradiction in my drinking wine without getting drunk. But now suppose that, as suggested above, pain is characterized as a property that is caused by bodily harm and that, in normal circumstances, causes things like wound-tending. So conceived, it is by definition impossible for pain to occur in normal circumstances and wound-tending not to occur. A world where pain does not cause wound-tending under the normal circumstances is not possible, as causing wound-tending under normal circumstances is part of the definition of "pain". This means that, if pain is characterized in terms of its causal role, it is not metaphysically independent of its effects.

Therefore, the argument goes, since pain (functionally characterized) is not metaphysically independent from its purported effects, it must be causally inefficacious. Otherwise, the causal explanation would be almost tautological: one would be saying that the fact about me that caused me to tend my wound was that I was in a state that caused me to tend my wound. The tautological character of this seems antithetical to causal explanations.This recalls Molière's satirical dialogue in *The Imaginary Invalid*, where a character explains that opium causes sleep because it has a "dormitive virtue". But then, if the functional characterization of pain is causally inefficacious, we still have the same problem as the reductionist, namely, how can one explain the regularities that involve pain?

A possible response to the necessitation problem is to bite the bullet and to claim that causes need not be metaphysically independent from their effects (e.g., Shoemaker 2001). I believe that this is a plausible response. Some have claimed that properties have "causal essences" (e.g., Shoemaker 1997). That is, some have claimed that even physical properties

have their effects necessarily. Electrons, for example, are claimed to repel other electrons by their essence, such that, if an electron were to cease repelling other electrons, it would cease to be an electron. This means that it is contradictory, and hence metaphysically impossible, for an electron to not repel other electrons. Someone who accepts this view could say, in a way inspired by Kripke (1980), that the fact that electrons repel other electrons is a necessary a posteriori fact. One could not know that electrons repel other electrons, and this could produce the illusion that this is a contingent fact, when in fact it is necessary. One cannot conceive of a world in which electrons don't repel each other; if one thinks one can, one is not really conceiving of electrons, but of other entities which share some similarities with them (Shoemaker 1997, pp. 131-132). Others have criticised this view and hold that properties, physical or not, are metaphysically independent from their effects (such as Armstrong, 1999). I will not go into the details of this dispute, nor take a stance on it.

If one accepts the properties can be metaphysically tied to their effects, the argument of necessitation loses force. If, even in the most paradigmatic cases of causation, properties are not metaphysically independent from their purported effects, then it is not a problem that pain necessarily involves wound-tending under normal circumstances. That is, whether or not one finds the argument from necessitation convincing seems to depend on what views one has about the relation between properties and their effects. If the causal functionalist takes a stance similar to Shoemaker's, she has a plausible response to the argument of necessitation.

4.2.2. The problem of over-generation

The causal functionalist can claim that functional properties can cause events, even if they are metaphysically dependent on those events. But if one admits that there can be causal relations in these cases, one is at risk of over-generating causes. That is, one is at risk of having to admit causes that one does not want to admit.

Take again the case of people who, coincidentally, get rich after reading a book. In this case, one could postulate a functional property, call it F, which by definition is caused by reading the book and causes an increase in wealth. In this case, all the people who got rich by reading the book share this state. If functional states are causally efficacious, then arguably being in state F caused the people to get rich. Here we are again confronted with the threat of overdetermination. Remember that one of the people got rich because they read a copy of the book which contained a check. In this case, there seem to be two sufficient causes of the same effect: one of them is reading a book that contains a check, and the other one is being in state F. On the one hand, reading a book with a check in it is enough to become rich, so there is no work left for property F to do. On the other hand, being in state F is by definition enough to become rich, even if one does not read a copy containing a check. There are, seemingly, two causes that don't need each other, but that happen to produce the same effect.

If one follows Yablo's solution to this problem, then one would have to say that F is proportional to the effect, whereas reading a copy of the book containing a check is not. In Yablo's terms, reading that copy of the book is not required to become rich, since there are other ways in which people can become rich because of the book (for example, by reading it at the right café). Just like in the case of pain, the specifics of the causing events would end up being irrelevant and mere determinates of the real cause, which would be the property F. Reading the book is, of course, not enough to become rich, since it is possible for someone to read the book and not become rich. Nevertheless, having property F is enough to become rich, since this property by definition causes people to become rich. This conclusion is clearly problematic: finding a check inside the book is, in fact, what caused this person to become wealthy.

Attributing causal efficacy to F has problematic implications for inductive inferences. If the cause of people getting rich is F, then it is fair to assume that other similar people will

also enter that state when reading that book. Since causal functionalism does not see these cases as distinct and independent, it would lead one to believe that it is no coincidence that all these people became rich after reading the book. Therefore, functionalism suggests that people who read the book in the future will continue to become rich. But this is not a good inference: if people got rich for independent reasons, then there is no reason to assume that there will continue to be a correlation between reading the book and becoming rich. Whereas reductionism suffers from a lack of systematicity, functionalism suffers from an excess of it. Reductionism mistakes valid inferences for invalid, and causal functionalism mistakes invalid ones for valid.

The causal functionalist can respond that some functional properties are causally efficacious and others are not. One could attribute causal efficacy only to properties like feeling pain, having eyes, and others, but not to properties like F. I am very sympathetic to the question, as will become apparent in the next chapter. Nevertheless, it does generate an issue for the causal functionalist, namely, what distinguishes causally efficacious and inefficacious functional properties?

An answer to this challenge is suggested by Weiskopf (2011), who distinguishes between functional properties that constitute natural kinds and those that don't. He says that, in the former case, the functional properties figure in successful explanations of phenomena. But, even if this might be enough for the purposes of Weiskopf's paper, it is not enough to avoid the problem of over-generation. The question seems to be, precisely, in virtue of what do some functional properties figure in successful explanations and others do not. Pain being causally efficacious explains why it figures in successful scientific generalizations, not the other way around. Therefore, one needs an account of why pain is causally efficacious that does not involve the role it plays in science. Another, related way of distinguishing between efficacious and inefficacious functional properties is to claim that causally efficacious

properties are those that can be used in valid inferences. The problem here is that one still should explain in virtue of what some functional properties have a role in valid inferences whereas others do not. In chapter 5, I will consider another method of demarcation that I take to be more successful, and which is based on natural selection.

A different solution that the functionalist could offer for the problem of over-generation is to claim that functional properties should not only be defined by what causes them and what they cause, but also by how this transition takes place. One could say, then, that having the same input and output is not sufficient for having the same property, and that the steps that lead from one thing to another should also be the same. It could be argued, then, that, although all instances of F share the same input (reading the book) and the same output (being rich), the intermediate steps are different in this case, which makes them count as different properties.

The central issue with this solution concerns how these intermediate steps are to be defined and individuated. If they are individuated by their physical constitution, then this view struggles to account for multiple realizability: octopus pain and human pain would, under this framework, count as different properties. It seems preferable, then, to define intermediate steps in non-physical terms, presumably in terms of their causal roles. Pain, then, would not consist merely in the coarse-grained transition from bodily harm to wound tending, but in a fine-grained sequence of steps, each contributing to the final behavior. If this transition does not occur via the appropriate intermediate steps, it would, according to this view, not qualify as pain (for a similar view, see Block 1980). One could wonder if properties such as feeling pain or having eyes really have common intermediate steps shared by all their instances. Initially, it is not clear why they should; if the physical realizers can vary, why couldn't the intermediate steps also vary? But a deeper problem is that, if the steps are defined functionally, then the problem of over-generation seems to arise again.

To illustrate how the over-generation problem arises again, consider two more cases. Suppose one of the people who read the book is inspired by a passage that describes France, so she decides to travel there. She stumbles with a museum and goes in. Since the floor is wet, she falls, hurts her back and receives a significant compensation, which makes her rich. Now take a different person who reads the book. He hates it so much that he decides to go to Canada, where the author lives, to tell him how bad it is. In Canada, he decides to visit a museum, in which he happens to find a check on the floor. Despite their clear differences, these two scenarios follow the same sequence of steps: a person reads a book, which causes them to go to a different county, which causes them to visit a museum, which causes them to be rich. Is this enough to say that the same property is at play in both cases? If the answer is 'yes', then the problem of overgeneration has appeared again: two unrelated cases are being lumped together. If the answer is 'no', then the functionalist still needs to explain what distinguishes them. It is not their functional role, as they have the same input and output, and it is not their functionally-defined intermediate steps, as these are also the same.

The problem of over-generation does not mean that we have to abandon causal functionalism. But it does mean that, to accept causal functionalism, one must provide an account of which functional properties are causally efficacious and which are not.

5. Natural selection: a possible way forward

Return to the case, described in sections 3.2.3 and 4.2.2, in which various people coincidentally become rich by reading the same book. I have argued that both reductionism and functionalism have difficulties distinguishing this case from non-coincidental cases, such as the case of pain causing wound-tending in different species. Reductionism seems committed to the claim that both these cases are coincidental, whereas causal functionalism seems committed to the claim that neither of them are. But, in fact, the book case is coincidental and the pain case is not. How should one distinguish them?

The way I have set up this problem heavily suggests a solution. The difference between both cases is that the pain case is a result of natural selection, whereas the book case is not. It is not a coincidence that humans and octopuses behave in similar ways; rather, they were selected to do so. In the book case, by contrast, it is a coincidence that the various people who read the book got rich, because natural selection was not involved. This difference suggests a demarcation criterion: that irreducible properties that result from natural selection can be causally efficacious, whereas those that do not result from such a process cannot.

It is of course possible that there is some way of distinguishing both cases that is not put in terms of natural selection. But, since it is not very apparent what this might be, and since natural selection seems like a very clear option, one should take this possibility seriously. In this section, I will suggest two ways in which this intuition could be developed into an account of the causal efficacy of irreducible properties. I will also present some implications of this idea and some issues for further research.

5.1. Two ways of integrating natural selection

There are two main ways in which one could use natural selection to avoid the problems discussed so far. One option is to amend functionalism. One can claim that functional properties that result from a selection process are causally efficacious, whereas functional properties that do not result from such processes are not. A functional property would then be causally efficacious iff it was selected to have the effects that it has. I uncreatively call this the "first theory".

What does it mean that a property *was selected to* have a particular causal role? For example, what does it mean to say that pain I experience was selected to produce wound-tending? The answer to this question could be formulated in terms of three conditions: (1) pain had that causal role for my ancestors, (2) the fact that it had that causal role contributed to their evolutionary success, and (3) that is the reason why I have that state (a very similar formulation is expressed by Neander, 1991, p. 74). In the human case, that state might involve a thalamus and in an octopus, it might involve something else, but both count as the same state because both were selected to have the same causal role.

The first theory, which is a kind of functionalism, can avoid the problem of over-generation. The property F, mentioned in the case of the book, is not a property that results from a selection process, and therefore it is not causally efficacious. According to this view, pain is causally efficacious not only by virtue of its functional properties, but also by virtue of its evolutionary history. Humans and octopuses both feel pain, not solely because they have the same functional role, but also because their functional properties result from natural selection. That is, there are two conditions that must be met for two organisms to share a property like pain: it must have the same causal role and the causal role must have been selected for.

Although the first theory avoids the problem of over-generation, it does not avoid the problem of necessitation. The first theory entails that causing wound-tending in normal circumstances is part of the definition of pain. As mentioned in section 4.2.2, for some this might not be such a bad thing. But if one is convinced that causes must be metaphysically independent from their effects, this problem calls for a different theory. An alternative to the first theory is to define pain only in terms of the functional role that it was selected to have, rather than on the one that it in fact has. Call this the "second theory".

The second theory avoids the problem of necessitation. It is metaphysically possible for me to be in a state that was selected to make me tend my wound in normal circumstances, without it actually making me tend my wound in normal circumstances. One could define pain, then, as a state that I inherited from my ancestors, which *in them* caused things like wound-tending, contributing to their evolutionary fitness, and which I inherited precisely because it contributed to their evolutionary fitness. Under this definition of pain, it is metaphysically possible for me to have pain under normal circumstances and not tend my wound. The cause and the effect are metaphysically independent. This definition also allows for multiple realization: an octopus can also be in a state that fulfills these conditions, since they both can be in a state that was selected to make them tend their wounds upon bodily harm.

Whereas the first theory defines pain both by its functional role and by its history, the second one defines it only in terms of its history. Which option one finds more appealing will probably depend on one's views of properties, namely on whether one takes properties to have a "causal essence" (like Shoemaker, 1997) or not (like Armstrong, 1999). I will not take a stance in the debate between the first and second theories.

Both theories are similar to so-called *etiological teleosemantics*, which purports to account for mental representation in terms of natural selection (a theory put forward, inter

alia, by Millikan, 1984, and Neander, 2012). According to this view, part of what it means for a being to have representational states is for it to have evolved in certain ways, for example, to have evolved to store information about the world (this is the version favored by Neander, 2012). In fact, if one takes representational state such as belief to be irreducible and causally efficacious properties, then the first and second theory will be committed to etiological teleosemantics, although not necessarily to any specific version of it.

The first and second theories are also similar to the views espoused by Macdonald (1992) and Papineau (1992), who claim that non-physical laws of nature can exist where there were processes of natural selection, which, according to them, explains why sciences like biology are explanatorily irreducible to physics. These accounts are similar but distinct from mine, as I discuss causation rather than natural laws. It is also worth noting that Macdonald and Papineau are concerned with the possibility of "autonomous sciences", i.e., with the possibility that sciences like biology and psychology can be practiced independently of physics. This would singly mean that biology and psychology can be practiced exclusively using irreducible properties. I make no claim about this. I have argued that there are irreducible and causally efficacious properties, but not that a science can be made solely using these.

5.2. The fate of Yablo's theory

The two theories offered above entail that some functional properties are not causally efficacious. But, as remarked in section 4.2.2, Yablo's account of causation seems to entail that every functional property is causally efficacious. Having the functionally-defined property F, for example, is by definition *enough* to become rich. It is also *required*, since, if one of the people who got rich had not had that property, they would not have gotten rich. It would seem, then, that if we claim that F is not causally efficacious, we have to abandon

Yablo's account of causation. This is a potential problem, as Yablo's theory offered a plausible response to Kim's argument, introduced in section 3.3.1, according to which accepting that irreducible properties are causally efficacious commits one to the claim that events are systematically overdetermined. If we abandon Yablo's theory, we are again vulnerable to Kim's argument.

One option would be to accept an account of causation different from Yablo's and respond to Kim's challenge in a different way. But it is also possible to hold on to Yablo's account, provided that one makes a small adjustment to it. To do that, one must add a condition to Yablo's theory that excludes undesirable properties from counting as causally efficacious. What exactly this condition looks like will depend on whether one favors the first or the second theory. If one accepts the second theory, then one could accept Yablo's account while postulating that an event can only cause another event if they are metaphysically independent, that is, if it is metaphysically possible for each to take place without the other.[14] This would rule out property F, but not pain, since the second theory does not define pain in terms of the effects it has, but in terms of the effects it was selected to have, and therefore makes it metaphysically independent from its effects.

If one accepts the first theory, one cannot make this move, since the first theory accepts the idea that effects are not metaphysically independent from the causes. One would have to add a different condition to Yablo's account. One option would be to establish a theory of natural kinds that takes into account natural selection, making pain a natural kind but not property F. If one makes this move, then one can establish that causation only takes place between events that can be defined in terms of natural kinds (either fundamental properties or properties that result from natural selection). This would, again, rule out F but not pain, as pain evolved and is therefore a natural kind.

[14] LePore and Loewer (1987, p. 635), for example, add such a constraint into their counterfactual account of causation.

Holding on to Yablo's theory would allow the first and second theories to resist Kim's argument. Octopus pain and human pain would count as different determinates of the same property, since they were selected to do the same things. Feeling pain simpliciter would then be proportional to wound-tending: had I not been in a state selected to make me tend my wound, I would not have tended my wound. This means that it is required. Additionally, I would have tended my wound independently of what exact such state I was in, as long as it was a state that was selected to have that causal role. This means that it is also enough. Specifically human pain, on the other hand, is not required, since I would also have tended my wound if I had been in a different, more octopus-like pain.[15] Therefore, human pain and pain simpliciter and not two independent causes that overdetermine my wound-tending. The cause of my wound-tending is pain simpliciter, and human pain is merely its determinate.

5.3. Some implications

The first and second theories allow for the irreducible properties to be causally efficacious, but only some of them. It is possible to accept that properties of biology and psychology are causally efficacious, since they are the result of natural selection. The case of social properties is not that straightforward and it is one of the main questions for future research regarding the first and second theories. These include properties used in sociology and economics, such as being a currency (this example appears in Fodor, 1974). Since various non-coincidental generalizations involve social properties, one would be well-advised to accept that these properties are causally efficacious. It is also very unlikely that they are reducible to the physical. One strategy would be to claim that social properties evolve through (something like) natural selection, and that they also have causal roles that they were selected to carry out (a similar view is that of Dawkins, 1976). A second option would be to

[15] This point is formulated with the second theory in mind. It also applies to the first theory, but the formulation would vary slightly.

define social properties in terms of the psychological and biological properties of the people involved, and then define those psychological and biological properties by their evolutionary history. Both these options will have to make some commitments about social properties and their nature. I solely point to these two options, but choosing between them and developing their implications is a question for further research.

Outside of biology, psychology and the social world, it seems that no properties would count as both irreducible and causally efficacious. Meteorological phenomena like tornadoes, for example, would have to be either defined in physical terms or deemed causally inefficacious. For some, this might be a problematic implication. Nevertheless, it is not entirely implausible. Suppose someone were to claim that tornadoes are multiply realized, that, say, tornadoes on Mars and on Earth have similar effects on the world but do not fall under a common physical definition. If someone were to claim this, it would raise the question of why tornados on Earth and Mars have similar effects. If it is due to physical similarities between them, then one could define a physical property through those similarities and use it to explain the similar effects. But if there are no physical similarities that account for these effects, it seems that it would be a mere coincidence that tornadoes on Earth and Mars have similar effects. In that case, just like in the case of the book, they should be explained differently. The property of being an Earth tornado would have to be counted as different from the property of being a Mars tornado, and each of them would have to be reduced to the physical.

Maybe there is a way in which tornadoes on Earth and Mars can produce similar effects non-coincidentally and without common physical properties that account for these effects. But it is not clear what this way would be. Properties that evolved through natural selection don't have this problem, since they clearly show how different physical states can non-coincidentally end up producing the same result.

There is another implication of the first and second theories which will probably dissuade some from accepting these theories. It involves a well-known example. Imagine that a lighting strikes in a swamp and creates a perfect replica of me, down to the last atom. This being would have all my physical properties, but it would not have my evolutionary history (this thought experiment was proposed by Davidson, 1987). This "Swampman" example has been extensively discussed in the literature, partly because it seems like a problem for etiological teleosemantics (hereafter just "teleosemantics"). Swampman has no evolutionary history, but teleosemantics claims that a necessary condition for having representational states is to have an evolutionary history. That is, teleosemantics is committed to the claim that Swampman does not have representational states, that he has neither beliefs nor desires nor intentions. Teleosemanticists usually bite the bullet and say that Swampman has no representational states (Millikan, 1996; Neander, 1996; although see Papineau, 2021, for a different approach). I don't have anything to add to this debate, but I do want to mention that it applies to the first and second theories as well. If one defines pain in terms of an evolutionary history, then it seems that Swampman feels no pain, since Swampman has no evolutionary history.

Some might be willing to accept that Swampman has no representational states, but not that he feels no pain. One could claim that a "correct" definition of pain must include swampmen. This has the potential for opening up an immense debate about what pain "really" is. But, as mentioned in section 4.1, words like "pain" can be used in various ways. If one defines pain functionally, then Swampman does indeed feel pain, since he can undergo states that have the same causes and effects as my pain. This would allow us to attribute pain to octopuses, humans and Swampmen. I have no problem with defining pain in this way. My suggestion is that pain, *qua causally efficacious*, should be defined with reference to its evolutionary history. This is compatible with accepting other definitions of pain, and even

with accepting that those other definitions do a better job at capturing what scientists and laypeople mean when they say "pain" (although I don't know if the functional definition achieves this or not). So, under a functional definition of pain, Swampman does feel pain. But, under that same definition, pain is not causally efficacious. As mentioned, many will find this counterintuitive, and it is likely that it will dissuade them from accepting the first or second theories.

7. Concluding remarks

I hope to have achieved three things with this thesis. The first is to cast some doubt on the doctrine of causal reductionism. I argued that there are some regularities which involve irreducible properties (such as the relation between pain and wound-tending), and that they will seem like mere coincidences if we accept causal functionalism. Secondly, I hope to have shown that, if we accept causal functionalism instead, we are at risk of having the opposite problem, that is, of regarding some coincidental regularities as non-coincidental. Thirdly, I suggested two possible solutions to these difficulties which I take to be plausible. Both of them claim that, for an irreducible property to be causally efficacious, it has to be the result of natural selection. One of the solutions takes this as an addition to functionalism, whereas the other takes it as a fully different theory. These two theories allow us to distinguish between coincidental and non-coincidental regularities. I also mentioned some open questions, some implications and some potential costs of this approach.

References

Ab Aziz, C. B., & Ahmad, A. H. (2006). The role of the thalamus in modulating pain. *The Malaysian Journal of Medical Sciences 13*(2), 11-18.

Antony, L. & Levine, J. (1997). Reduction with autonomy. *Philosophical Perspectives* 2, 83-105.

Armstrong, D. (1999). The causal theory of properties: Properties according to Shoemaker, Ellis and Others. *Philosophical Topics 26*(1/2), 25-37.

Batterman, R.W. (2000). Multiple realizability and universality. *British Journal for the Philosophy of Science, 51*(1), 115-145.

Block, N. (1980). Troubles with functionalism. In N. Block (Ed.), *Readings in the Philosophy of Psychology* (pp. 268-305). Harvard University Press.

Cao, R. (2022). Multiple realizability and the spirit of functionalism. *Synthese, 200*(6), 1-31.

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

Crook, R. J., Hanlon, R. T., & Walters, E. T. (2013). Squid have nociceptors that display widespread long-term sensitization and spontaneous activity after bodily injury. *Journal of Neuroscience, 33*(24), 10021–10026.

Davidson, D. (1970). Mental events. In L. Foster & J.W. Swanson (Eds.), *Experience and Theory* (pp. 79-101). University of Massachusetts Press.

Davidson, D. (1987). Knowing one's own mind. *Proceedings and Addresses of the American Philosophical Association, 60*(3), 441–458.

Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese 28*(2), 97–115.

Galdino, G., Veras, F. P., & Dos Anjos-Garcia, T. (2024). The role of the thalamus in nociception: Important but forgotten. *Brain Sciences 14*(8), 741.

Hempel, C.G. (1969). Reduction: Ontological and linguistic facets. In S. Morgenbesser, P. Suppes, & M. White (Eds.), *Philosophy, Science, and Method: Essays in Honor of Ernest Nagel* (pp. 179–199). St. Martin's Press.

Kim, J. (2000). *Mind in a Physical World*. MIT Press.

Kitcher, P. (2001). The World As We Make It. In *Science, Truth and Democracy* (pp. 43-54). Oxford University Press.

Le Pore, E. & Loewer, B. (1987). Mind matters. *Journal of Philosophy, 84*(11), 630-642.

Levin, J. (2023). Functionalism. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy* (Summer 2023 Edition). Stanford University. https://plato.stanford.edu/archives/sum2023/entries/functionalism/

Lewis, D.K. (1983). New work for a theory of universals. *Australasian Journal of Philosophy, 61*(4), 343-377.

Lewis, D.K. (1986). *Philosophical Papers: Volume II*. Oxford University Press.

Molière. (1999). *The imaginary invalid*. M. H. Richey & R. B. Bosley (Trans.). Hackett Publishing. (Original work published in 1673)

Macdonald, G. (1992). Reduction and evolutionary biology. In D. Charles and K. Lennon (Eds.), *Reduction, Explanation, and Realism*. Oxford University Press.

Menzies, P. & List, C. (2010). The causal autonomy of the special sciences. In G. Macdonald & C. Macdonald, *Emergence in mind* (pp. 108-129). Oxford University Press.

Millikan, R.G. (1984). *Language, Thought, and Other Biological Categories*. MIT Press.

Millikan, Ruth Garrett (1996). On swampkinds. *Mind and Language 11*(1), 103-17.

Neander, K. (1996). Swampman meets swampcow. *Mind and Language 11*(1), 118-29.

Neander, K. (2012). Toward an informational teleosemantics. In D. Ryder, J. Kingsbury & K. Williford, (Eds.), *Millikan and her critics* (pp. 21-41). Wiley.

Paley, W. (1809). *Natural Theology: Or, Evidences of the Existence and Attributes of the Deity* (12th ed.). J. Faulder.

Papineau, D. (1992). Irreducibility and teleology. In K. Lennon & D. Charles, *Reduction, Explanation, and Realism*. Oxford University Press.

Papineau, D. (2022). Swampman, teleosemantics and kind essences. *Synthese 200*(509).

Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, Mind, and Religion* (pp. 37–48). University of Pittsburgh Press.

Rupert, R. (2006). Functionalism, mental causation, and the problem of metaphysically necessitated effects. *Noûs, 40*, 256-283.

Shanthanna H. (2018). A case report of a thalamic stroke associated with sudden disappearance of severe chronic low back pain. *Scandinavian Journal of Pain 18*(1), 121-124.

Shoemaker, S. (1997). Causality and properties. In D.H. Mellor & A. Oliver, *Properties*. Oxford University Press.

Shoemaker, S. (2001). Realization and mental causation. In C. Gillet and B. Loewer, *Physicalism and Its Discontents* (pp. 74-98), Cambridge University Press.

Strevens, M. (2012). The explanatory role of irreducible properties. *Noûs, 46*(4), 754-780.

Subramanian, S.; Tseng, B.; Del Carmen, M.; Goodman, A.; Dahl, D.M.; Barbieri, R.; Brown, E.N. (2024). Monitoring surgical nociception using multisensor physiological models. *Proceedings of the National Academy of Sciences of the United States of America, 121*(40).

Weiskopf, D.A. (2011). The functional unity of special science kinds. *British Journal for the Philosophy of Science, 62*(2), 233-258.

Woodward, J.F. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Woodward, J.F. (2010). Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biology and Philosophy, 25*(3), 287-318.

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