Durham E-Theses

The chances of higher-level causation: an investigation into causal exclusion arguments

KERTESZ, GERGELY

The chances of higher-level causation:

an investigation into causal exclusion arguments

Thesis submitted for the degree of
Doctor of Philosophy

by Gergely Kertész

Department of Philosophy
University of Durham
2019

Gergely Kertész: The chances of higher-level causation

It is well-known that non-reductive physicalism suffers from internal tensions between physicalist and antireductionist commitments. This thesis reconstructs Jaegwon Kim's classic causal exclusion arguments that aim to demonstrate the tension between the putative causal autonomy of multiply realized higher-level properties and basic commitments of physicalism and investigates some solutions that aim to dissolve the paradox highlighted by exclusion worries.

The main goal of this thesis is to appraise a solution, developed by Menzies and List, according to which, while the causation of an effect via a higher-level realized property is possible, it is incompatible with the causation of the same effect by a distinct realizing property. On this incompatibilist view causal exclusion is a contingent matter and can be directed both downwards and upwards, but it presupposes a difference-making account of causation. On the one hand, the thesis provides justification for preferring a difference-making account of causation over productive accounts presupposed by Kim's exclusion argument, on the other hand, it develops internal criticism against the approach suggested by Menzies and List.

Two strands of original arguments are formulated against this view. First, it is shown that downwards exclusion claims rest on the tacit assumption that realization and multiple realization can be modelled on the determinable-determinate relation, a premise explicitly rejected by Menzies. Independent arguments are also developed against this premise. It is shown that if the premise is unavailable Menzies has no viable argument for the claim that lower-level realizer properties cannot be proper causes of an outcome when it is also caused by the relevant realized property. Second, it is argued that higher-level causal autonomy is much less likely to occur than Menzies and List claimed it to be and inter-level causal compatibility is a lively option in the incompatibilist framework they developed.

The chances of higher-level causation:

an investigation into causal exclusion arguments

Thesis Submitted for the Degree of

Doctor of Philosophy

by Gergely Kertész

Department of Philosophy

University of Durham

2019

Declaration

I confirm that no part of this thesis has previously submitted for any degree at this or any other university. All of the content is the author's own work, except discussion of others' work, which has been indicated in the text. Some of this work is under revision in a journal, some appeared in print before, the latter is referenced below.

Unrecognizable parts of chapter 2 and 4 were published (in Hungarian) **as:**

Gergely Kertész (2015): *On the problem of causal autonomy*. In: Márton Miklós–Molnár János–Tőzsér János (eds.): Realizmus, magyarázat, megértés. L'harmattan Kiadó

Statement of Copyright

The copyright of this thesis rests with the author. No quotation from it should be published without the author's prior written consent and information derived from it should be acknowledged.

Word count:
Preliminaries: 2328
Main text: 89856
References: 2803
Sum total: 94987

Thesis outline

The rise of non-reductive physicalism in modern philosophy resulted in an internal tension between physicalist commitments and antireductionist commitments. This thesis investigates causal exclusion arguments that aim to track this tension and also some moves that try to handle the problem highlighted by the tension. It argues that the concept of causation utilized plays a more important role in the evaluation of these arguments than Jaegwon Kim thought when he developed the first versions of the exclusion argument. The thesis finds that the application of difference-making concepts of causation to resolve said tension is a well-justified move. There are at least two interestingly different ways of reformulating the exclusion worry based on difference-making theories of causation. The main goal of the thesis is to appraise the so-called incompatibilist path worked out by Menzies and List, according to which multiply realized higher-level causes exclude their lower-level realizers from causal efficacy. Remaining inside that framework, formulating internal criticisms, it is shown that higher-level causal autonomy is less likely to occur than Menzies and List claimed it to be. This insight pushes the framework close to a compatibilist interpretation of the relation between higher and lower-level causation.

The first chapter does the groundwork, it reconstructs the problem of mental causation and extends it to higher-level causation more generally as it was conceived in the second half of the last century. It introduces the antireductionist arguments developed based on the concept of the multiple realizability of higher-level properties and the resulting non-reductive physicalist position. It also introduces the most important interpretations of multiple realizability.

The second chapter reconstructs Jaegwon Kim's classic causal exclusion arguments against the causal autonomy of mental and other higher-level properties. It reviews Kim's

response to the paradox created by exclusion arguments that favours local reductions in the spirit of the mind-brain identity theory. Other possible treatments are explored as well. The chapter puts forward original arguments against the solution that allows for frequent overdetermination of outcomes by causes at different levels showing that this road is only open for emergentists who allow that the physical realm is not physically closed, a position that is inconsistent with physicalism. It is also argued on multiple grounds that the notion of productive causation utilized by Kim defending his version of the exclusion argument should be replaced with a more viable difference-making account.

The third chapter is an excursion that investigates a challenge for counterfactual theories of causation concerning transitivity. It argues in the context of the contrastive theory of causation, that contrary to what Johnathan Schaffer claims, causation is an intransitive relation. However, this doesn't create a problem for causal inferences in general because those special cases where causation fails to be transitive can be identified on structural grounds and one can decide whether a causal chain is transitive or not.

The fourth chapter reconstructs the incompatibilist solution to the causal exclusion problem worked out by Menzies and List relying on the notion of proportionate difference-making causation. According to these philosophers, causal exclusion is a contingent matter and we can see this simply by replacing the notion of a cause in Kim's exclusion argument with the notion of proportionate difference-making. Contrary to Kim's version, the reformulated argument allows for three options. Multiply realized higher-level properties exclude their realizers from causal efficacy with respect to certain outcomes. Some lower-level properties exclude higher-level properties realized by them when the higher-level property is not specific enough for the outcome in question. Inter-level causal compatibility occurs in cases when lower and higher-level properties are identical. The first of these options

shows that higher-level causal autonomy is an empirical possibility. Exploring this conceptual framework, the chapter finds a surprising further option. It shows that, at least theoretically, the criteria developed for inter-level causal compatibility allow for causal compatibility without the identity of higher and lower-level properties, even when the higher-level property is multiply realized. This contradicts the assumption that causally relevant multiply realized higher-level properties always exclude their realizers from causal efficacy.

The last chapter puts forward two strands of original internal criticism against the reformulated exclusion argument. Both are connected to the conclusion of the previous chapter. First, it shows that the claim that multiply realized properties downwards exclude their realizers, rests on a tacit assumption that is explicitly rejected by Menzies. Without presupposing that the realization relation can be modelled on the determinable-determinate relation Menzies and List have no viable argument for the claim that lower-level realizer properties cannot be proper causes of an outcome when it is also caused by the relevant realized property. The chapter also provides an original argument against the determinable-determinate interpretation of the realization relation.

The second line of argument shows, analysing the concepts of realization sensitivity and insensitivity of causal relations, a distinction crucial for the reformulated exclusion argument, that realization sensitivity, contrary to what Menzies claimed, might apply to multiply realized higher-level causal relations as well. This is an independent argument against the view that it is impossible to have cases where both a multiply realized property and one of its realizers come out as a proportionate cause. The chapter closes with the conclusion that the possibility of top-down exclusion is still empirically plausible, however it is less likely to occur than Menzies and List claim it to be. So, there is a good chance that we find cases of higher-level causal autonomy, but multiple realization is not sufficient for that.

Chapter 1: Mental causation, multiple realization and physicalism

1.1   On the importance of mental causation

According to Jerry Fodor's dramatic declaration, abandoning mental causation would have dire consequences. It would mean that "practically everything I believe about anything is false and it's the end of the world!" (1990:156). This is a fairly theatrical way of stating the problem but summarizes the gist of it. If mental properties were devoid of causal powers, it would be hard to make sense of many of our crucial everyday beliefs. Our common-sense concept of agency presupposes that our beliefs can be causes of physical effects and other mental states as well. An agent can act according to her will and can form intentions based on her beliefs and desires. If mental properties are epiphenomenal, if they lack causal powers on their own right, then wilful action is an illusion and our actions are caused independently of what we have in mind. Remembering, reasoning and other cognitive capacities that involve mental states causing further mental states would be similarly useless.

Most of modern psychology would be in the very same soup. Some classic behaviourist psychologists in the first half of the 20th century tried to dispense with cognitive content, but the language of conditioning processes proved to have paralyzing limitations. Many basic experiments in e.g. social psychology are performed and interpreted on the assumption that the subjects go through inner experiences, make decisions based on their beliefs and values and carry out actions motivated by further mental entities. From the 1950s onwards it became common sense in psychology that if such conceptual tools are rejected much of human nature remains elusive.

Maybe the most influential and important early experiments in the history of social psychology were the ones measuring the effects of cognitive dissonance on human behaviour

(See: Baumeister and Finkel 2010). Cognitive dissonance theory was on the forefront of the cognitive revolution in psychology and it was among the first theories that rendered classical psychological behaviourism obsolete. Attitudes and attitude change are indispensable categories for this theory and for most of later cognitive psychology. Therefore, it will serve as a good example for a scientific theory that is thoroughly committed to the existence of causally potent mental states.

There are countless modifications on the early version of this theory, but for present purposes it is enough to see the essence of the idea. The occurrence of cognitive dissonance can be a consequence either of an action that contradicts already held personal beliefs and values or of confrontation with new information that contradicts said beliefs and values. A person experiencing internal inconsistency experiences mental discomfort and is motivated to reduce said cognitive dissonance (see: Festinger 1957).

Dissonance reduction can follow various pathways, just to mention two: the person can actively avoid situations inducing the dissonance or can justify the deeds that induce the dissonance. In the first kind of case a person trying to lose weight might avoid places where there is easy access to sweets and junk food. In the second kind of case a person knowing that he has health problems connected to his/her obesity might invent ad hoc justifications for eating unhealthy food, like "I had a hard day, I deserve something that helps me to relax" or "this is a rare occasion, I'm only doing it under special circumstances".

Beliefs and attitudes are crucial for cognitive dissonance theory and for the description of the processes it is interested in. The dissonance stems from inconsistencies either between attitudes, beliefs or between an attitude or belief and an action carried out. The reduction of the dissonance takes place via adopting new beliefs or attitudes that help to create consistency between already held beliefs, attitudes and those that created the dissonance.

The notion of belief utilized in social psychology is quite close to the idea of a propositional attitude widely relied on in analytic philosophy and more specifically in the philosophy of mind. The main difference is that in social psychology attitudes are supposed to have intensity over and above content and an attitude towards the content. In any case, for present purposes the causal role these attitudes and beliefs play in the theory is the most important aspect to highlight. Inconsistency between them causes the dissonance and in turn the dissonance causes the adoption of further beliefs as a remedy that reduces the dissonance.

It is clear therefore, that the causal role of the mental is crucial for both our everyday sense of self and also for much of scientific psychology. When psychology explains attitude change as a result of mental conflicts it is committed to the existence and efficacy of mental causes. Thus, questioning mental causation results in sceptical questions concerning both the validity of causal explanations in psychology and at the same time everyday explanations of behaviour. For many, that seems to be a too high price to pay.

Much of the discussion in this thesis accepts what many would call mental realism. Realism about the mental has at least two related, but importantly different interpretations. First, it implies the view that the language we use to describe the mental aims to grasp, to reflect mental reality. This view does not imply that mental discourse succeeds in doing so, however independently of its level of success, it presupposes that the main function of mental discourse is to reflect mental reality.

Surprisingly, this is not the only stance one could adopt concerning the function of mental discourse. According to some so-called mental fictionalists, mental discourse might have a different function altogether. For them mental discourse serves the function of developing social coordination, but not by describing the mental world properly, it is rather a

tool to influence the attitudes, emotions and moods of others. This fictionalist interpretation presupposes that such influences can be performed by mental discourse without succeeding or even aiming at a realistic description of the mental realm of reality (see: Demeter 2013).

Those who believe that mental discourse aims to describe the mental world properly still have a few interesting options to interpret the status of mental discourse. One radical option is to claim that even though mental discourse aims to describe the mental realm, it fails to do so, and the failure is so spectacular that we should dispense with mental discourse and turn to e.g. pure neuroscience instead (see: Churchland 1981). There are those moderate realists who think that mental discourse utilizes useful fictions, idealizations, but the discourse as a whole is still capable of accounting for real patterns in human behaviour (see: Dennett 1991). On the other end of the spectrum there are those, such as Fodor whom I quoted earlier, who believe that mental discourse is basically successful in grasping mental reality and mental categories are indispensable for anyone who aims to provide an exhaustive picture of reality. This is the second sense in which someone can be a realist with respect to mental discourse (see: Demeter 2009).

On this doubly realist view, it is both true that mental discourse aims to describe mental reality and that entities described by mental terms exist and make differences to the course of events in our world on their own right. Realism concerning both the function of mental discourse and mental entities is made plausible by the fact that there are folk and scientific practices which are seemingly successful in predicting human behaviour that, at least implicitly, commit themselves to the existence of mental entities. However, many philosophers provided reasons for the contrary view according to which mental entities and properties should not be taken so seriously.

In this thesis, I will focus on a particular challenge to mental realism, that originated with non-reductive physicalist views concerning the mental, according to which even though mental properties are closely tied to physical properties they cannot be reduced to the physical, because they are multiply realizable by different configurations of physical properties. Interestingly this non-reductive physicalist view became something like an orthodoxy among contemporary philosophers, even though, as I will show in subsequent sections, this development resulted in a tension between the basic commitments of physicalism and the commitment to mental realism.

The classic version of the so-called causal exclusion argument worked out by Jaegwon Kim (1998, 2005, see also: chapter 2) was formulated to highlight this tension and to make the problem as visible as possible. It posed a challenge to Fodor and like-minded mental realists who wanted to maintain physicalism at the same time. It showed that their commitments combined would deprive mental states of their causal efficacy and reality. One prominent counterargument against that challenge was developed by Peter Menzies and Christian List (Menzies 2003, 2007, 2008, Menzies and List 2009, 2010, Menzies 2013, see also: chapter 4 and 5). This reformulated version of the exclusion argument turned the whole idea of causal exclusion upside-down providing an argument with the promise that antireductionist convictions can be maintained together with mental realism. In the chapters that follow I will analyse and evaluate both arguments. But first, let us see, what created the mentioned tension.

1.1.1 Origins of the contemporary problem of mental causation

The contemporary problem of mental causation originates from problems that first occurred with the advent of Davidson's anomalous monism. In his view, causation requires the

operation of strict laws. The interaction of mental and physical events cannot be covered by strict laws. Therefore, to be able to interact causally a mental event has to have a physical description that allows for that. According to Davidson every event could be redescribed as both a physical event and a mental event, but we are talking about the same event[1] in both cases. It is in virtue of the physical aspect of the event that it can enter into causal relationships, because the causal connections between events can only be mediated by strict laws and according to the contemporaries of Davison there are no strict psychophysical laws.

The properties mentioned in mental and physical descriptions are not identical, not even commensurable. They cannot be mapped on to each other. The mental is anomalous[2], so mental properties cannot be reduced to physical properties. But still, there is an intimate relation between token mental events and token physical events. The tokens of a mental and a physical event are identical to each other when they occupy the same space-time region. This allows for what is called token physicalism, as every mental event token is identical to a physical event token.

Anomalous monism was widely criticized for not being able to account for the causal relevance of mental properties[3]. In causal explanations, not all properties of an event are taken to be relevant for a particular outcome. The redness of a tomato is irrelevant to its

---

[1] Events are still the most popular candidate relata for the causal relation. There are important differences between the different concepts of events proposed (Lewis 1973, 1986; Kim 1973; Davidson 1970) but in all of them events are concrete particulars occupying a region in space-time. According to most interpreters, for Davidson the space-time region itself provides the identity conditions for events. For Kim events are property instantiations by objects in a region. And for Lewis an event is a property of a space-time region. There are theories that take the causal relata to be different kind of entities, but I won't consider such theories in this thesis.

[2] The exact reasons provided for this view are not important for present purposes.

[3] Early proponents of this criticism are Honderich (1982) and Kim (1993). The problem had a great influence on the philosophical community from the 1980s onwards.

rolling on a table, it rolls because of its sphericity. Suppose that redness and sphericity are the analogues of the physical and the mental properties respectively. If mental properties are real properties, they should be causally relevant at least to some outcomes even if in certain cases they are irrelevant. So, there should be a case when the redness of the tomato is causally relevant. However, according to the theory of causation sketched above anomalous mental properties don't even stand a chance of being causally potent. It does not help that mental and physical events are token identical, if causation between distinct events can only be explained by highlighting physical properties, then mental properties are never causally relevant, they are never part of the explanation answering the question why this token event caused that token event. As Kim puts this:

> "… *the fact that m is a mental event* - that it is the kind of mental event it is - *appears*
>
> *to have no role in determining what causal relations it enters into*. Event m's causal
>
> relations are fixed, wholly and exclusively, by the totality of its physical properties."
>
> (Kim 1998:34, my italics)

Some would argue that this standard reconstruction is a misunderstanding of Davidson's view (e.g. Gibb 2006, 2017). I won't try to take sides in debates concerning the right interpretation of his philosophy. What I summarized above is the received interpretation that defines the origins of the mental causation debate in the contemporary context.

The modern debate on mental causation goes on in the foreground of a monist worldview according to which in the end everything is physical or consists of basic physical components. So, the motivation and the questions posed are different from those defined by Descartes and later Enlightenment philosophers. As the name suggests, classic substance dualists assume that the mental and the physical are different substances. Their questions

concern the intelligibility of interactions between the unextended thinking substance and the extended substance of the physical. It would be wrong to say that there is nothing in common between the old and the modern problem of mind-body interaction, but in this thesis, I will barely mention the connection between the two and I will concentrate on modern issues.

If mental properties are devoid of causal efficacy on their own terms, then they are mere shadows of physical existence without having an effect on physical goings on. They are there, but nobody should really care about them. In philosophical parlance this view is called epiphenomenalism. As Kim summarizes:

> "… *epiphenomenalism* strikes most of us as obviously wrong, if not incoherent; the idea *that our thoughts, wants, and intentions might lack causal efficacy of any kind* is deeply troubling, going as it does against everything we believe about ourselves as agents and cognizers. […] even if we had to acknowledge it as true, *could not serve as a guide to life*; it cannot serve as a premise in our practical reasoning" (Kim 2005:70, my italics)

As the quote highlights the weirdest thing about this view is its impracticality. Suppose that mental properties are causally impotent, how should we change our lives in light of that knowledge? It seems that, even if it were true, we wouldn't be able to get rid of either mental explanations or the science of psychology. But, it seems that we would be working with languages that are empty shells not reflecting anything, much less the joints of nature. As again Kim puts it:

> "there may really be not much difference between these two options, *eliminativism and epiphenomenalism*. For a plausible criterion for distinguishing what is real from what is not real is the possession of causal power. As Samuel

Alexander said, something that 'has nothing to do, no purpose to serve' - that is, *something with no causal power might as well, and undoubtedly would in time, be abolished*." Kim (1998:119, my italics)

Therefore, in agreement with Kim, I do believe that we should save the mental realm from epiphenomenalism. In the following chapters, I will survey approaches to mental causation and higher-level causation in general that try to preserve its dignity. As I already said, most of this thesis is devoted to the critical analysis of theories that aim to preserve mental causation one way or another.

My main focus will be on the causal autonomy solution that originates with Peter Menzies. He and his allies try to convince us not only of the autonomy of the irreducible higher-level taxonomies of our world, but of the causal efficacy and relevance of the properties posited by mental and other higher-level descriptions of the world. I don't think that this autonomy solution is successful as it was originally formulated, so I will try my best to test it, to push it to its limits by formulating mainly internal, but also external criticisms of it as this is the only way to see its strengths and weaknesses. Menzies formulated his solution as a criticism of Kim's reductionist solution to the exclusion problem, therefore, after doing some introductory groundwork in this chapter, I will start to explain the exclusion problem by analysing Kim's approach in chapter 2.

1.2   Identity, multiple realizability and the special sciences

To be able to understand how philosophers ended up in the mess of epiphenomenalism with respect to higher-level properties like mental properties, one should also tell a short story of reductionism and antireductionism in modern philosophy of mind and philosophy of science more generally. The story usually starts with the so-called identity theory and the idea of reductive physicalism.

1.2.1   The mind-brain identity theory and some of its challenges

In the beginning there were identity theories of mind and brain or at least this is the point where the modern story starts. In the 50s a series of modern materialists (Feigl 1958, Smart 1959) proposed such theories. According to these philosophers, mental states, states of consciousness, sensations are identical to certain brain states. The classic semi-fictional example is the identity between pain and C-fibre firings in the nervous system. The idea is that as a gust is wind is the movement of a mixture of oxygen, nitrogen, CO_2 and some other molecules, similarly the experience of pain is a coordinated cascade of C-fibre firings.

Note, that this is a claim of type identity, identity between kinds grasped at different levels of reality. In other words, we could have said that all gusts of wind are the movement of the mentioned kind of mixture of gas molecules, and all experiences of pain are a certain kind of coordinated cascade of C-fibre firings. For many, this seemed to be a scientifically respectable rendition of the mind-body problem which implied the ontological reduction of higher-level kinds preferably to lower-level physical kinds. One of the goals of scientific reduction is reduction in the ontology of things and properties, entities in general. The only brand of reduction that can achieve this goal is the one based on inter-level property identification.

But for the identity theory to work there must be a one-to-one correspondence between the kinds, types or properties identified by psychology and neurology or physics, the kinds in the science that deals with the higher-level phenomenon to be reduced and the kinds in the other one that deals with the supposed reduction base. This is the assumption that Putnam (1967, 1975) famously challenged. Putnam doubted that pain, even if in humans it is based on C-fibre firings, is based on the same kind of physical-chemical state in other mammalian species not to mention molluscs equipped with more substantially different neural circuits.

We should note that Putnam's is an empirical, more specifically an abductive argument. In its original form it starts from the premise that many different living beings experience the same kind of experience, namely pain. The next premise is based on the plausible conviction that many of these living beings have considerably different nervous systems that realize different tokens of pain. The argument aims to tell us, that in light of the fact that the pain experiences of neurologically different living beings are the same while the corresponding nervous systems, physical realizers are plausibly quite dissimilar, the truth of the mind-brain identity theory is less likely than the truth of a so-called non-reductive physicalist view according to which everything is physical or is made up of physical parts and there are higher-level entities not type-identical with physical things. This argument is the first formulation of the so-called multiple realizability objection against the mind-brain identity theory. It relies on empirically plausible counterexamples to mind-brain identity claims. It aims to show that in fact higher-level kinds or properties are multiply realized.

The same basic idea was extended to more far-fetched sci-fi cases as well. Even though the structure of this argument is the same as that of the first version this one is less convincing as these sci-fi examples, even though they are conceivable, have no serious empirical

plausibility. Anyway, the structure of the argument is the same as in the case of the previous one. Silicon-based alien lifeforms, if they exist, might have silicon brains that differ from ours, intelligent robots, imbued with full-fledged mental lives, would probably have electrical computer brains. In spite of the mentioned differences, presumably these beings could experience a similar kind of pain as humans. If this is so, the truth of the mind-brain identity theory is less likely than the truth of a non-reductive version of physicalism. However, even if the argument is sound it proves only that pain is probably multiply realizable, not that it is in fact multiply realized.

Multiple realizability blocks type-type identity claims or the reduction of higher-level kinds to lower-level kinds, but as in the case of the Davidsonian view in section 1.1.1, it allows for token identity claims. It is consistent with multiple realizations to say that any token of a higher-level property is identical with a token of a lower-level physical property. The token identity claim is a necessity for those who would like to remain physicalists, and most contemporary philosophers do want to maintain some form of modern materialism. So-called token physicalism or non-reductive physicalism is still probably the most popular view among contemporary philosophers.

The arguments summarized above convinced enough people that by the end of the 1970s multiple realizability and non-reductive physicalism became basic notions in the toolkit of philosophers, especially in the philosophy of mind. Another, closely similar version of non-reductivism emerged in the context of philosophy of science, in discussions about reductive explanations in the sciences.

1.2.2 Identity and reductive explanation in the sciences

In other formulations the multiple realizability argument is presented as a refutation of reductionism in the sciences more generally. Fodor's seminal paper on "The disunity of science as a working hypothesis" (Fodor 1974) presents a challenge against Nagel's account of theory reduction based on bridge principles and the identification of properties described in the reducing and the reduced theory[4] (see: Shapiro 2004).

Traditionally, bridge principles are taken to be empirically motivated biconditional statements connecting properties in the reduced and the reducing theory. Theory reduction takes place when the kinds/types and explanatory generalizations of the theory to be reduced can be captured or closely approximated by the reducing theory and the latter theory is better in terms of explanatory power, delivers more precise predictions and/or has a greater scope.

Usually Nagel's theory of theory reduction is taken to be a theory of reductive explanation, not of ontological reduction. It is true that on the surface bridge-laws serve explanatory purposes for their function is to create connections between different theories using different descriptions, allowing the derivation of the reduced theory from the reducing theory. The possibility of such derivation does not imply the necessity of identities. However, according to Nagel (1961, 1970) in most cases connecting principles also declare the identity of the properties referred to by the different descriptions (see: Fazekas 2009:305-306).

Nagel was particularly interested in so-called heterogenous cases of reduction, where:

---

[4] From the title of the paper, one would guess that Fodor aimed to criticize Oppenheim & Putnam (1958), a classic paper that formulated the reductive unity of all sciences as program. However, that program was far more liberal than the identity theory that came to prominance at about the same time. The original unity of science project is probably compatible with the later non-reductive physicalist view. It is the conceptual framework used by Fodor that makes it apparent that he aimed at attacking Nagel's view, more precisely a type reductionst reading of Nagel's view.

"the distinctive traits that are the subject matter of the science [that is reduced]

fall into the province of a theory that may have been initially designed for handling

qualitatively different materials" (Nagel 1961:340)

This qualitative distinctness is what makes reduction especially interesting. When temperature in gases is reduced to the mean kinetic energy of molecules, to use Nagel's favourite example, the bridge principle connecting the two terms aims to "maintain that what are prima facie indisputably different traits of things are really identical" (Nagel 1961:340). This implies that the two terms refer to the very same thing, in this case plausibly a property.

The qualitative distinctness emphasized by Nagel, and also by more recent accounts of reduction (see e.g.: Gillett 2010, 2016)[5], is quite straightforward in the case of temperature. No lower-level particle has a temperature. That term is meaningless in the realm of atoms or molecules. Among the more basic properties of particles there is only one that is relevant for the reduction of temperature, the kinetic energy of particles. In statistical thermodynamics the temperature of a volume of gas is equated with the mean kinetic energy of the ensemble of particles that make up the volume of gas in question[6]. However, the reduction achieved in this example seems to imply that the terms temperature and mean kinetic energy refer to the very same thing under the two qualitatively different descriptions from theories describing different levels of reality.

---

[5] Carl Gillett introduced the term „piercing explanatory power" to highlight the importance of qualitative distinctness between different levels of description in composition based reductive explanations or, as he calls them, scientific reductions.

[6] The physics of temperature is more complicated than the simplistic account presented here or by Nagel himself. There are important further assumptions the ensamble of particles has to satisfy (see: Sklar 2015). However, these details are largely irrevant for the issue of qualitative distinctness, so I set them aside here.

To be able to evaluate different antireductionist arguments based on multiple realizability considerations it is useful to differentiate at least two important aspects of reduction (see: Crane 2001a). First, *explanatory reduction*, when what is explained by a higher-level science gets explained, is made more intelligible based on a lower-level science. This expresses an asymmetric relation. The lower-level science, the description it provides explains a term or description of the higher-level science, but not the other way around. Explanatory reduction does not require inter-level identity between the entities assumed to exist, referred to under different descriptions by the sciences in question.

The second sense in which we can talk about reduction is *ontological reduction*, when it is shown that a term in the higher-level theory refers to the same thing (kind, property) that the lower-level reducing theory is talking about, under a different description. This relation is symmetric. Two terms in different theories are ontologically reduced to each other if the terms in question co-refer. Note that the identification of two entities does not eliminate either of them. Claiming that Charles Bronson is Charles Dennis Buchinsky does not imply that either Charles Bronson or Charles Dennis Buchinsky is non-existent. So, although reductionists like to use the phrase that something is "nothing but" or "just is" this or that, reduction in terms of identity is not the same as elimination. Identity reduces the number of autonomous entities we should accept but allows that there are different scientific categories in sciences at different levels that pick out the same thing.

Fodor's argument, much like Putnam's, is best understood as putting forward arguments against ontological reduction and type-identity claims, not against reductive explanations in a broader sense. Arguments against ontological reduction claim that the co-reference of terms in sciences interested in different levels of reality is highly unlikely. Fodor's classic argument requires two main steps.

The first premise it accepts is that proper reduction cannot be done if the natural kind terms in some ideally completed special science and the natural kind terms in an ideally completed physics are not co-extensive, do not refer to the same set of entities. The most important contemporary critics of this premise are the so-called new wave reductionists like Bickle (1998). According to them explanatory reduction without identity statements is possible and is part of scientific practice, so they consider the first premise of the present argument to be irrelevant for the reductionist project. However, when it comes to the relations between the entities the different sciences are committed to, Fodor still has a point, a point that concerns the issue of ontological reduction. One might want to dismiss this as practically unimportant for scientific explanation, but to answer the question „what are mental properties", or to answer similar questions in any special science, physicalists need to provide an answer that considers the question of identity.

The second premise in Fodor's argument is based on multiple realizability considerations. Examples of multiple realization from different sciences are put forward to show that the co-extension of kinds in sciences dealing with different levels of description required by the first step is implausible. Fodor follows Nagel's lead in talking about inter-level relations. He requires symmetric, biconditional bridge-laws for connecting the kinds of psychology and physics[7]. The proper reduction of a special science law to a physical-level relation requires two such bridge-laws:

---

[7] I should note that Nagel's view concerning bridge principles is ambigous. Even though his much discussed examples of reduction require biconditional connecting principles between the terms of the reducing and the reduced theory at certain points he accepts that asymmetric, one-way bridge-laws are sufficient for reduction, for the deriveability of the reduced theory from the reducing theory (see: Nagel 1961:355, footnote 5). Fodor's understanding of Nagel is definitely biased in one direction and considers only reductions based on biconditional bridge-laws that allow for the identification of the relevant terms of the reduced and the reducing theory.

Special science law: S1 → S2,   Physical law: P1 → P2

Bridge laws required for reduction: S1 ⟺ P1, S2 ⟺ P2

The one-way arrows above represent causal relations in a broad, unrestricted sense. For Fodor these are lawlike connections, but because in later discussions I will focus on causal interpretations of those relations, I will refer to them as causal relations[8]. Two-way arrows represent bridge laws connecting kind terms in the relevant sciences. However, because of multiple realizability it is unlikely that such kind of bridge-laws, reductions can be established in most special sciences.

According to Fodor special-science kinds have different extensions compared to lower-level sciences. His main example comes from economics, the science that, among other things, is interested in the laws of monetary exchange[9]. Money as a kind is multiply realized by different metal coins and different kinds of paper notes. Therefore, to connect the higher-level causal relation to lower-level kinds one needs realization specific bridge-laws for all different cases of realization:

Special science law of monetary exchange: S1 → S2

Physical laws of monetary exchange: P1 → P2, P3 → P4, P5 → P6

---

[8] Fodor is quite obscure in his talk of special science laws and physical laws. He is not alone with that, as we will see in chapter 2, Kim's later discussion of causal exclusion in terms of causal sufficiency suffered from the same problem.

[9] Fodor's example is Gresham's law which states that bad money drives out good money. The Wikipedia entry on this the subject sums it up perfectly: "if there are two forms of commodity money in circulation, which are accepted by law as having similar face value, the more valuable commodity will gradually disappear from circulation." (see: https://en.wikipedia.org/wiki/Gresham's_law)

One-way bridge laws based on simple conditionals:

P1 ⇒ S1, P3 ⇒ S1, P5 ⇒ S1   P2 ⇒ S2, P4 ⇒ S2, P6 ⇒ S2

P1, P3, P5 are lower-level kinds (let's say coins made of different metals) diverse realizers of special science kind S1. The higher-level kind term refers to all of the lower-level kinds, whereas all lower-level kinds have a relatively narrow extension where this narrow extension is a proper subset of the extension of the higher-level kind P1. The predicate that is the disjunction of all the lower-level kind terms that correspond to the higher-level term has the same extension as the higher-level kind money. However, the disjunctive predicate cannot be constructed without knowing the higher-level kind as there is no information available on the lower-level that could inform the collection of lower-level kinds into a meaningful bundle. It is also highly unlikely that the disjunction of possible lower-level realizers would be a proper kind term in any lower-level science, and also that any genuine law of a lower-level science would apply to that disjunction.

The point is simply that the kind terms of the lower and higher-level sciences themselves refer to distinct kind of things, therefore reduction in terms of identity between kinds or properties is blocked. Note that even though identity-based reduction is blocked, explanation for higher-level properties in lower-level terms might still be possible. E.g. we can predict or manipulate the occurrence of a higher-level property knowing that it is realized by a particular lower-level realizer without knowing anything about other realizers. So, non-identity between the kinds at different levels is compatible with the possibility of inter-level reductive explanations. However, such explanations are partial or local, as they only apply to certain realizers of a higher-level special science property.

What we cannot tell by knowing only that a higher-level property is instantiated on a particular occasion is what particular realizer lies below. Multiple realization is a many-one relation where a higher-level property can be realized by many, or potentially an infinite set of, different realizers.

It should be emphasized that multiple realization for a psychological kind requires sameness in terms of the psychological kind but possible alternatives in terms of underlying neuroscientific or other realizer kinds. A psychological kind is multiply realized if "quite different neurological structures can subserve identical psychological functions across times and across organisms" (Fodor 1974:113). This idea is neatly summed up by the formula: "same but different". The same special science property is realizable by different realizers.

Block and Fodor (1972) introduced three still much discussed empirical examples for multiple realization in the realm of psychology. First, consider brain plasticity, the brain's capacity to realize the same function in different brain regions. This shows itself when, usually as result of an accident, a brain region loses its original capacity to serve some function. E.g. when the auditory cortex gets injured, another region that normally serves other purposes, and has a different neuroanatomical structure, takes over the lost functionality. The second widely discussed case is the convergent evolution of certain psychological traits. It is analogous to the convergent evolution of biological traits, where the same function gets tinkered by the evolutionary process from different genetic, anatomical and physiological starting points. A good example is the body shape of sharks and dolphins. Sharks are cartilaginous fishes, dolphins are mammals: their physiology, even their basic organs are substantially different from each-other, but their bodies are streamlined, shaped in a surprisingly similar fashion because of their quite similar predatory behaviour. The third example is the all-time favourite in the literature, the possibility of artificial intelligence which

was already discussed in relation to Putnam's ideas. These examples are still much discussed. In the meantime, numerous philosophers have levelled challenges against their use as examples of multiple realization, but they are still the most influential examples of the idea of "same but different".

Let us summarize the antireductionist conclusion that follows from the failure of inter-level kind identity statements. Loosely following Francescotti's formulation of the issue (Francescotti 2014:6-7) it can be summarized in terms of the following argument:

MULTIPLE REALIZABILITY: Special science properties are multiply realizable with respect to properties of more fundamental sciences.

NO BICONDITIONALS: If special science properties are multiply realizable with respect to properties of more fundamental sciences, then there are no nomologically necessary[10] biconditionals linking the former to the latter.

NON-IDENTITY: If there are no nomologically necessary biconditionals linking special science properties to more fundamental properties, then the special science properties cannot be identical with properties of more fundamental sciences.

ONTOLOGIAL ANTIREDUCTIONISM: Special science properties are not identical with properties of more fundamental sciences.

---

[10] Expecting only nomological necessity I follow Francescotti's formulation, however I should note that nomological necessity is the bare minimum a physicalist has to accept, and many would require, something stronger, like metaphysical necessity (see my section 1.3.1).

So, the kind of antireductionism that can be reasonably derived from multiple realizability is a rejection of identities between properties at different levels. This conclusion does not speak against the possibility of inter-level explanations, or local explanatory reduction based on knowledge of some realizer of a higher-level property. It rejects the possibility of ontological reduction achieved thought type-identities. But if ontological reduction is impossible the road to the most straightforward form of physicalism is blocked as well. This is the main reason why most contemporary physicalists are non-reductive physicalists. Believers of this view usually think that there are distinct, non-physical properties that are nevertheless entailed by or in some sense determined by physical properties. This dependence is usually spelled out in terms of supervenience, the subject of the section that follows.

1.3   Physicalism and the physical determination of higher-level properties

Physicalism is a metaphysical view concerning the nature of our reality, a form of modern materialism, that most contemporary philosophers subscribe to, including Jaegwon Kim and Peter Menzies, whose causal exclusion arguments are the main topic of this thesis. Physicalism tells us that prima facie non-physical properties like mental ones are in fact physically acceptable in the sense that these properties are necessitated by basic physical properties. In physicalist circles it is also common to say that prima facie non-physical facts are nothing over and above the physical facts.

However, the precise nature of the relation between prima facie non-physical things and physical things is problematic. If one only considers options that do not deny the existence of prima facie non-physical properties, the clearest interpretation of the phrase "nothing over and above" would be identity. But as we already saw in the previous section identity between properties relied on in the special sciences and basic physical properties is highly unlikely. The challenge of multiple realizability means that one cannot be a physicalist on the cheap. As a result, the most popular form of physicalism today is non-reductive physicalism. To be able to formulate what physicalism means if identity statements are unavailable, philosophers had to come up with new conceptual means.

1.3.1   Supervenience physicalism

David Lewis formulated the idea of physicalism in a quite intuitive manner. He says that by making a copy of the physical realm, of all properties physical, we would make a copy of all the facts of the world. This metaphor contains the seeds of an idea of physical necessitation that is usually spelled out in terms of the supervenience relation. The core of this idea is usually expressed in the phrase: "There cannot be an A-difference, without a B-difference".

According to physicalism all properties in the world are either physical or are determined by basic physical properties. Spelled out in terms of supervenience[11]:

> **Physical supervenience**: if any physical system instantiates a non-physical property NP at time t, it is necessary that there exists a physical property P such that the same system instantiates P at t, and necessarily anything that instantiates P at any given time instantiates NP at the same time.

If there are two distinct properties NP and P, where NP supervenes on P, then two systems, or even possible worlds cannot be different with respect to NP without being different with respect to P. It is easy to see that this relation is compatible with reductive physicalism. If NP is identical with P, then NP trivially supervenes on P. The great advantage of the supervenience-based approach is that it is also compatible with non-reductive views, those that accept that higher-level mental and other properties are multiply realizable or in fact realized by distinct kinds of physical systems. The asymmetric one to many relation implied by the multiple realizability of higher-level properties is nicely accommodated by the supervenience relation.

The supervenience of the mental on the physical plays a central role in the causal exclusion arguments this thesis aims to investigate. It provides the connection between higher and lower-level properties and allows for moves from higher-level candidate causes to

---

[11] It is importnat to see that even though the supervenience relation accomodates the property variance pattern implied by multiple realization, supervenience implies different things than one-to-many realization. Supervenience is a reflexive relation, anything trivially supervenes on itself. This is not true of realization or multiple realization, nothing realizes itself. Supervenience is transitive, this is debated in the case of multiple realization (see: Polger 2008). Supervenience is non-symmetric, as sometimes it is symmetric, every reflexive case of supervenience is a symmetric case trivially, however the mental supervenes on the physical in an asymmetrical sense which is also consistent with supervenience.

lower-level ones and vice versa. So, before bringing in any new conceptual scaffolding, it is important to pin down exactly what form of supervenience is required for physicalism.

The supervenience claim above expresses what Kim called *strong supervenience*. According to that formulation the relation between NP and P holds in all possible worlds. So, no two individuals in any possible world can differ in terms of mental properties without a difference in physical properties. A weak notion, restricted in its scope to the actual world only, proved to be problematic because that would require us to show how is it that two things in only slightly different possible worlds can be the same in all physical respects while different in terms of supervenient properties. Strong supervenience also has the advantage that if that thesis holds then, seemingly, the physical entails the mental and other higher-level properties. However, there are two problems faced by a supervenience-based definition of physicalism or, as it sometimes called, supervenience physicalism.

1.3.2   The insufficiency of supervenience physicalism

First, as it has been argued by a number of philosophers (Crane 2001b, Kim 1999) it is not clear that supervenience physicalism can tell itself apart from emergentism, its most important traditional rival. Below, I will talk manly about an influential strand of emergentism that originates with C. D. Broad (1925). Emergentism, similarly to physicalism, is a form of substance monism, combined with property dualism. Emergent properties are both fundamental in the sense that they are not resultants, derivatives of physical properties, and dependent in the sense that they require certain underlying physical configurations for their emergence. They are also considered to be genuinely novel, in the sense that they have causal powers of their own not conferred by the underlying physical properties. These powers are expressed by emergent higher-level laws, so-called intra-ordinal laws. Emergent causal

powers exert so-called downward causal effects on physical properties. This view might sound weird for some, but there is a wider agreement today that it is a consistent view of nature (see e.g.: McLaughlin 1992, Hendry 2010b).

Emergent properties are connected to the physical base via synchronic trans-ordinal laws formally indistinguishable from a Nagelian bridge-law[12]. However, the metaphysical content they express is different from mere connecting principles as the properties they connect to the physical base are not resultants of physical properties. This means that emergent properties supervene on the physical base, but the modal strength of the relation should be modified if one aims to grasp the content of emergentism in contrast to physicalism. For the emergentist trans-ordinal laws are fundamental laws much like the laws of physics. However, two possible worlds might differ with respect to trans-ordinal laws, without a difference in physical laws. So, according to the emergentist two worlds, identical in all fundamental physical respects, can differ with respect to their supervenient properties. To tell emergentism and physicalism apart philosophers introduced modifications into the modal strength of the supervenience relation, therefore strong supervenience usually comes in two flavours:

Metaphysical supervenience: NP supervenes on P in all metaphysically possible worlds.

Nomological supervenience: NP supervenes on P in all possible worlds governed by the same laws as those governing the actual world (including trans-ordinal laws).

---

[12] The view, especially in contemporary formulations, allows for both biconditional or one-way trans-ordinal laws. This is also consistent with Nagelian formalisms. Interestingly, Nagel (1961) aknowledges that emergentism and reductionism can be captured by the very same formal descriptions.

Among others, Papineau (2008) believes that metaphysical supervenience expresses the commitments of physicalists, while nomological supervenience grasps the emergentist tenets. For emergentists, or even for full-blown dualists, it is possible that higher-level properties are absent from worlds were the required trans-ordinal laws are not in place. Contemporary dualists like Chalmers (1996) believe in the possibility of phenomenal zombies, beings that are identical to us in all physical respects but devoid of phenomenal mental states[13]. The same applies to any higher-level property an emergentist identifies as emergent, like biological properties in the case of C. D. Broad (1925).

Jessica Wilson (2005) launched an influential attack against this distinction arguing that accepting the plausible view of dispositional essentialism, according to which physical properties are individuated by their causal powers and those causal powers are essential to the properties, one cannot make sense of having the same physical properties in a word with extra emergent causal powers and in a world without those. If physical properties are individuated by what they do, by their causal powers, then without emergent laws being replicated as well, it is impossible to instantiate the very same physical properties in a different world. According to this view, laws of nature express the causal powers and so properties are individuated by the laws of nature that apply to them. If there are laws over and above physical laws those emergent laws also express the causal powers of physical properties as emergent properties depend on configurations of physical properties. If Wilson

---

[13] The problem of phenomenal consciousness is central to contemporary debates in the philosophy on mind. There seems to be no way of explaining how phenomenal states can arise from physical states. On the one hand, phenomenal states are non-functional, so they resist the means of standard scientific methods, on the other hand, zombies physically identical to conscious humans, devoid of phenomenal conscious experience are at least conceivable (see: Chalmers 1996).

is right, then the nomological/metaphysical distinction cannot be used to differentiate physicalism from emergentism.

Fazekas (2014) has a notable counterargument to this line of thought. He objects that emergent laws don't express the causal powers of the physical properties in the supervenience base of the relevant emergent property. This is because trans-ordinal laws are not causal laws, only higher-level, emergent intra-ordinal laws express causal powers and these causal powers are genuinely novel, fundamental powers by definition. They belong to emergent properties that are not resultants of the base properties. What is done by an emergent property is not done by the subvening physical properties. According to Fazekas, the reason why Wilson thinks that emergent laws express the causal powers of physical properties it that she, along with some important interpreters of emergentism like O'Connor & Wong (2015)[14], fails to distinguish between emergent causal (intra-ordinal) and non-causal (trans-ordinal) laws. A full-fledged dispositional essentialist might want to hold the view that a trans-ordinal law is just an expression of an essential potentiality that belongs to a subvening physical property, but that leaves no room for making sense of the emergentist claim according to which emergent properties instantiate genuinely novel causal powers in reality. If there are genuinely novel, emergent causal powers these cannot be essential potentialities of the physical base, because in that case they are not novel, not additions relative to the fundamental powers of physical entities. Therefore, such kind of essentialism would beg the question against the emergentist position.

---

[14] The referenced O'Connor & Wong text is an influential SEP entry edited by them. In section 3.1. discussing the „Standard Ontology of Emergence" describing mainly C.D. Broad's version, they tell the reader: „newness of property, in this sense, entails new primitive causal powers, reflected in laws which connect complex physical structures to the emergent features. (Broad's trans-ordinal laws are laws of this sort.)" This recapitulation of the emergentist view overlooks the trans-ordinal vs. intra-ordinal distinction.

The question concerning the utility of different supervenience relations in defining physicalism is probably not settled yet, but assuming that Fazekas is right, strong metaphysical supervenience seems to be a satisfactory way to express the core, or as it is usually called, minimal commitment of physicalism. There are further worries concerning strong supervenience physicalism, but because the question is not central to my investigation in this thesis, I will set them aside here.

1.3.3 Realization Physicalism

The second serious issue for supervenience physicalism stems from the fact that supervenience is not an explanatory concept. It is important to highlight that for a long time supervenience was thought of as some kind of more substantive determination or dependence relation, but in fact it only expresses a covariation between variables. As Kim puts it:

> "Supervenience itself is not an explanatory relation. It is not a 'deep' metaphysical relation; rather, it is a surface relation that reports a pattern of property covariation, suggesting the presence of an interesting dependence relation which might explain it" (Kim 1993: 167)

So, a supervenience claim in itself does not entail any "in virtue of" claim. It does not explain the relation between fundamental physical properties and higher-level properties, but it would be definitely useful to get an explanation of why e.g. mental properties supervene on physical properties or in other worlds why does the physical necessitate the existence of the mental. When such kind of explanation is available, as Horgan (1993) called it, the relation of superdupervenience holds. The best option to gain such understanding seems to be to develop a theory of realization that sheds light on how higher-level properties

are anchored in basic physical properties, on the reason why we think that higher-level properties have a more intimate relation to basal properties that supervenience is capable of representing (see: Baysan 2015, forthcoming). The need for a theory of realization is also necessitated by the multiple realization thesis, because if type-identity between basal and higher-level properties is not an option, one is required to provide an alternative explanation for the physical determination of higher-level properties.

What a realization relation expresses is that the realized property exists in virtue of the realizing property. To make this more tangible, let us take a closer look at one prominent theory of realization. Shoemaker (2007, 2013) defines a concept for property realization in the following way. A property P subset realizes a property NP if an only if the causal powers of NP are a proper subset of the causal powers of P. This is the so-called subset view of realization. In accordance with this definition if NP has more than one realizer capable of conferring the same subset of causal powers, then NP is multiply realizable. This solution works only on the dispositionalist assumption that properties can be identified with sets of causal powers. One obvious advantage of the view is that it provides a straightforward interpretation of the "nothing over and above" claim that is so close to the heart of all physicalists. If the powers of NP are inherited from P then NP has no autonomous causal powers, but it can still retain its distinctness from the base property as it consists of a different set of causal powers.

Subset realization also explains the supervenience of the NP property on the P property. Imagine that property NP is identical to the following set of causal powers CP {cp11, …, cp15} and it can be realized by the set of properties PR {P11, …, P19} because all properties in PR share the powers in CP. What differentiates them from each-other is that they have further causal powers as well. It is easy to see that whenever a particular property in PR, e.g.

P11 gets instantiated NP gets instantiated as well, but not the other way around. This conforms to the covariance pattern peculiar to supervenience as there cannot be an NP-difference, without a P-difference. So, this concept of realization explains both the intimate relation between NP and P properties and also the supervenience of NP on P.

There are a few other theories of realization available (see: Baysan 2015), the functional realization view, the determination view, according to which, realized and realizer properties are the determinables and determinates of the same property space (section 5.3 will detail that option), and there are mereological theories or realization. Here, I simply wanted to highlight the theoretical function of such a theory in defining physicalism.

The discussion above shows that if one wants to maintain a physicalist position then one should at least conform to supervenience physicalism. However, if that proves to be insufficient for physicalism, a question I won't try to settle here, then one way out is to subscribe to a theory of realization that explains the intimate relation between the physical base and the higher-level properties anchored to them.

1.3.4 The causal closure of the physical

Another way of putting the idea of physicalism, independently of supervenience and theories of realization, is the following. Once God had created the laws of physics and the initial conditions (physical properties) of the universe, the work was done: other facts or complex properties came for free (Crane 2001a). This provides the basic intuition for the principle of the causal closure of the physical. The idea is roughly that everything that happens in our world is bought about by physical causes. Nowadays causal closure is considered to be more and more important as a defining characteristic of physicalism as fewer and fewer people believe that supervenience in itself can provide a strong enough formulation of physicalism

(see: Wilson 2005, 2010), one that can differentiate it from emergentism, its main rival. Even though not everybody agrees with this worry, this is one important reason why the causal closure principle became a more important cornerstone of physicalism.

Physical closure is an important premise in the causal exclusion argument, the main topic of this thesis. Even though it seems to grasp the content of physicalism nicely, as Lowe (2000) has shown, it is not easy to formulate the principle in a satisfactory manner. Another problem is that the motivation behind the principle is empirical and the backing one can muster for it as of today is still questionable (see: Hendry 2010b, 2017). As this principle is highly important for all versions of the causal exclusion argument, the main topic of this thesis, I will devote a great deal of attention to it in chapter 2 section 2.3 after I discussed Kim's argument for the causal exclusion of higher-level properties.

1.4   Kinds of multiple realization

In the following sections, I would like to detail the idea of multiple realization and to involve further insights from recent debates around it. In section 1.2 I have summed up the idea of the multiple realizability of special science properties under the phrase "same but different". So far, I concentrated on groundwork and only well-known basic issues were touched upon. Now, further questions should be asked concerning the conceptualization of the required difference between the available realizers of a realized higher-level property[15]. The question I aim to answer here is the following: what kind of difference should there be for two realizers to be considered separate, distinct kinds of realizers? Is it enough that two realizers occupy different regions in space-time, or should they be different systematically in further, more specific lower-level physical properties? Is there a principled way to answer this question?

To answer the main question, let us start from intuitions derived from a relatively straightforward scientific example, one that I borrow from Richardson (1982)[16]. Hydrogen or oxygen atoms, to mention two well-known elements, are considered to be unified categories of atoms. However, isotopes of both hydrogen and oxygen are differentiated at a higher resolution of analyses. Isotopes of a chemical element in general differ in the number of neutrons while they are the same in other respects, in their number of protons and electrons. Isotopes occupy the same place in the periodic system (hence the name isotope meaning

---

[15] This is not an easy task. Realization is philosophical term of art, and it was used without much theoretical clarification in the early days of the debate about identity theory and multiple realization by Putnam, Fodor and others (see: Shapiro 2004). There are no firm pretheoretical intuitions philosophers can rely on in formulating the notion of realization or multiple realization. The only way to validate a formulation of this notion is to test whether it can accommodate received examples of inter-level explanations.

[16] It is useful to rely on such examples in the discussion of multiple realization as most examples from psychology are way more controversial and their proper evaluation requires more complicated preparations.

'same place'), while they differ in terms of their mass number because of the neutrons. The differences between isotopes have further measurable consequences with respect to some of their physical properties.

In this example hydrogen is the higher-level kind and its isotopes are the lower-level realizer kinds that might realize it. In this case it is straightforward that the distinct realizers differ from each other, but at the same time they are closely similar to each other as well. More than that, there are certain respects, the number of protons and electrons, in which they are not only similar, but the same. Those aspects in which the isotopes are the same allowed chemists to explain in lower-level terms why they put them into the same higher-level category and why do they share a certain set of chemical properties.

1.4.1 Mild multiple realization

Hydrogen and its isotopes are only one received example of multiple realizability, but it highlights something worth emphasizing: the fact that two realizers are physically different while belonging to the same higher-level kind does not entail that they do not do so because they have something in common (Polger 2004:10). If they have something in common and if that commonality explains the reason why they are put in the same category at a higher-level, and it explains how all otherwise different realizers can confer the causal powers required to fulfil a function, capacity attributed to the higher-level realized kind then the commonality in question is able to serve as a basis for a biconditional bridge-law between the higher and lower levels of description. If those parts or aspects of the distinct realizers that make them distinct don't play a role in realizing the higher-level property, then we are justified in connecting that higher-level property to the lower-level commonality.

As the argument for ontological antireductionism in the last section clearly shows, type-identity statements are possible only if biconditional bridge-laws can be established. Biconditionals are not sufficient for establishing identity, as biconditionals only express a one-one correlation between kinds or properties at different levels, but without biconditionals identity cannot be established. If it is possible to find biconditional bridge-laws connecting the higher-level realized kind to its various realizers, the road is open for establishing identity as well. In the case of isotopes, it is clear that the lower-level description of atoms allows for finding strictly definable commonalities and those commonalities also form a basis for explaining the common capacities of hydrogen atoms in forming chemical bonds with other atoms. As a result, it is motivated to accept the commonalities as a basis for biconditional bridge-laws. Furthermore, in the case of atoms and their parts scientists usually suppose an intimate compositional relation between the levels that allows for identification as well. Being a hydrogen atom is the same thing as having one proton and one electron either with or without extra neutrons in the nucleus.

So, even though the hydrogen atom is multiply realized, as Francescotti (2014) puts it, it is only *mildly multiply realized*. Mild multiple realization is accompanied by biconditional links between the different levels of description, in spite of the fact that the lower-level realizers can be distinguished from each other systematically. Francescotti provides a useful starting definition[17]. F is *mildly multiply realized* by the property class C iff[18]:

---

[17] According to this definition higher-level properties supervene on physical properties only with nomological necessity, which might not be enough for physicalism, but such kind of covariance is indeed necessary for physicalism, so I will accept Francescotti's liberal approach for the discussion of multiple realization.

[18] Here, and also later in this text 'iff' abbreviates the expression 'if and only if'.

"there are members, G and H, of C such that G ≠ H and it is nomically possible that:

an instance x of G instantiates F and an instance y of H instantiates F" (Francescotti 2014:7)

This is a good definition for multiple realization in general, but, in my view, it is not a satisfactory definition for the mild version of it. However, it is easy to amend it and Francescotti provides the key to do so when he formulates what he means by *robust multiple realizability* which, as we will see, excludes the possibility of finding commonalities between the realizers that might form the basis of a biconditional bridge-law. The definition of the mild version should be restricted to cases where there is a property J that x and y both instantiate, a property that is common only to instances of F. If there is such a common property, then F and J can be connected by a biconditional and the bridge-law F $\Leftrightarrow$ J has the potential to establish identity between those properties.

Remember the example of hydrogen. The only kind of atoms that have only one proton and one electron are hydrogen atoms. This is true of all of its isotopes, the available realizers for the kind hydrogen. When it comes to the number of neutrons in a hydrogen atom the picture changes. Hydrogen isotopes have different numbers of neutrons and there are other atoms that have the same number of neutrons as some hydrogen isotopes. So, it seems there is only one commonality among the isotopes of hydrogen that explains the reason why they share most of their chemical properties and it is understandable that nothing else has the same kind of chemical properties as, no other element shares the same commonality with them. So, a more satisfactory definition, call it MMR1, should say that:

F is **mildly multiply realized** by the property class C iff there are members, G and H, of C such that G ≠ H and it is nomologically *necessary* that: *if* an instance x of G instantiates F and an instance y of H instantiates F *there is a property J that both x and y instantiate and is common only to instances of F.*

In the last conjunct of the definition the "common only to" clause is required because without that restriction J could be a feature common to all physical things in the universe, like occupying a region of space-time. We need lower-level commonalities that have the potential to explain the capacities of the higher-level property we are interested in. To rule out overly general, empty commonalities the clause restricts the commonalities to those that apply only to those lower-level systems that realize the higher-level property F.

So far so good, however, there might be a further reason, one that Francescotti did not discuss, to be dissatisfied with the definition. Ask the following question: is there any principled reason to exclude the possibility that the lower-level commonality referred to by J is an inseparable feature of all and only the realizers of F while still being irrelevant for the realization of the realized higher-level property? At first sight, there seems to be no such reason. It also seems possible to have a perfect correlation between F and another feature that all its otherwise different realizers share, call it B, without this realizer-level feature playing any role whatsoever in doing what F does. It is also possible that there are two features, B and J, common to all and only the realizers of F, while only J has a role to play in realizing F. This might be a result of at least two scenarios. Let's see whether the amended definition is capable of dealing with them.

First, the covariation of F and B might be a coincidence. In that case it is a contingent fact of our world that B always goes together with J and F, but it is nomologically possible for J and F to occur without B, while it is at least nomologically necessary that J goes together

with F[19]. Note that both J and B are features of the lower-level system or entity that realizes F. So, B and J together with B and F are in fact metaphysically separable property pairs and their constant conjunction in our world is a mere coincidence. Therefore, while the biconditional B $\Leftrightarrow$ F is true in our world, empirical induction would be insufficient to justify the identity of B and F because there are nomologically possible worlds where F is not accompanied by B. This kind of coincidental covariation is plausibly excluded by the amended definition for mild multiple realization as it requires the nomological necessity of the covariation. The existence of a covariation between F and B in all nomologically possible worlds makes the covariation non-coincidental as it is necessitated by the fundamental laws of physics.

Let's turn to the second kind of scenario that might provide a counterexample to the amended definition of mild multiple realization. It is possible that whenever J realizes F the realizing system also produces B causally as a side-effect, and B can only be produced by other parts of a same kind of realizing system. B in such a case is like the smoke produced by a steam engine, the only difference being that smoke is not only produced by steam engines. So, the example is imperfect, but it is still instructive to spell it out in more detail. The steam engine pushes the locomotive forward by converting heat generated from coal into mechanical energy via cylinders and pistons. However, this is impossible without generating smoke as a

---

[19] This is a version of the old problem of distinguishing accidental generalizations from those based on laws of nature. Recall two classic examples that go back to Carl Hempel (1966). All gold spheres are less than a mile in diameter. All uranium spheres are less than a mile in diameter. Both statements are true, but the truth of the latter is direct consequence of the laws of physics, as there is no world with the same laws where a uranium sphere can reach a mile in diameter, while the former expresses a mere coincidence in this world (see: van Fraassen 1989:27). Even though I can't provide an example, theoretically the same kind of distinction might apply to biconditional bridge-laws as well.

by-product (B) and that smoke is a part of the realizing system until it leaves through the chimney.

In this scenario the realizing system exemplifies properties J, B and F, where F is realized by J. More precisely, certain parts and properties of the realizing system together (J) realize the powers of property F and the presence of J and F is always accompanied by B as a by-product of J at least in our world. At the same time nothing excludes the possibility that F and B covary by nomological necessity, so, for the sake of the argument, let us suppose that they do. Even in that case, B is not an operating component in the realizing system J: it does not realize F, as it is just a necessary side-effect of the operations of J, the property relevant for the realization of F. So, even if we accept that it is nomologically necessary that F and B are correlated one to one, it seems that a biconditional connecting F to B would not equate the right terms. F is necessarily accompanied by B, but it is irrelevant for the realization of F[20].

Intuitively, to exclude both the coincidence and the side-effect scenarios, it should be enough to point out that F is not realized by B or that B is not a realization-relevant part of the system that realizes F, and only lower-level properties that are relevant for the realization of the higher-level property we are interested in should be utilized in a bridge-law. This idea

---

[20] There are well-known problems in the philosophy of causation that are structurally similar. Regularity theories of causation tried to convince us that causation can be reduced to the constant conjunction of C type events with E type events, but now it is a textbook platitude that such theories have a hard time dealing with scenarios where the putative cause (C) and effect (E) events don't stand in a causal relation with each-other but are caused by a common cause (J). In such cases it is true that whenever event C is present, E is present, but still C doesn't have the capacity to influence the occurrence of E. At the same time, J is a cause of both events and can be used to manipulate their occurrence. In the case of a steam engine, the working components of the engine play an analogous role to the common cause. They produce the smoke (B) and at the same time they realize the capacities of the steam engine (F) as a whole. Even though B and F stand in a different relation to J the relation between B and F is analogous to what we have in the case of a common cause scenario as B doesn't have the capacity to influence the occurrence of F whereas intervening into the workings of J has that potential.

can be utilized in the definition of mild multiple realization if we can show that the existence of a realization relation implies more than strict covariance between properties at different levels and that extra content can be spelled out properly. Covariation is already a requirement in the last formulation of the definition. Fortunately, philosophers of science working on inter-level mechanistic explanation and realization in the sciences have come up with two criteria that together capture a sufficient condition for the existence of a realization relation between two properties. Together these are called the mutual manipulability criteria for constitutive relevance between levels (see: Craver 2007:152-157)[21].

The two criteria provide a reliable guide when it comes to the identification of the realization base for any higher-level property. As Craver (2007:162) puts it, his account is plausible as it fits well with experimental practice in the sciences, among others cognitive neuroscience, and because it is an extension of the more general notion of causal relevance.

The first criterion (MMi) says that changing certain things about the system of organized components realizing F would make property F absent. In my notation J describes those aspects, components and properties of the realizer system that together account for the powers of property F. Therefore, changing the state of the realizer system with respect to J must result in changes with respect to F.

The second criterion (MMii) says that changing the realized property F to non-F would change the state of the realizer system of organized components from one that conforms to

---

[21] More recently, extensive discussions took place in the literature concerning the validity of the mutual manipulability criteria. The first version put forward by Craver (2007) was based on the interventionist theory of causation, and it turned out that that theory has a hard time including variables (designating properties) that are not independent of each other in one consistent causal model. The original formulation went through some modifications, but withstood criticism (see: Krickel 2018). In this thesis, I will only utilize the basic idea and won't go into critical discussions of the principle.

J to an alternative state that does not. If both MMi and MMii hold true for a property pair, then it is justified to suppose that the properties tested on grounds of these criteria stand in the realization relation.

Naturally, not all lower-level properties or parts of a realizing system are relevant for the realization: only those, the modifications of which have a bearing on the presence of the realized property. Let's go back to the example of isotopes and atoms. The common feature of all isotopes of oxygen is the number of protons (J). By changing that, e.g. taking a proton away, we can change oxygen (F) into a different element, but one cannot change F, that something is an oxygen atom, by adding a neutron to the structure. However, by changing an oxygen atom into a nitrogen we would change the number of protons as well.

At least one of the two criteria fail to apply to the B and F property pair in both cases discussed above. If B cooccurs with F only coincidentally, not realizing F, then in line with MMi there are no possible changes in B that would result in a change to F as B is irrelevant for the realization of F. Similarly, in line with MMii, F can be changed to non-F leaving B fully intact, as B is not relevant for the realization of F. In this case, both criteria fail to hold which shows that connecting F and B by a bridge-law would be improper.

If B co-occurs with F because it is a side-effect of the operations of the system realizing F, then according to MMi there are no changes in B that would result in a change to F and therefore B is irrelevant for the realization of F. Whilst, in agreement with MMii, F cannot be changed to non-F leaving B intact, as B is a by-product of the operations of the realizing system in a state described by J and J realizes F. So, in this case only criterion MMi fails to hold, but it this is enough to show that a biconditional bridge-law connecting B to F would be a mistake.

So, to make the definition of mild multiple realization secure it should be amended in light of the above discussion. Let us call it MMR2:

F is **mildly multiply realized** by the property class C iff there are members, G and H, of C such that G ≠ H and it is nomologically necessary that: if an instance x of G instantiates F and an instance y of H instantiates F, there is a property J that both x and y instantiate and is common only to instances of F and *the mutual manipulability criteria (MMi & MMii) hold true for the J and F property pair.*

Now that we have a working definition for mild multiple realizability we can see that this is not the kind of multiple realizability that licences ontological antireductionism. Ontological antireductionism requires the failure of biconditional bridge-laws, MMR2 allows for them. If the commonality identified among realizers of the same higher-level property accounts for the unified treatment of the realizers at higher levels, then the antireductionist promises Fodor and Putnam offered us are in danger.

It is useful to clarify what exactly belongs to a realizer system. The lower-level commonality identified in realizer systems can be called the core realizer. But the core is not enough for the realization of the higher-level property and we usually think that realization should explain the metaphysical dependence of the realized on its realizer. Think of C-fibre firings as realizers of pain. The presence of those fibres in themselves is insufficient for the presence of pain. One can put those fibres into medical saline water in a Petri dish where they might survive for a while, but under such circumstances they won't realize pain states. The core realizer usually requires a lot more to be able to serve its function (see Shoemaker 2007:21). The total realizer[22] includes at least the core realizer and other parts with which the core realizer interacts functionally. Some philosophers argue that if the total realizer is

---

[22] The distinction between the core realizer and the total realizer that goes back to Shoemaker (1981) is widely used in the literature.

intended to be a metaphysically sufficient condition, then the total realization base is wider than the salient system, e.g. the body of an organism, and includes certain background or environmental conditions without which, the higher-level property would not be realized as the core realizer would not be operational. Energetically open systems, like organisms are the best examples for such wide total realizers (see: Kirchhoff 2015)[23]. If the total realizer is sufficient for the realized property, the core realizer is a necessary part. If a property is multiply realized in a robust sense, when there is no common core feature in all realizers, Mackie's (1974) analysis of causation in terms of INUS conditions is helpful to analyse the situation. All the different core realizers can be analysed as insufficient but necessary conditions as part of unnecessary but sufficient systems of conditions.

1.4.2   Robust multiple realization

Fortunately for those who aim to argue for the autonomy of the special sciences, there is the possibility of what Francescotti (2014) calls robust multiple realizability. According to his definition, which I will call RMR, it obtains when there are no exclusive commonalities among the realizers of a higher-level property. *Robust multiple realization* obtains for higher-level property F and the class of realizer properties C iff:

> „there are members, G and H, of C such that G ≠ H and it is nomically possible that:
>
> an instance x of G instantiates F and an instance y of H instantiates F, and there is
>
> no property J of C that x and y both instantiate and that is common only to
>
> instances of F" (Francescotti 2014:9, *notation changed to match mine*)

---

[23] For such systems constant external energy flow is vital, without that they cannot maintain their integrity. They are energetically coupled with their environment, so the core realizer together with other working components of the organism would still be insufficient for the realized property without certain features of the environment.

There is an insight from the discussion of mild multiple realization one should utilize here to amend this definition. Even though it is highly unlikely, it is possible that all realizers of F have something in common, but that commonality is irrelevant for the realization of F. In light of this possibility I would amend the definition as follows. RMR1:

F is **robustly multiply realized** by the class of realizer properties C iff there are members, G and H, of C such that G ≠ H and it is nomologically *necessary* that: *if* an instance x of G instantiates F and an instance y of H instantiates F, and there is a property J of C that both x and y instantiate and that is common only to instances of F *the mutual manipulability criteria (MMi & MMii) fails to be true for the J and F property pair.*

Robust multiple realization unquestionably justifies serious ontological antireductionist conclusions, but it is controversial that there are good examples of it. According to one influential camp of authors (Polger & Shapiro 2016), robust (or interesting) multiple realization is hard to find, while others (Aizawa & Gillett 2009) think that choosing the right framework it is easier to find examples. I will get back to this disagreement in the next section. Note, that if a property is robustly multiply realized, then there is no unified way to predict and explain the higher-level property in question based on lower-level information. This also means that it is impossible for the different realizers of the same higher-level property to be in the same kind of physical state. If the different realizers shared the same kind of lower-level state one could find a commonality among them that would allow for the use of a biconditional bridge-law.

There is at least one straightforward example in the literature that seems to fit the bill: the case of temperature. Nagel (1961) used the case of temperature as his paradigm

example for proper reduction, but later it became a much-debated case (see: Enç 1983; Bickle 1998). Let us take a closer look at this example. According to classical thermodynamics, temperature is only assigned to objects or volumes of gases in thermodynamic equilibrium. The temperature is identified to, is a measure of the motion energy of particles, more precisely the mean kinetic energy of the particles that comprise the volume of gas or object in question. However, the kinds of motion relevant for temperature are different for different kinds of matter, and the calculation of temperature from lower-level information requires different mathematical equations for different kinds of matter. This is the reason why philosophers in the 1980s started to cite temperature as an example of a multiply realized property. Bickle (2019: section 2.2) summarizes the example in a short passage, and unfortunately his reference authors like Enç (1983) do the same. It is worth providing a description that is a bit more detailed.

Temperature in an ideal gas is the mean molecular translational kinetic energy. This theory is well-known. But there is a difference even between how the temperature of different kinds of gases can be properly accounted for, as the above definition is only applicable to monoatomic, more precisely to noble gases. In gases that consist of diatomic molecules or of more complex molecules, rotational movement and vibrations of the compound structures also have to be accounted for. These molecules have more degrees of freedom, so they have more ways to absorb energy.

At a first approximation, temperature in a solid is the mean maximal molecular kinetic energy of the constituents. The molecules of a solid are bound in a lattice structure and restricted to vibratory motions. In actual fact the proper physical description utilizes so-called phonons (see: Fai & Wysin 2012:389-390). A phonon is a collective excitation in a periodic arrangement of particles in condensed matter, but the phonon-based description requires a

quantum mechanical treatment, quantization of the vibrations. So, temperature in solids is something entirely different from temperature in a noble gas.

Just to take a glance at further alternatives, radiation has temperature as it can be in thermodynamic equilibrium with a solid or a gas, and it is also something entirely different compared to solids or gases. There are no atomic or molecular constituents in radiation, its components are energetically coupled electromagnetic waves and for that reason that phase is dominated by electromagnetic interactions (see: Sklar 1993:351-353, 2015: section 6).

It is instructive to think through the example of temperature using the basic distinction between mild and robust multiple realization. Mild multiple realization requires less than what was discussed in connection to temperature above. Many would think that temperature in a gas is multiply realized by all the different micro-configurations that realize the same temperature value. This implies extremely small differences between realizers that are taken to be distinct realizers. According to those who think this way, even the smallest compositional difference between realizer systems is enough to reference them as distinct realizers. But it is also true that there is something invariant in all those different samples of gas, this is the reason why scientists could produce a bridge-law connecting temperature to an overall property of the constituents of a gas sample. The commonality is expressed by the mean (translational, etc.) kinetic energy of the constituent atoms. From this point of view, temperature in a volume of gas, temperature restricted to volumes of gas, is only mildly multiply realized. However, temperature as such in an unrestricted sense is robustly multiply realized, as it is captured by different basic models, explanatory mechanism in different kinds, phases of matter. The monoatomic gas with freely moving particles has different kinds of mechanisms to absorb energy compared to a solid with its phonon vibrations.

A more entrenched example of robust multiple realization is the case of human brains and computer brains, which one finds in almost all textbooks. Almost everybody agrees that if intelligent computers could be made, they would constitute an essentially different kind of realizer for intelligence. Only one condition needs to be satisfied for this to be the case. If the basic procedures by which a human brain and a computer carries out its operations are seriously different, then the view is indeed plausible.

1.4.3   Two mainstream accounts of multiple realization

Below, I would like to introduce two basic concepts of multiple realization that dominate the literature in the last decade. I won't to argue either for or against these views. Instead of choosing between them where multiple realization becomes important in my discussion of causal exclusion, I will investigate the problem from the point of view of both theories.

1.4.3.1   A flat view of multiple realization

The first concept relies on a functional account of realization, where the higher-level function (F) of an entity (x) is realized by a lower-level causal mechanism (CM). F is multiply realized if and only if it is realizable by different kinds of underlying causal mechanisms (CM1, CM2, etc.) (see: Polger & Shapiro 2016; Shapiro 2004). A different way of putting this is to say that property F is multiply realizable if there exist at least two distinct functional analyses of F[24].

---

[24] Conditions for the multiple realization of kind F in higher-level taxonomic system S1 by lower-level scientific taxonomic system S2 (see: Polger & Shapiro 2016:67-77):

(1) Ps and Rs are of the same kind in higher-level scientific taxonomic system S1. (higher-level sameness)

(2) Ps and Rs are of different kinds in lower-level scientific taxonomic system S2. (lower-level difference)

(3) The factors that lead the Ps and Rs to be differently classified by S2 are among those that lead them to be commonly classified by S1 (demands that lower-level kinds should be different in a realization relevant manner)

(4) the relevant S2-variation between Ps and Rs is distinct from the S1 intra-kind variation between As and Bs. (demands that variation in the realized kind is distinct from variation between realizer kinds)

The view is called "flat" as the realized and the realizer property are attributed to the very same individual entity (see: Gillett 2003).

According to this view, multiple realization is a phenomenon distinct from mere variation. There is a lot of variation in nature and between levels of nature, but not all of that variation is relevant for our taxonomic systems. On this account, mild multiple realization points to uninteresting variability in realization. In such cases, the core realizer is always the same in terms of the lower-level taxonomy, so biconditional bridge-laws can be established as there is one unified way of explaining the higher-level property by a lower-level mechanism. However, robust multiple realization is taken seriously by this account. In such cases the core realizer of the realized kind cannot be defined in a unified manner.

Polger & Shapiro (2016:64) argue that "the question of multiple realization is a question about actual sciences, and it is always specific and contrastive". This means, that the question concerning multiple realizability should be relativized to available levels of description. So, a mental property can be singularly realized with respect to the neurological level, but multiply realized somewhere below that level[25].

Figure 1.4-1

Let me show the most important features of the theory on a toy example Shapiro (2000) introduced, and that is recited by many in the literature. Consider the following kind of variation among corkscrews: (a) a waiter's corkscrew and a winged corkscrew (see: Figure 1.4-1), (b) two waiter's corkscrews, where one is made of aluminium and the other is made of iron, (c) two waiter's corkscrews

---

[25] There is a hot debate among theorists on whether the realization relation is transitive or not. Polger (2008) denies it, Gillett & Aizawa (2009) object, but this problem has no real bearing on my interests in the thesis.

that differ only in colour. The kind corkscrew, supposing it is a real kind, is (robustly) multiply realized by waiter's and winged corkscrews (a), because they realize the same function of removing corks by different mechanisms. The waiter's version works by a lever mechanism, the winged version works by a rack and pinion mechanism. In case (b) there is variation between the two realizers, but they belong to the same mechanistic kind. In the terminology introduced earlier, this would be a case of mild multiple realization, as the items share core realizer features. As Shapiro (2000:645) explains, relative to the function in "a corkscrew, rigidity screens off the differences between steel and aluminum". In the last case (c), the difference in colour is simply irrelevant for the realization of the function, so the kind corkscrew is univocally realized by the two items. To highlight the differences, in the next section, I will show how the other conceptual framework deals with the same examples.

1.4.3.2   The dimensioned view of multiple realization

The second concept of multiple realization relies on a compositional account of realization. According to this view, higher-level property F of an entity x is realized by a set of lower-level properties and relations between x's components. F is multiply realized, when at least two, non-identical sets of property instances, {P1-Pn} and {P*1-P*n}, at the same level of composition determine another property instance of higher-level property F (see: Aizawa & Gillett 2009, Aizawa 2018)[26].

---

[26] The precise definition of multiple realization according to the dimensioned view: "A property F is multiply realized if and only if (i) under condition $, an individual x has an instance of property F in virtue of the powers contributed by instances of properties/relations P1-Pn to x, or x 's constituents, but not vice versa; (ii) under condition $* (which may or may not be identical to $), an individual x* (which may or may not be identical to x) has an instance of a property F in virtue of the powers contributed by instances of properties/relations P*1-P*m to x* or x*'s constituents, but not vice versa; (iii) P1-Pn ≠ P*1-P*m and (iv), under conditions $ and $*, P1-Pn and P*1-P*m are at the same scientific level of properties" (Aizawa & Gillett 2009:188, *notation changed to bring it closer to my usual notations in the thesis*).

The view is called dimensioned because the component parts of the whole that bears property F, the lower-level entities, their properties and relations are said to constitute and realize the whole and its properties. Gillett rejects the view according to which realized and realizer properties belong to the same subject, the kind of view he dubbed the "flat view" (Gillett 2003). This feature characterizes all present-day alternatives of the composition-based view. Along with all flat views, Gillett (2010) rejects the notion of a structural property used implicitly or explicitly by many philosophers to describe the complex lower-level state that is identified as the relevant realizer. On the composition-based view it comes out firstly, as an unnecessary addition to the ontology of the system and secondly, as an ontological chimera that includes features of both the higher and lower-level ontology.
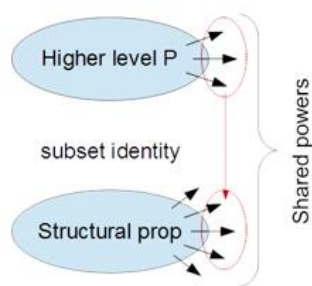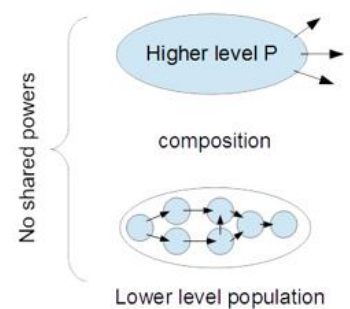
Figure 1.4-2

Figure 1.4-3

Gillett's direct criticism was pointed towards a specific version of the flat view, the subset view of realization (see: Figure 1.4-3 and section 1.3.3 for a summary of the view), but the same basic insight holds for the mechanism based functionalist view as well. We simply need to replace the notion of a structural property with the notion of a lower-level mechanistic kind Shapiro (2004) prefers.

For the dimensioned view, all kinds of inter-level identities are impossible. The realization relation is an asymmetric, one-to-many composition relation between realms of qualitatively distinct properties (see: Figure 1.4-2) similarly to the case of temperature and kinetic energy as we already saw earlier (section 1.2.2). Naturally, not all constituents of the whole that bears the realized property are relevant for realization. In many cases, some constituents are parts of the core realizer, some constituents are not.

According to the dimensioned view, variation in the composition of a higher-level property provides a case for real multiple realization. Almost any level of compositional difference between realizer systems is taken to be relevant. This is because on this view ''realization is a transitive relation'' (Aizawa & Gillett 2009:194) as composition is a transitive relation[27]. Let's take the example of mechanical hardness. On this account even the small differences in the metal lattice of two samples of the same original cast iron bar count as a case of multiple realization for the same hardness property, because the exact configuration of the constituent particles and the exact positions of point and line dislocations in the metal lattice are different. This is maybe the most frequently criticized aspect of the theory[28]. Here, I will take it as an accepted theoretical option.

Let us evaluate the example of corkscrews in the context of the Dimensioned view. For case (a) the verdict is unambiguous, waiter's and winged corkscrews multiply realize corkscrews. However, I should highlight that in this framework the realizer is not the winged corkscrew as whole, instead, the pointed helix, the arms, and the rack, and their properties together realize a variant of corkscrew. The relevant properties of these parts are qualitatively different from the realized property (capacity for cork removal). In scenario (b) where there is a difference in the material composition of the two items, there is real multiple realization again, as the items are made of different metals, even though mechanically they are the same. On the compositional view, there is no principled difference between mild and robust cases of multiple realization or in a different parlance, between uninteresting variation and multiple

---

[27] If one takes the realization relation as a species of determination relation the assumption is even more natural.

[28] In this view multiple realization is way more ubiquitous than on the flat view. Critics like Polger and Shapiro say that it is correspondingly less obvious that it is metaphysically significant as the original question by Fodor was relativized to scientific kinds at this or that level of description.

realization proper. The only case the dimensioned view dismisses is (c) where the otherwise identical items have different colours, because the variation between the realizer instances is irrelevant for the realized property in question.

1.5   Concluding remarks

This chapter aimed mainly at setting the stage for later investigations. It introduced the problem of mental causation. It summarized the objections to the identity theory and the rise of non-reductive physicalism motivated by the theory of multiple realization. It summarized the ways in which one might define physicalism in the absence of inter-level identities. It introduced the most prominent theories of multiple realization.

Chapter 2 turns to the analyses of Jaegwon Kim's causal exclusion argument. Because of the consequences Kim drew from his premises it is usually considered to be an argument for type physicalism. Kim tries to convince us that non-reductive physicalism is not tenable, and we should fall back on some restricted version of the identity theory. The chapter develops detailed arguments for two crucial premises of the exclusion argument; however, it identifies a hidden premise concerning the interpretation of the concept of causation and argues that critics of Kim's project are justified in preferring a difference-making account of causation over productive accounts Kim favoured in has later writings.

Chapter 4 and chapter 5 develop a critical discussion of Menzies' reformulated, non-reductivism friendly approach to the causal exclusion problem, one prominent reformulation that is based on a difference-making notion of causation. The differences between different notions of realization and the flat and dimensioned views of multiple realization will become more relevant for those investigations.

Chapter 2: Kim's causal exclusion argument against non-reductivism

2.1 Causation, conditions, events and powers

Before discussing the causal exclusion argument in detail certain general clarifications concerning causes, effects and causation in general are due. Here, I will only provide a broad-brush characterization of causation, and only commit myself to basic things that would be true in most theories of causation. I will start with preliminaries concerning causal relata. There are convincing arguments for the view that causation is basically a token-level relation between individuals and only derivatively a relation between types (see: Hausman 2005, Woodward 2003). Therefore, as a default, I will take type causal statements to be results of generalizations over token causal relations.

2.1.1 The relata of causation

It is natural to think that a short-circuit caused a fire or that a stone broke a window. In either case we talk about a particular event or an object. Of course, we can talk about causes and effects in terms of types as well. But it would be weird to think that the object type stone caused the window breaking or that the event type short-circuit caused the fire. Surely, it was a particular stone following a particular trajectory that resulted in a particular breaking. We do say things like 'short-circuits cause fires', but by that a firefighter would mean that particular tokens of short-circuits cause particular token fires, not that these event types are causing each other.

It would be similarly weird to say that properties are the causes of other properties. A nominalist about properties thinks that properties are sets of concrete objects resembling each other. It is natural for a concrete object, a member of that set, to cause something, but sets are unlikely candidate causes. The universalist account of properties faces different

problems. If it is taken to be a universal, the causally relevant property of the short-circuit has multiple locations, it is located wherever it is instantiated. But then the short-circuit in my cellar that does no harm at the moment might cause a fire somewhere on the other side of the planet. This is definitely not how we think about causes and effects.

Most philosophers of causation believe not only that it makes sense to talk about temporal relations between causes and effects or spatio-temporal connectedness between them, but also that this is necessary to make sense of everyday and also most of scientific causal talk. So, it seems, that a causal relation takes place between particulars. Causes and effects are bound to exist in space and time. Causes bring about effects and both of them are situated somewhere and at a certain point in time. Also, the relevant properties exemplified by them are not abstract particulars with multiple spatio-temporal locations, but property instances or tropes.

Suppose again that causation is a relation among abstract properties, not located in time and space. If it were, how could we talk about causal precedence in a meaningful manner? Causal precedence is important for our practice of finding causes and effects. In most contexts[29] we think that causes precede their effects. When searching for causes in a large set of statistical data that describes correlations between variables it is a huge help to have temporal information. This is because we make two assumptions. First, the cause can bring the effect about but not the other way around. And second, in the temporal dimension the cause comes first, and the effect follows. Knowing what type of event came earlier and

---

[29] I won't consider cases like that of gravity under the Newtonian interpretation, or the relationship between temperature and pressure in closed systems, because in cases like these the relation between the variables is (1) not asymmetrical and (2) it is simultaneous. Most of the causation literature has an explicit agreement not to include such relations into the set of causal relations for at least the two mentioned reasons.

what came later in the mentioned dataset helps us by eliminating bad candidates for causal relations. From this it follows that our conviction concerning the asymmetry of the causal relation is intimately connected to our conviction that causes and effects are localized particulars.

2.1.2   Causes and conditions

An important, but less discussed factor that plays a role in understanding causation is the set of background conditions under which a certain causal relation holds. In some theories this is made more explicit, in others the role of background circumstances is straightforward but less emphasized. In his classic book, Mackie (1974) talked about this in terms of causal fields. Take the example where a short-circuit creates fire. When making a singular causal statement like this the presence of oxygen and combustible material is assumed. The latter conditions are part of the causal field in which the short-circuit starts the fire. Every singular causal relation between events is relative to a similar field of background conditions. This means that something that counts as a cause with respect to one field may not be a cause in relation to another. In a causal field without oxygen, no short-circuit can start a fire.

So, causal statements are made holding certain background conditions fixed. In case we are interested in what caused the fire (compared to having no fire) in the house we hold fixed everything that seems to be common among the two situations. By pushing the presence of oxygen and combustible material to the background we get an answer.

In more contemporary theories, like in Woodward's interventionist framework, background conditions are taken to be definitive of causal relations. For Woodward (2000, 2003) causal relations are bound to particular background conditions under which the relation holds. In his later works, he introduced more fine-grained traits to characterize causal

relations on grounds of their sensitivity to changes in background conditions (Woodward 2006, 2010). This is important as such characterization provides explanation for our tendency to reject or to accept causal relations as good explainers. If a causal relation is fragile, meaning that it holds only under a fairly narrow set of background circumstances, we are more likely to say that the explanation that refers to it is unsatisfactory. Whereas robust causal relations that hold under a wide range of background circumstances are more attractive explainers for us. To sum up, to conceptualize a causal relation between particulars like events properly we also need to attend to the background conditions under which the relation holds. Hausman (2005) characterises causal relations in accordance with everything said so far:

"a causes b if and only if (1) a and b occur and (2) for some properties A and B instantiated by a and b and *some kind of circumstances K* instantiated by *the circumstances in which a and b occur*, instantiations of A in K that bear the proper spatio-temporal relations to instantiations of B cause those instantiations of B's." Hausman (2005:44, my italics)

According to Hausman this definition expresses a general condition for causal relations to hold, quite independently of what particular theory one subscribes to. It emphasizes that causes and effects are locally instantiated in space-time and also the importance of background conditions ("circumstances" in the quote above) for causal relations to hold. Even more than that, it highlights that all token causal relations are instances of causal generalizations (where a generalization says: instances of type A cause instances of type B under circumstances type K). Hausman equates causal generalizations with type causal statements which are generalizations over actual and possible token causal

relations. This means that both token and type causal statements talk about or refer to token causal relations.

2.1.3 Causation, properties and powers

We already established that causes and effects are particulars like objects and events. These particulars have causal powers in virtue of the properties they instantiate. This is what allows them to be potential causes. As I said above, properties are not causes, therefore, it is not a property as such that bestows those powers on the particular in question, it is the property instance. This description in terms of property instances is useful as it helps us single out the causally relevant aspects of causes. The stone that broke the window instantiates a lot of properties. It has a certain mass, opacity, electrical conductivity, velocity, hardness, probably it is brittle and has a particular colour as well, but most of these properties like colour, brittleness, opacity, conductivity are irrelevant with respect to the window breaking. Mass, velocity and hardness are those properties that bestow the powers necessary for breaking the window.

There are a wide range of available views on what causation amounts to and I will discuss various views on the topic throughout the thesis. But I think most people would agree that causes do what they do by manifesting powers conferred by the properties they instantiate. If a particular has a power to do something, then it is a potential cause. If some particular manifests a power, then it becomes a cause. Whenever that happens the property instance that grants the power is considered to be causally efficacious. The short-circuit has the power to start a fire even when there is no oxygen in the building where it takes place, but under the right circumstances it manifests its power to light the fire. In that case the relevant, causally efficacious property of the short-circuit is the heat generated.

In what follows, when talking about causation I aim to mean token causation, if not specified otherwise. By type causation I will mean causal generalizations over token causal relations. I will take the causal relata to be token events, instantiating properties that confer powers to do things. Following common practice, for brevity's sake throughout the thesis, I will talk about properties instead of property instances, but by that I aim to mean property instances if not indicated otherwise.

2.2   A summary of Kim's exclusion argument

The causal exclusion argument is presented as a reductio against the irreducibility of mental or any other higher-level properties. This has a further consequence. It puts the popular non-reductive physicalist position into jeopardy. All who would like to call themselves physicalists, non-reductionists and reductionists alike, must accept the starting points of the argument, and if they do, it seems they are forced to follow Kim towards certain inconvenient consequences.

In most construals the argument has six starting points. Each point has independent support. The argument aims to show that the starting points are inconsistent with each other and one should reject at least one of them to preserve the others. According to Kim all physicalists are committed to the view that properties are either physical or determined by basic physical properties. So, they would agree on the following:

I.   **Supervenience** (of mental properties): All mental properties supervene on basic physical properties.

This means at least two things. First, whenever we have a mental property M instantiated by an entity e necessarily there is an underlying physical property P accompanying it. Second, that mental property M cannot be changed without changing the subvenient property P (see section 1.3.1). Physicalists would also agree, that physical events are caused by other physical events:

II.   **Causal closure** (of the physical realm): Every physical event has a sufficient physical cause. So, physical causation is complete.

There are further starting points that not only the physicalist is bound to accept. One very important premise among these is the so-called exclusion principle. It states that events

are not caused twice over or at least only on the rarest of occasions. This premise is based on the no (systematic) overdetermination premise. As we will see later in more detail, the literature on causation considers overdetermination of events an important case of causation, although it is considered to be a rare phenomenon. The kind of overdetermination the mental causation debate is interested in is of the inter-level kind. In this kind of case properties at different levels of reality, e.g. physical and mental, are considered to be simultaneous causes of the same outcome.

However, according to most physicalists there is some intimate relation connecting the properties at lower and higher levels, be it identity or realization. The existence of such a relation between levels makes it less plausible to think that these properties can work as independent causes for the same effect. No overdetermination and the exclusion principle that follows from it is regarded by Kim (2005:51) as "virtually an analytic truth with not much content", but even he acknowledges that others have raised issues concerning its validity. For this reason, section 2.4 will provide detailed motivation to disregard the possibility of serious overdetermination and at the same time for the exclusion principle.

III. **No overdetermination**: Mental causes do not overdetermine their effects.

IV. **Exclusion Principle**: If an event e has a sufficient cause c then no event distinct from c is a cause of e. "No single event can have more than one sufficient cause occurring at any given time — unless it is a genuine case of overdetermination." (Kim 2005:42)

Furthermore, we are all convinced that our mental states, beliefs and desires are difference-makers in the physical world, that they are real, causally efficacious properties. Without that moral decisions, exchanges of values or even market exchanges would be

hopelessly meaningless. We would become mere shadows of something behind the scene of our lives. As non-reductive physicalists, motivated by the arguments of Putnam, Fodor and their followers we should also believe that mental properties and physical properties are distinct existences endowed with causal powers of their own.

For non-reductivists mental properties are numerically distinct from physical properties and the main support for the distinctness claim comes in the form of the multiple realization argument. As we saw, in its early formulations multiple realization was put forward as a challenge to reductive physicalism. Fodor aimed to challenge Nagelian reduction and the idea of identity statements connecting higher-level properties to lower-level properties. Putnam formulated a similar argument aiming to refute the mind-brain identity theory in the philosophy of mind (see section 1.2). According to them the presence of different realizers in this world and other possible worlds prevents identification.

V. **Mental causation** (or later higher-level causation): mental properties are real and causally efficacious.

VI. **Distinctness** (mental properties are not identical to physical properties): mental properties are numerically distinct from physical properties.

Kim's exclusion argument runs as follows: Supervenience, Causal closure, No overdetermination, the Exclusion principle and Mental causation are not consistent with Distinctness. One of the premises has to give way. Supervenience and Causal closure have to stay, as both starting points are among the basic commitments of physicalism. There is no good reason to think that there is systematic inter-level overdetermination delivered by supervenient and subvenient properties together. So, the exclusion principle kicks in and we are forced to choose between lower and higher-level causes. It would be too high a price to

make mental causation a mere shadow of existence, but because of completeness we must choose the subvenient cause. Therefore, according to Kim (1998, 2005), to be able to save mental causation Distinctness has to go.

Kim developed two successive versions of the exclusion argument. His aim was to show that mental to physical causation is not possible as mental events are pre-empted by the physical events realizing them. He developed another argument that builds on the exclusion argument to show that even mental to mental causation is problematic for the same basic reasons. I take it along with many philosophers that the exclusion argument generalizes, and it has consequences beyond the mental causation debate. It is an argument that applies to any multiply realized higher-level property that supervenes on the physical. So, in many cases I will talk about mental and other realized properties interchangeably.

Later in this chapter, I will scrutinize some of the crucial premises to see how serious their motivations are. But before that, let me go through the two versions of the exclusion argument to show how they are supposed to work.

2.2.1   The 1st version of the exclusion argument

The first version of the exclusion argument (Kim 1998) starts with a scenario where a mental event, an instance of a mental property M1, causes a physical event, an instance of a physical property P2, let us suppose it is a behavioural outcome. Naturally, M1 has a physical supervenience base P1. Kim contends that one should accept that P1 is also a cause of P2 regardless of one's interpretation of causation as sufficiency or in terms of counterfactuals.

The sufficiency approach suggests that P1 is a cause of P2 because P1 is sufficient for P2. This sufficiency for P2 is based on supervenience: M1 supervenes on P1 and we already supposed that M1 is sufficient for P2. Kim claims that P1 is a cause of P2 on a counterfactual account of causation as well. He thinks that based on supervenience it is a safe assumption that if P1 had not occurred, then M1 would not have occurred either. M1 brings P2 about as without the presence of M1 P2 would not have occurred. So, P1 is a cause of P2.
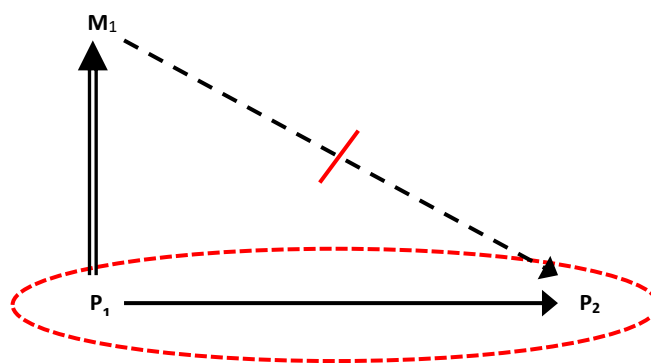
Figure 2.2-1

Under both interpretations we have two causal statements to choose from. M1 causes P2, but P1 causes P2 as well. Because of the multiple realizability of M1, we can assume the distinctness of mental properties from physical properties, so P2 has two putative causes. M1 supervenes on P1 and both M1 and P1 cause P2. The exclusion principle tells us that one of the putative causes must be excluded. According to the causal closure principle it is not an option to exclude P1. P2 has a sufficient cause in the form of M1 but the closure principle dictates that if P2 has a sufficient cause then it has a physical sufficient cause which would be

P1 (see: Figure 2.2-1[30]). So, M1 has to give way. Getting rid of M1 results in P1 being the only cause of P2 and M1 a causally empty property.

The argument started out with the assumption that M1 Causes P1 and ended up contradicting that assumption. This clearly shows that the premises (I-VI) of the argument are inconsistent.

2.2.2 The 2nd version of the exclusion argument

This second version of the argument that appears in Kim (2005) is only slightly different form the first version but in a quite important way. It relies more heavily on the causal closure
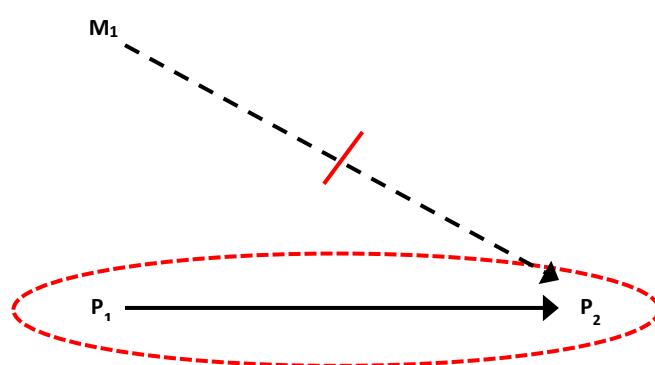
Figure 2.2-2

principle. It starts with the same assumption that M1 causes P2 and then immediately turns to the closure principle according to which every physical effect has a sufficient physical cause, so P2 has to have a sufficient physical cause in the form of P1. Again, we have two competing causes M1 and P1. We know that M1 is distinct from P1. By the exclusion principle we must dump one of the two causes. The causal closure principle forces us to choose P1 over M1 (see: Figure 2.2-2). And the result is the same as in the former version of the argument.

Notice that the supervenience relation between M1 and P1 wasn't utilized in this version of the argument. Exclusion took place solely based on the closure principle. The supervenience relation wasn't denied, the argument is compatible with it, but as it wasn't

---

[30] On all such figures single tail arrows depict causal connections, arrows with dashed tails signify causes that are made empty by another causal connection and two tailed arrows refer to the supervenience relation.

used as an assumption even a failure of supervenience would be compatible with it. This means that the failure or rejection of the supervenience thesis wouldn't be effective against this version of the argument. This version relies solely on causal closure as an anchor into the physical realm. This highlights that from the point of view of Kim's exclusion argument Closure is crucial. It is the principle that maintains the solid connection to physicalism and on the other hand it is the principle that forces the choice between the candidate causes.

2.2.3 Excluding higher-level causation as such

The exclusion argument can be applied also against mental to mental or higher-level causation in general. Imagine that M1 causes M2. In the context of mental causation there are many such examples. Suppose that my belief that Europe is falling apart politically (M1),
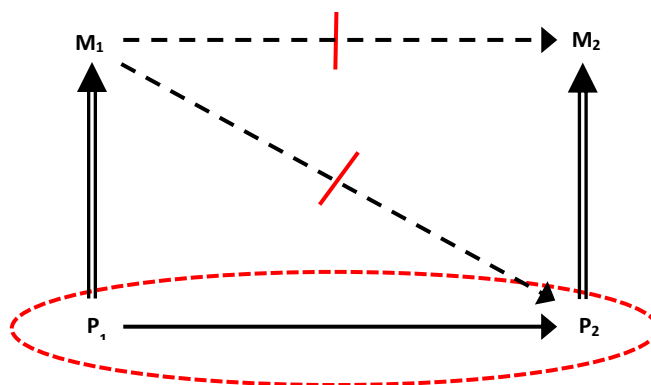
Figure 2.2-3

makes me think that we should expect a second financial crisis (M2). Both mental properties have a physical supervenience base P1 and P2 respectively. Why is M2 instantiated in this scenario? Is it because M1 caused it directly? Or is it because M1 caused P2, the supervenience base of M2? Kim argues in the following way:

> "To cause a supervenient property to be instantiated, you must cause its base property (or one of its base properties) to be instantiated. To relieve a headache, you take aspirin: that is, you causally intervene in the brain process on which the headache supervenes." (Kim 1998:42)

The message is clear: whatever happens, higher-level supervenient properties are caused by bringing about their subvenient properties and so we can go back to either version of the exclusion argument and close the argument straightforwardly in favour of P1, the sufficient physical cause of P2 (see: Figure 2.2-3).

Another possible way to build the argument in the spirit of Kim's exclusion arguments would be the following. What is responsible for the instantiation of M2? Suppose it is M1 either directly, or indirectly by causing P2. However, M1 also has a supervenience base P1 that is a sufficient cause for P2 and by causing P2, P1 would cause the instantiation of M2. This is where the exclusion principle kicks in. P2 cannot be overdetermined, so we are forced to choose between M1 and P1. Closure forces us further to choose P1 among the two. As P2 is already caused by P1 and M2 supervenes on P2 there is no further causal role left for M1 to play. When a physical property causes a mental property, the latter is caused indirectly by bringing its supervenience base about. There is no autonomous role for higher-level properties. Causation only takes place between P1 and P2. Devoid of autonomous causal powers, M1 and M2 are mere epiphenomenal shadows of their realizers. So, again we started with the assumption that there is mental to mental causation but relying on the six premises (I-VI in the intro to section 2.2) of the exclusion argument Kim claims to be well-motivated we arrived at a contradiction.

2.2.4 Kim's solution to the paradox of the exclusion argument

Kim's exclusion argument was presented as a reductio against the Distinctness premise, against non-reductive physicalism. In case one is committed to physicalism (Closure and Supervenience), and aims to preserve mental causation, this seems to be the only path to follow. Multiple realizability leaves us with a dilemma: either we should deny physicalism because there are efficacious non-physical properties irreducible to lower-level properties, or higher-level properties are inefficacious, and so epiphenomenal. The latter option is what we wanted to avoid in the first place, the former requires us to give up on physicalism. Therefore, as physicalists we are forced to find a way to reconcile mental causation with physicalism.

> "What we want — at least, what some of us are looking for — is a philosophical account of *how it* [mental causation] *can be real in light of other principles and truths that seem to be forced upon us*." (Kim 1998:72, my italics)

In Kim's view, to preserve mental causation via some kind of reduction offers the best deal. But before accepting such a deal he has to offer us some satisfactory answer that can accommodate our convictions concerning Distinctness as well. We can't just give up on Distinctness without a satisfactory bargain. Epiphenomenalism can't work, mental causation should be preserved, but if one lacks the resources to properly argue against multiple realization then some explanation or an error theory of multiple realizability claims and our intuitions concerning Distinctness is required. Or, in other words, there should be a satisfactory answer to the question, what makes special science statements, like those of psychology or even statements of macro-level physics true or useful?

What is nice about Kim's approach to the problem of mental causation is that he has an answer to this question, whether we like it or not. He accepts the challenge from Fodor

and Putnam and agrees that it is not possible to find a single broadly physical, complex neural property shared by all organisms experiencing pain. But instead of accepting that pain is a realization-independent mental property he went for the elimination of this generic, unified property and replaced it with so-called local, species-specific mental properties (Kim 1992). Such properties are reducible to their neural bases.

So, there is pain-in-humans, pain-in-foxes, pain-in-magpies, but there is no species or structure independent pain as such. As pain-in-humans is identical to some specific kind of neural realizer (probably a set of satisfactorily homogeneous particular realizers) the causal powers of pain-in-humans are identical with those of the realizer. The same goes for all such species-specific mental properties.

So, on the one hand Kim is a reductionist and on the other hand a revisionist eliminativist as well, as he thinks that there is no such thing as a generic mental property pain. It is fair to ask at this point, where is the explanation for our intuition concerning Distinctness? His answer is that even though there is no generic pain property behind it, we have a generic concept of pain. This concept is useful as it encompasses a group of phenomena that stand in some kind of family resemblance relationship to each other, but closer attention shows us that the different species-specific pains are causally heterogeneous to a large extent[31]. This view is a consequence of another view he holds. On Kim's account of higher-level physically realized properties they inherit their causal powers form their realizers. The causal powers of a higher-level property instance are the same as the causal powers of its realizer. As

---

[31] Even though the idea is applied in a piecemeal manner this approach is close to Heil's, according to which the mental-physical disctinction is not onologically deep, as it is a difference in conception only. He emphasises the distinction between predicates and properties (Heil 2013). Predicates might refer to properties but usually they do not, as in the case of mental discourse. But predicates might have truth-makers independently of this issue.

supervenience is a weak principle and probably insufficient for physicalism Kim supplemented it with the principle of causal inheritance[32] to guarantee a metaphysically serious connection between realized and realizer properties. As there are interesting physical differences between different realizer properties, on grounds of this principle we are forced to think that the realized property instances are causally heterogeneous. If respectable scientific kinds are individuated on grounds of their causal powers, as both Fodor and Kim agreed, realized properties with physically diverse realizers are unfit candidates. The concept of pain remains meaningful as a synonymy class of relatively diverse predicates, but not as something that pics out a proper scientific kind.

This answer is satisfactory in the sense that it accounts for the underlying convictions that make Distinctness plausible for us and allows for causally efficacious mental properties at the same time. Local, domain specific reductions eliminate the problem created by distinctness on pain of the reconceptualization of multiply realized kinds. Higher-level kinds are mere concepts, as these kinds are causally heterogenous. They are not proper scientific kinds as such.

2.2.5 Dissolving the paradox by other means

Naturally, there are other ways of dissolving the paradox, but these strategies are or would be rejected by Kim. Others have proposed to give up No overdetermination (Sider 2003, Schaffer 2003) and to accept that certain outcomes can be heavily and systematically overdetermined. Allowing the systematic overdetermination of physical outcomes provides a solution because without the prohibition of overdetermination we are not forced to choose

---

[32] This is how Kim formulates the causal inheritance principle: "If mental property M is realized in a system at t in virtue of physical realization base P the causal powers of this instance of M are identical with the causal powers of P" (Kim 1992:18)

between mental and physical causes. In section 2.4 I will devote more attention to this thought. I think that this approach to dissolving the paradox is highly problematic and one can provide strong arguments against it.

Others proposed to give up the exclusion principle based on a different and more convincing argument (Árnadóttir & Crane 2013) finding motivation for thinking that the issue of overdetermination is simply irrelevant for the non-reductive physicalist as the distinctness of the mental and the physical can be maintained without maintaining that the mental and the physical causes involved in the exclusion argument are separate causes in a sense that would require to talk about them as proper overdeterminers. They are distinct in a similar way as parts are distinct from wholes or as the statue is distinct from the lamp of clay that it is made of, but for the same reason they are not separate or independent in the sense in which the different shooters in a firing squad are separate and independent causes. I find this to be a promising approach and it does have the potential to disarm Kim's version of the exclusion problem as on this account mental causes don't need to have causal powers that doesn't belong to the realizer system or some of its parts. However, I won't consider the latter option in this thesis as I decided to focus on other options.

The rest of this chapter will be interested mainly in the validity of certain premises in Kim's exclusion argument, with a special focus on causal closure (section 2.3), and I will put effort into making more sense of the No overdetermination idea he proposed (section 2.4). After that I will focus on Menzies' approach to exclusion which is similar to Kim's in that they both believe in an incompatibilist approach to the exclusion problem. The difference is that to run Menzies' version he doesn't need the No overdetermination idea. Distinctness and a reformulated version of the Exclusion principle are doing the job for him. As my main interest

is in the evaluation of Menzies' approach, the discussion of other solutions to the exclusion problem has to be done on another occasion.

The option Menzies took is to launch an attack on a hidden premise that is presupposed by many of the premises in Kim's exclusion argument. The presupposition that went unnoticed for some time concerns the concept of causation the argument relies on. A bunch of authors formulated similar objections finding this niche in the course of the last decade (Menzies 2008, 2013; Menzies and List 2009, 2010; Raatikainen 2010; Woodward 2008, 2015)[33]. The common point of these arguments is this: the notion of a 'sufficient cause' Kim utilized in his argument is defective, proper analyses shows that it is so weak that it fails on some basic tests for causal relations. If it does, then it is advisable to replace it with a notion that shows better performance on the tests mentioned and see what comes out of that experiment. Section 2.5 takes up the job of walking the reader through different interpretations of Kim's concept of causation and the possible arguments against those. In that section motivation is provided for replacing Kim's notion with a broadly counterfactual notion of causation. Chapter 4 turns to the thorough investigation of the approach favoured by Menzies according to which the exclusion principle should be based on a slightly modified version of the classical counterfactual analyses according to which causes are proportionate to their effects. As we will see the exclusion argument propped up with proportionate causation delivers the verdict that mental and other higher-level causes can exclude their realizers from causal efficacy even if it can happen the other way around as well.

---

[33] Woodwards approch is more substantially different from the others, because he choose a compatibilist approach according to which, both realizer and realized properties are causes of the same outcome even though these causes are not equally informative or effective in changing the outcome. So, he rejects the exclusion worry as such.

2.3   The causal closure of the physical

There are two important issues to clarify when it comes to the principle of physical closure. First, one should find a formulation of closure that is satisfactory in the sense that it allows physical closure without excluding the existence of non-physical causes by stipulation. As Lowe (2000) has shown, it is a tough challenge to find a solution that is not too weak and not too strong to do this job, but without that the exclusion principle wouldn't be useful as a premise in the exclusion argument. The main rule is that one shouldn't exclude mental causes in a question-begging manner.

The second issue to tackle concerns the empirical backing one can gather for the closure principle. From around the middle of the 20th century physicalism became a standard view in philosophy. This change was driven by modern physical science, its advancements and discoveries, so the plausibility of this view can only be appraised by a more careful examination of the evidence provided by science. Whatever formulation one settles with, causal closure is an empirical premise the faith in which must be proportionate to the empirical support one can gather for or against it.

2.3.1   Formulating the causal closure principle

The causal closure principle is at the heart of physicalism. The basic idea is as follows:

"One way of stating the principle of physical causal closure is this: If you pick any physical event and *trace out its causal ancestry or posterity, that will never take you outside the physical domain*." (Kim 1998:40, my italics)

As Kim summarizes it in his earlier writings:

Causal closure: Every physical effect has a sufficient physical cause.

To deny it would mean that physical properties, either fundamental or broadly physical properties, are insufficient to provide a complete explanation even of physical effects. According to most physicalists the success of the physical sciences provides good inductive evidence to think otherwise (see: Kim 2005:70). Kim formulated this succinctly:

"If you reject this principle, you are ipso facto rejecting the in-principle completability of physics - that is, the possibility of a complete and comprehensive physical theory of all physical phenomena. For you would be saying that any complete explanatory theory of the physical domain must invoke nonphysical causal agents. Never mind a complete physical explanation of everything there is; there couldn't even be a complete physical explanation of everything physical. *It is safe to assume that no serious physicalist could accept such a prospect*." (Kim 1998:40, my italics)

Unfortunately for physicalists, this simple formulation is probably too weak to capture the whole idea. It does not rule out the possibility of a causal chain between two physical events, C and E, that includes a non-physical link D. But this would mean that the physical domain is not closed, as to account for the physical event E we would need to evoke something non-physical, e.g. a mental cause. In case causation is transitive, the above formulation of closure allows for non-physical causes without running into a contradiction. Even though E has a non-physical cause D, by transitivity it also has a physical cause C (see: Lowe 2000).

Let me draw attention to something that usually goes unnoticed with respect to this counterargument against the simple formulation of closure. In many popular theories of causation, the relation is assumed to be transitive, but at present this is considered to be a

vexed issue. There are serious, stubborn counterexamples to the transitivity of causation, therefore intransitivity seems to be a plausible option (see: Hitchcock 2001, Maslen 2004, Hall 2004a). In chapter 3, I argue for the plausibility of a version of the latter view[34]. However, if causation is not transitive, there is no guarantee that when C causes D and D causes E then C is also a cause of E.

Applying this to Lowe's criticism of simple closure, there is no guarantee that when a physical event has a non-physical cause it has a physical cause as well just because the non-physical cause itself was caused by a physical cause. If causation is intransitive any causal chain might fail transitivity and we are faced with the duty to check what follows from this to the acceptability of the simple formulation of the closure principle.

To make the project meaningful, we should accept that even though we might not be able to decide which particular causal chain is transitive and which is not, due to epistemic limitations, the transitivity of a particular causal chain is a metaphysical fact. On the one end of the spectrum, it is possible that in all those cases when a physical effect is caused by a mental cause and in turn the mental cause is caused by a physical cause, transitivity fails. In possible worlds where such scenarios are realized, there are at least some physical effects without a sufficient physical cause and therefore the simple closure principle excludes them

---

[34] The transitivity of causation is complicated issue. In recent literature there is a debate concerning the assumption because from the 1990s onwards philosophers up came with more and more persuasive counterexamples to transitivity. These counterexamples drove philosophers to develop more nuanced counterfactual theories like versions of the contrastive theory in the 2000s. Others working in the interventionist framework reacted to those problems in a different fashion. While contrastivists like Schaffer (2005, 2013) and Maslen (2004) argued that the counterexamples can be explained away, interventionists like Hitchcock (2001) argued for intransitivity. My own investigation in the context of contrastive theories arrived at a result close to Hitchcock's. As I see the problem causation is intransitive, but there is a way to identify those special cases where causation fails to be transitive and one can provide sufficient conditions to secure transitivity for other kinds of cases.

from the set of physically closed worlds. On the other end of the spectrum, it is possible that in all cases when a physical effect is caused by a mental cause and in turn that mental cause is caused by a physical cause transitivity holds. In possible worlds where this kind of scenario is realized every physical effect has a sufficient physical cause via an intermediate mental cause and therefore the simple closure principle includes them into the set of physically closed worlds even though intuitively, they should be held outside.

So, even if causation is intransitive, there is still good reason to reject the simple closure principle. If we cannot decide what kind of word is our word in this respect, we should assume the worst and according to that option this is a world where every physical effect has a sufficient physical cause, but in many cases the connection is mediated via a non-physical cause.

This is good enough against the simple formulation of closure in itself, even if causation is intransitive. But it would be even better if we could identify and differentiate deviant causal chains, chains where transitivity fails, from normal causal chains where it holds. Fortunately, it is possible to do that. As my investigation in chapter 3 shows, one can provide sufficient conditions for the transitivity of causation to hold in causal chains. The condition is as follows: if the occurrence of an earlier event in a causal chain is not a precondition (or to use more familiar language, a background condition) for the existence of a later causal connection in the same chain then transitivity holds[35]. It fails in other cases, where earlier events in a causal chain are preconditions for the existence of a later causal link.

---

[35] In the causation literature such cases are called short-circuit scenarios (see: Hall 2004a, 2004b). The condition I developed abstracts away from the content of concrete examples and formulates the problem at the level of theoretical language shared by most theories of causation.

Let us translate this to our present case. If mental cause (D) of the final physical effect (E) in a chain only causes the final effect if it was (C) that caused (D) providing the necessary background conditions for (D) to be able to cause (E) at the same time, then we are faced with an intransitive chain where (C) is not a cause of (E). In ordinary examples of mental causation this doesn't apply. Suppose that my mental pain is caused by a dog bite and that pain causes me to swallow some painkillers. The difference-making ability of my pain to make me swallow painkillers is independent of the occurrence of the dog bite.

If ordinary causal scenarios involving mental causation, conform to my condition for transitivity, and they seemingly do, then most causal chains are transitive and the argument against the simple formulation of closure goes through. However, there is a way out for the defender of closure. The principle can be restricted to immediate causes of physical effects. It is helpful to use a time-based restriction on the definition as immediate causation is hard to define by other means:

Causal closure: Every physical effect at time t has a sufficient physical cause at time t.

There are other formulations of the principle similar in spirit, but this slightly amended formulation is considered to be more or less canonical. Unless indicated otherwise in the next section on the empirical backing for closure I will assume this version of the principle. But as we will see, certain further modifications will be required in order to incorporate certain empirical facts concerning the physical world.

2.3.2 The empirical backing for physical closure

Only a small handful of people tried to think through the empirical, inductive support for the causal closure principle, based on our best scientific knowledge. The most important text taking up the job of evaluating arguments for the causal closure principle is Papineau (2001). According to him, the real turning point in the story of physicalism happened in the middle of the 19th century with Helmholtz's establishing the conservation of energy in 1847.

With this new synthesis all previously established conservation principles concerning physical forces were brought together to form a universal principle of the conservation of energy. But synthesis in this case meant more than just bringing together all known conservation equivalences. Helmholtz considered cases where apparently non-conservative forces like friction were given microscopic interpretations in terms of conservative fundamental physical forces. The third element that made his work exceptional was his interest in physiology. Helmholtz was a physician by training and had a close connection with eminent German physiologists who were committed to a reductionist program the goal of which was to show that the same laws operate in the organic realm as in the inorganic realm. These three pillars formed the bases for many reductionist/physicalist programs that followed later. Papineau's reconstruction of the reasons why most philosophers have yielded to physicalism in the last fifty years also follows this tripartite approach.

2.3.2.1 The conservation of energy premise

First and foremost, physical closure is supported by (i) the conservation laws themselves. A suitable general formulation of the idea of a general conservation law can be found in Gibb (2010):

„Every physical system is conservative or is part of a larger system that is conservative (where a system is conservative if its total amount of energy and linear momentum can be redistributed, but not altered in amount, by changes that happen within it.)" (Gibb 2010:367)

This formulation says nothing about the forces involved and below, following Papineau and Gibb I will consider the conservation of energy and momentum only in abstract terms as for my purposes it doesn't matter which forces are involved in a general conservation principle. This abstract principle tells us that there are no other ways of changing things in a physical system then (1) by changing the amount of energy or momentum which can only be done from a more extended conservative system or (2) by redistributing energy or momentum inside the system. Normally anything that happens inside a closed physical system can be described as some kind of redistribution. (1) and (2) provide us with a good general characterization of the physical realm. It has a fixed amount of energy and momentum rendered in a pattern we might also call a distribution. According to this picture, whenever a physical change or a physical event takes place in that realm it is a change in the pattern, the distribution of energy or momentum. Naturally, for the purposes of physicalist philosophers, a suitable general law of conservation excludes the possibility of interventions by forces from outside the physical realm itself. More on this later, first we should examine some further points.

2.3.2.2 The successful reductions to fundamental forces premise

The second source of support (ii) for physical closure is the success of reductive attempts at explaining apparently non-conservative or dissipative forces like friction in terms of basic conservative forces. Here, it would be natural to add further examples to see the wider consequences of the issue. Even though Papineau restricts his discussion to friction, there are a handful of other non-conservative forces there are reduced to more basic conservative forces: viscosity, tension, compression or drag in fluid mechanics are the most important.

The reduction of certain other physical macro-properties (not necessarily forces) like temperature or properties of solids like hardness[36] are similarly interesting as today these are considered to be reduced to complex fundamental level electrical and mechanical forces. In the case of temperature its conservativeness was proven independently of its reduction. Micro-level kinetic energy formed the reduction base for temperature and this fact is even more interesting if we add that most non-conservative, dissipating forces dissipate energy via releasing heat.

The case of hardness requires some clarification. Hardness is not a force, it is the resistance of a solid to deformation, in other words a theory of hardness describes the behaviour of solids under forces like compression or sheering. So, it is not a force, but it is directly connected to activities of certain non-conservative forces. What the reduction of hardness properties, like deformation hardness, aim to explain is how compression or sheering forces are resisted by molecular or ionic bonds. This is where a problem of macro level mechanics is translated to the language of fundamental forces. The different aspects of

---

[36] Since ancient times hardness is one of the most important macro properties for human purposes. From engineering to architecture, it is indispensable to understand hardness. It is a complicated property of solids that is usually divided into three categories: scratch hardness, elastic hardness and deformation hardness.

hardness, resistance capacities of solids, are reduced to complex, bonding and crystal structure type dependent[37] conservative micro and meso-level electrical and mechanical forces.

Together such developments suggest that because more and more macro or higher-level forces and properties turned out to be amenable to reduction in terms of conservative fundamental forces other properties (if not forces) will probably also yield to scientific inquiry and will "reduce to a small stock of basic physical forces which converse energy" (Papineau 2001:27). I added hardness and temperature to the list to show that there are way more interesting cases relevant to the inductive base of this argument than friction. So, to sum it up: many apparently special forces and properties in nature were reduced to the workings of a small stock of basic conservative physical forces, therefore we have good inductive reasons to think that what we learned while investigating certain meso and macro-level properties can be extended to mental and other higher-level properties.

2.3.2.3 The support from physiological research

The third source of support (iii) for physical closure comes from physiology, from inductions based on advancements in the life sciences. It is a negative argument stating that there is no evidence for the workings of alternative forces inside living systems. Or in other words, as we get to know more and more about the inner nuts and bolts of organisms, we see the same basic physical laws at work as in the inorganic realm. This is contrary to what many scientists thought before the middle of the 20th century. From the middle of the 19th century onwards more and more evidence was found supporting the view that living systems fall under the

---

[37] Advancements in the reduction of hardness properties and the disunited, multiply realized nature of hardness properties are relatively new results in science. Originally physicists expected a unified explanation. (see: Gilman 2009, introduction).

scope of the second laws of thermodynamics[38], contrary to all appearances. The issue became more straightforward in the second half of the twentieth century with the advent of biochemistry, biophysics and later neuroscience. As Papineau says "detailed physiological investigation failed to uncover evidence of anything except familiar physical forces." (Papineau 2001:31) „...there is no direct evidence for vital or mental forces. Physiological research reveals no phenomena in living bodies that manifest such forces. All organic processes in living bodies seem to be fully accounted for by normal physical forces." (Papineau 2001:27) This amounts to a strong inductive argument against traditional vitalism that postulated and expected the existence of special vital forces over and above basic physical forces in special contexts. If there had been such forces scientists should have found deviant processes in living systems, processes the unfolding of which deviate from the course set by basic conservative physical laws.

---

[38] This is a historically very important case. It seems that life on Earth has a tendency to maintain a highly ordered state and what is even more to increase order and complexity. One simply has to consider the biosphere as a whole to see this second wonder. Both tendencies seem to go against the second law of thermodynamics. This seemed to require explanation and for some time it was taken to be a special or "vital" property only living systems exhibit. As these phenomena seemed to go against a basic physical principle people thought that their explanation requires the postulation of special vital forces. One important milestone in this story is Schrödinger's 1944 book "What is Life?". In it, he explains away a few apparent conflicts between physics and biology, among them the conflict with thermodynamics. This account later became the received view in biology. Schrödinger observes that living systems are open systems capable of increasing complexity locally but not globally. Relying on constant energy inflow provided by our Sun living systems can increase complexity even though this achievement is more than paid for by the increase of disorder outside, so the phenomenon is sustained by a net increase of disorder in the physical universe as a whole.

2.3.3   The empirical/inductive argument for closure

Some philosophers argued on good grounds that Papineau's argument requires further premises to make it work. Gibb (2010) shown that the argument from conservation laws as it was presented by Papineau is inconclusive. Her reconstruction brings out two important necessary presuppositions. Below, I follow Gibb (2010:374) although I introduce modifications with respect to the formulation of the conclusion, I use different labels for the premises and include extra clarifications between square brackets. The three Helmholtzian arguments for causal closure can be brought together and summarized in the following manner:

CONSERVATION: Every physical system is conservative or it is part of a larger conservative system.

NO-ENERGY: There is no non-physical energy [supported by (iii) negative evidence from physiology and (ii) positive from reductions to fundamental forces]

CLOSURE: Therefore, the causal closure of the physical is true

The conservation of energy and the no non-physical energy premises are insufficient for the conclusion in themselves. As we will see two extra assumptions and some additional clarifications are due. The conclusion speaks about causes whereas the premises are concerned with momentum and energy flow. Before anything can be said about the validity of the argument the two vocabularies should be linked up properly.

2.3.3.1 Connecting causal closure and its backing via physical causation

There are at least two ways of achieving a match. One is to reformulate causal closure in terms of physical causation. The new principle could rely on a suitable version of the so-called transference theories of causation (Dowe 2008, Salmon 1984) stating that the causal relation can be reduced to the transference of some conserved physical quantity. This option has serious drawbacks.

Let us start with a problem pointed out by Gibb (2010:377). Reliance on a transference theory of causation in the premises would make the first premise redundant and render the whole project based on conservation laws empty. The best way to show this is to include the transference-based definition of causation as a premise and see what happens:

CONSERVATION: Every physical system is conservative, or it is part of a larger conservative system.

NO-ENERGY: There is no non-physical energy

CAUSATION: Causation is transference of a physical quantity from cause to effect

----

CLOSURE: Therefore, the causal closure of the physical is true

The premise concerning causation entails causal closure without the conservation and/or no-energy premises. However, this comes at a cost; the resulting new argument for causal closure is analytically valid but less informative than the former version. It basically says: everything that is caused is caused by the transference of physical energy or momentum therefore it is impossible for things to happen because of other factors. This might be true independently of what we know about conservation laws. A probably unwanted consequence

of all this is that making the conservation premise redundant in the argument makes the original question concerning the empirical support for closure obsolete. To have a meaningful discussion concerning the support conservation laws provide for closure and physicalism, which is the only properly worked out empirically based defence, one has to find a different way of formulating the whole argument. Instead of getting rid of the original premises, one should make the relation between the body of supporting evidence relied on by most physicalists, conservation and no-energy, and the conclusion, causal closure, as explicit as it can be.

I think, it would be false to say that there is no motivation for relying solely on physical theories of causation as a starting point. Proponents of transference theories like Dowe or Salmon had reason to go with this idea. Being naturalist philosophers, they turned to science for answers concerning causation and their reasons for believing that physics might provide a good answer are similar to the reasons we can provide for believing in physicalism as a general metaphysical doctrine. The no-energy and conservation premises do provide direct motivation for the causation premise. If all physical energies/quantities are conservative, there are successful reductions of certain macro-level properties and seemingly non-conservative physical forces reduce to fundamental level conservative forces and there is no evidence for the existence of other kinds of energy then it is plausible to think that in this world the causal relation reduces to the transference of conserved physical quantities inside the physical realm. So, one way to go is to regard the causation premise as an intermediary conclusion between the original premises and the original conclusion:

CONSERVATION: Every physical system is conservative, or it is part of a larger conservative system.

NO-ENERGY: There is no non-physical energy

CAUSATION: Therefore, causation is transference of a physical quantity from cause to effect

CAUSATION: Causation is transference of a physical quantity from cause to effect

----

CLOSURE: Therefore, the causal closure of the physical is true

Note, that the argument for physical causation is as vulnerable as the argument for closure was. For example, one might say that there is still room for finding special energies in some context or another we still don't know enough about. But, this is uninteresting, in Papineau's treatment the whole idea of closure is predicated on the plausibility of the no energy premise, so to reject physical causation we need points that only apply more specifically to the argument for physical causation.

The first line of argument against the physical reduction of the causal relations might be the following. It is fair to say that starting with physical causation would commit one to a radically reductive view of causation, a view that begs the question against any metaphysical views opposing physicalism, like interactive dualism or ontological emergence. If causation in general is identified with physical causation, then no other "mechanism" for causal "doing" is possible. Even if one believes in some form of physicalism about our world it does not follow that there are no possible non-physical worlds or that there are no non-physical entities in our world in causal interaction with each other even though they are not in causal interaction with the physical realm. Even the causal closure of the physical realm applied to our world is compatible with the existence of non-physical things isolated from physical goings on and

causation between them that is not based on physical mechanisms. Therefore, even if there are good empirical arguments for reducing causation to physical causation at least in our world it would be too hasty to do so. If possible, it is advisable to find other ways to argue for the causal closure of the physical.

Furthermore, to rely on less metaphysical, more sober-minded arguments it is enough to attend to some empirical facts concerning higher-level causal relations in our world. Phil Dowe, a major contemporary proponent of the transference view of causation, says this:

"...to suppose that the conserved quantity theory will deal with causation in other branches of science [outside physics] also requires commitment to a fairly thorough going reduction, since clearly there is nothing in economics or psychology that could pass for a conservation law." (Dowe 2008, section 6.6)

The idea is the following: if, at present, most higher-level properties/quantities don't seem to reduce to basic level physical goings on in any straightforward manner and, as Dowe pointed out, there are no insights concerning many important higher-level quantities suggesting their conservativeness then it is an open question whether causal relations determined for these quantities can be reduced to the transference of basic physical energy. This thought provides empirical motivation for the view that it would be too hasty to accept physical causation on the evidence at hand and therefore a more permissive theory of causation should be relied on.

There are also plausible arguments against accepting physical theories of causation as good theories of causation. Transference theories of causation suffer from well-known internal problems. They have certain advantages compared to the metaphysically more neutral counterfactual theories, as they can avoid problems concerning negative causes or

effects or those of late pre-emption (more on this in section 2.5.3), but at the same time they suffer from a serious and notable disadvantage: transference theories have a hard time handling problems of causal relevance. A theory of causation should be able to differentiate aspects of the cause that are relevant for bringing about some effect from aspects that are not. But from the point of view of most physical theories any physical transmission between the cause and the effect counts as a causal relation which results in so called causal misconnections, failures of causal relevance (see: section 2.5.4). This is the main reason why most philosophers agree that physical theories of causation only provide necessary conditions for causation, but the theory is not good enough to provide sufficient conditions. Metaphysically more neutral theories like counterfactual theories are better in handling this issue. And they have a further advantage, they can incorporate non-physical causes into causal discourse. Taking the above worries seriously provides sufficient motivation for the conclusion that even if there are no positive arguments for the existence of special higher-level or non-physical causes, we should leave the door open for that option.

Let us turn to the second step of the argument, the inference from physical causation to closure. If causation is physical causation, then it is obvious that there can be no non-physical causation. So, mental causes are excluded from the get-go. The good news is that the premise seems to entail physical causal closure. But, is this really so?

There is at least one argument against the validity of this inference. According to our best scientific knowledge it is possible that there are uncaused events in the physical realm (the received example is radioactive decay in the case of a particular atom where it is impossible to predict the exact time of decay[39]). This is something that the causal closure

---

[39] For interesting interpretations of uncaused physical events and their non-causal powers see Lowe (2013), especially from page 158 to 161.

argument wouldn't necessarily allow for in its canonical formulation. If every physical effect at time t_1 has a sufficient physical cause at time t_1 then there can be no uncaused physical occurrences as every physical effect has a cause. However, this problem can be easily avoided by clarifications of the terms used. In case the term "physical effect" refers to a physical occurrence that is caused by something, either a physical or a non-physical cause, then uncaused physical occurrences fit neatly into the picture.

Unfortunately, there is also bad news concerning the argument from physical causation. Not only the conservation and no energy premises are made redundant by this premise in the previous version of the argument, but in case one plugs this into Kim's physical causal exclusion argument, replacing more permissive versions of the causal closure premise, it makes redundant the no overdetermination premise. In case the no overdetermination premise is dispensed with, the exclusion argument becomes quite empty or in other words it becomes more of a statement than an argument. (See: Lowe 2000:571-573).

Kim formulated causal closure in a way that allows for non-physical causation. His version of the causal closure principle only states that every physical effect has a sufficient physical cause. This does not exclude mental events as causes of physical events. However, formulating causal closure in terms of a transference theory excludes mental causes from the get-go and by doing that begs the question against possible opponents. From physical causation it jumps to the conclusion that mental causes cannot be distinct from physical causes as there are no mental causes. I agree with Kim (2005:51) that there is a philosophical gain to be had from separating the causal closure principle from the no overdetermination and exclusion principles. These have different sources of support. Closure only has empirical support or opposition whereas the no overdetermination premise of the exclusion principle is considered to be analytically true by Kim, while others like Menzies and List (2009) and

Raatikainen (2010) think that it is far from it (more on this in sections 2.4 and 4.1). By keeping the two premises separate we render the exclusion argument more interesting in terms of its content and avoid begging the question against interactive dualists and other opponents of physicalism. This is a serious reason why physical causation is a bad candidate for a starting point if one wants to argue for physical causal closure.

To sum up, physical causal closure is entailed by the physical causation premise, but the premise is too strong for two main reasons: (1) it makes the conservation premise redundant and (2) that version of the argument pre-empts the causal exclusion argument. Therefore, we should try a different approach.

There is one interesting insight from the discussion of closure based on physical causation we should incorporate into an amended formulation of physicalism. The plausible existence of uncaused physical events requires a modification that allows for them. Reacting to arguments from Lowe (2000), Kim (2005:43) defines closure in the following way:

> **Causal closure:** If a physical effect has a cause that occurs at t, it has a physical cause that occurs at t.

2.3.3.2 Connecting causal closure and its backing in a non-reductive manner

The other way of matching the language of the conservation and no energy premises with the conclusion is to include a few further premises concerning the possible ways physical systems can be affected by anything. This is how Gibb (2010) reformulated and supplemented the argument. The inclusion of further premises is necessary if one wants to make the inference from the premises to the conclusion valid.

First, it is plausible that a physical system can be changed by physical energy or momentum transference from without. Note two things: according to the conservation

premise this transference can only start from without the affected system, from a more extended system and if the whole physical realm is one conservative system then from without this system no physical energy flow is possible. Even if there is something outside the physical realm that realm cannot contribute physical energy to the physical realm as that would violate the conservativeness of physical energies. In other words, nothing in the physical realm/system happens because something outside of it changes the amount of energy or momentum inside the realm.

However, there is a second plausible way a physical system can be changed. The energy and/or momentum distribution of the system can be changed from without. Note that redistribution is a good general description for the normal evolution of any closed physical system that behaves in line with the conservation of energy principle and basic dynamical laws. In all such systems energy and momentum gets redistributed as a result of the system's evolution in time. But normal evolution should not count as an interesting change introduced into the distribution of energy or momentum as usually when talking about vital or mental causation we are talking about interventions originating from without the physical system causing redistribution. So, the third premise to include says that there are (at least) two ways to affect a physical system:

AFFECTABILITY: The only ways that something could affect a physical system is by

(1) affecting the amount of energy or momentum within it or by (2) redistributing

the amount of energy or momentum within it.

We still owe an answer to the question of what kind of things could bring about changes in the distribution of energy or momentum. This question should be treated with much caution. It is plausible that it is brought about, caused by the transference of energy or

momentum. According to the conservation principle, in that case, the impulse should come from without the physical system in question. If our system is the physical realm as a whole, then the existence of such impulses is inconsistent with the conservation premise. If the system in question is a subsystem of the physical realm then this could happen in accordance with the conservation premise. So, it seems that what the fourth premise should say is this:

REDISTRIBUTABILITY: Redistribution of energy and momentum cannot be brought about without supplying energy or momentum.

Accepting these four premises provides us with an argument for the causal closure of the physical that appears to be valid. However, following Gibb (2010), I don't want to commit myself to the sufficiency of this argument, the only real commitment I would like to make is that these four premises are required for the argument from conservation to work.

CONSERVATION: Every physical system is conservative, or it is part of a larger conservative system.

NO-ENERGY: There is no non-physical energy

AFFECTABILITY: The only way that something non-physical could affect the physical is by (1) affecting the amount of energy or momentum within it or (2) redistributing the amount of energy or momentum within it.

REDISTRIBUTABILITY: Redistribution of energy and momentum cannot be brought about without supplying energy or momentum.

CLOSURE: Therefore, the causal closure of the physical is true

Now that we have a compact argument that seems to work let's turn to the examination of the new premises. The present formulation of the redistribution premise would be supported by the acceptance of a physical theory of causation, but as we already saw that would make the conservation premise redundant, so we need independent support. It is important to mark that the premise concerning redistribution says something dubious. I have already mentioned that redistribution can result from the normal evolution of a closed physical system without the introduction of extra energy, so the premise seems to say more than it should, it should allow for further possibilities.

In most possible closed physical systems, a kind of uninteresting or default redistribution occurs without impulses from the outside. This process is driven by basic physical laws. While the closed system runs its course, it evolves from an initial state towards others states or an end state. This evolution of the system can be expressed as a pattern of redistributions of energy and momentum in time where earlier distributions determine later distributions or states. There is exchange of relevant properties/quantities between components of the system, but there is no change in total amounts. This default redistribution does not require energy or momentum from the outside it is simply a reflection of the internal evolution of the physical system in question.

2.3.3.3 Evidence for physicalism and the room for emergence

Transference of energy or momentum from the outside would change the distribution of energy or momentum inside the system by changing the amount of energy or momentum at the same time, but this is not required for redistribution to take place. Redistribution only requires the transference of energy and momentum inside the system from one region to another. This is an important point for the discussion of physical closure. If redistribution

could happen in a way that deviates from normal evolution but takes place without energy transfer into the system from the outside, we would have an interesting third case at hand that is in no contradiction with conservation laws.

There is good reason to liberalise the redistribution premise in a way that includes that latter possibility (see Gibb 2010). We had reason not to start with a physical theory of causation, therefore in our analyses and reformulation of the premises, where it is possible, we should rely on a more permissive, neutral theory of causation like the counterfactual theory or the powers theory. In case we accept that causes are difference-makers whatever relation realizes their capacity to make a difference the redistribution premise becomes compatible with redistribution without the transference of physical energy from the outside and maybe more than that:

REDISTRIBUTABILITY*: Redistribution of energy and momentum can be brought about without supplying energy or momentum.

Conservation laws put no restrictions on what could initiate redistributions of energy or momentum. If the amount of energy and momentum remains constant in a system conservation is not violated. The former is true even if the redistribution that takes place is not in line with the normal evolution of the physical system driven solely by basic physical laws. This is exactly the point where classical and modern emergentists alike found a gap in the physicalist/mechanist worldview exploitable for their purposes. The last great British emergentist, C. D. Broad in his discussion of mind-body interaction and the conservation of energy principle suggested the following:

"[facts concerning the conservation of energy] suggest that all the energy of our bodily actions comes out of and goes back into the physical world, and that minds

neither add energy to nor abstract it from the latter. What they do, if they do anything, is to determine that at a given moment so much energy shall change from the chemical form to the form of bodily movement; and they determine this, so far as we can see, without altering the total amount of energy in the physical world."

(Broad 1925:109)

So, conservativeness might require the exclusion of certain kinds of interventions into a physical system, but the road is still open for the emergentist to say that physical energy can be redistributed by non-physical difference-making causes in line with physical conservation laws. So far, so good, but this has further important consequences we should make explicit. The emergentist, to be able to accommodate interventions of non-basic level physical origins, should deny the exclusiveness of basic physical laws[40]. This does not require the violation of any basic physical law. According to most emergentists, physical laws apply to all things but under certain circumstances, when certain base level conditions are satisfied, some further laws/forces also apply to them in concert with base level physical laws. This is called downwards or top-down causation by most proponents of such views. This is how Hendry (2010) explains the idea succinctly:

"…to say that some system exhibits downward causation is to make a counternomic claim about it — that its behaviour would be different were it determined only by the more basic laws governing the stuff of which it is made."

Hendry (2010a:185)

---

[40] For a detailed discussion of the issue concerning the scope of physical and emergent laws in the context of modern chemistry see: Hendry 2010a, especially page 188

The view is consistent with the ubiquity of basic physical laws, but it assumes special laws that are activated only in the context of certain higher-level phenomena. In those special contexts the behaviour of physical particles is determined not only by basic physical laws, but by special higher-level laws as well. Therefore, their behaviour cannot be predicted based solely on basic physical laws.

„Under the ubiquity of physics, physical principles constrain the motions of particular systems though they may not fully determine them." Hendry (2010b:217)

It is important to note that it doesn't matter whether emergent special laws are expressions of conservative or non-conservative vital, mental or other energies. Either way they don't change the amount of physical energy in the system, what they change is only its distribution. There is some reason to think that they would be conservative. As Papineau explains,

"...the nature of other fundamental forces provides inductive reason to suppose these *sui generis* forces will be conservative in their own right." (Papineau 2001:30)

And historically, it is true that most of those who believed that there are non-physical energies believed them to be conservative. But conservativeness is not a difference-maker when it comes to the question of empirical backing for emergent laws. The decisive question is, how do we get to know about the existence of emergent laws? A passage from C. D. Broad combined with insights already on our table helps to formulate an answer. Below, he talks about emergent chemical powers:

„If the emergent theory of chemical compounds be true, a mathematical archangel,

gifted with the further power of perceiving the microscopic structure of atoms as

easily as we can perceive hay-stacks, could no more predict the behaviour of silver

or of chlorine or the properties of silver-chloride without having observed samples

of those than we can at present." (Broad 1925:71-72)

According to Broad, to be able to differentiate emergent from non-emergent properties first we should have access to the most precise descriptions of basic level goings on, the exact microscopic physical structure of atoms and the basic-level ubiquitous laws that apply to them. In case the behaviour of chemical compounds cannot be derived from this information then we are faced with an instance of an emergent law. The source of our failure of derivation cannot be rooted in practical limitations, only if the derivation is impossible even for, as Broad (1925:70) formulates this, a mathematical archangel who has limitless computational powers can we conclude that we have empirical evidence for the existence of an emergent law. Using Hendry's counternomic formulation, we can also say: in such a case, the calculated behaviour of a system governed only by basic physical laws would be different from its actual behaviour. The deviation from the results gained by the purely basic level ideal calculation could be assigned to the governing force of an emergent law.

We should highlight an important difference compared to basic physical laws. Empirical evidence for the existence of emergent laws can only be gathered under those special circumstances where they apply as their scope is limited to the special phenomena to which they apply. If such a form of emergentism were true, the behaviour of special systems like chemical systems would be calculable from the joint work of basic physical and emergent laws according to some rule of composition (See: McLaughlin 1992:31). Emergentism, conceptualised this way, is consistent with conservation laws and does not involve violations of basic physical laws.

But is it compatible with the causal closure of the physical? It is clearly not, but the contradicting premise as Papineau (2001:30) and Gibb (2010:379) both point out, is not the conservation premise in the argument for closure, as people might expect. It is the no energy premise. If there is no non-physical energy or in other worlds there are no non-physical forces, then there is nothing that could do the work of redistributing physical energy or momentum in a manner that deviates from the normal evolution of the system solely determined by basic physical laws. According to emergentists following the footsteps of Broad, a special science law would be a description of a physical energy redistribution pattern that cannot be accounted for solely in terms of basic physical laws but, as we already saw, can be explained using basic laws combined with emergent laws in accordance with certain rules of composition allowing the computation of what happens when something is acted upon by both basic and emergent laws.

According to Papineau, the no energy premise is based on positive inductive evidence from putative cases of successful reduction and on negative inductive evidence from physiology/biology. Obviously, the evidence for it is not fully conclusive, but if one accepts no energy as a starting point it excludes the possibility of emergent laws and also the possibility of redistribution not caused by energy or momentum transfer form without the local physical system under investigation.

According to some emergentists, like Hendry (2010a, 2010b), the empirical evidence is not fully convincing, especially when it comes to the classical example of the reduction of chemistry to quantum physics. In his classic article McLaughlin (1992) argues that even though emergentism is a coherent belief system, the reason why it went out of fashion after the 1920s is the development of quantum chemistry. As I briefly mentioned, for Broad and for most former emergentists like John Stuart Mill, chemistry provided a convincing case where

special science phenomena seemed to go beyond what physics could explain. In McLaughlin's reconstruction, this new success in the reductive unification of the natural sciences brought about convincing new evidence for physicalism and at the same time against emergentism.

However, this presupposes that Broad's mathematical archangel has done her job and relying on the Schrödinger equation calculated chemical behaviour based solely on lower-level information about physical particles below the chemical level of composition. Hendry (2010a) shows in persuasive manner that this is definitely not so. Even though the reduction of simple atoms was done in an acceptable way, the only clear case being the case of the hydrogen atom, the symmetry properties of even the simplest of molecules evade straightforward reduction. Existing "reductive" models simply put those properties "in by hand" (Hendry 2010a:185). The so-called Born-Oppenheimer approximation that scientists use to derive features of molecules "makes only a small difference to the calculated energy of the molecule, but it makes a big difference to its symmetry properties" (Hendry 2010a:185), therefore a significant addition is required to achieve the expected results. As symmetry properties are crucial to recover important chemical capacities and are only available on the chemical level this practice cannot be called a real, satisfactory approximation or reduction. The physicalist might wait for further theoretical developments on meta-inductive grounds, but the present state of affairs provides less support for unificationist optimism then usually claimed. Therefore, Hendry (2010a, 2017) declares a stalemate in the debate between the reductionist and emergentist sides.

2.3.4   Conclusions about the status of physical closure

My summary of the moral is the following. There is growing empirical evidence for physicalism, more specifically for the causal closure of the physical. The existence of

conservation laws, the reduction of dissipating forces to conservative forces and discoveries that have happened in physiology, especially in the 20th century. Some important developments also took place with the advent of quantum chemistry, despite the fact that this project has achieved significantly less than what was announced. On grounds of the available evidence from scientific practice physicalists should be more careful with self-confident assertions concerning the prospects of physical closure.

I do think that acknowledging all this is highly significant in evaluating physicalism as a view about the nature of our reality. However, in the context of causal exclusion arguments it is enough to see that there is growing empirical support for closure, so the premise has both inductive and meta-inductive support even though it is far from being certain. There is still ample room to question the no energy premise in the argument for physical closure, which via weakening the plausibility of the causal closure principle weakens Kim's exclusion argument as well, but naturally even if the road is still open no convinced physicalist would choose to walk the path of arguing against closure just to solve the paradox of the exclusion argument. As we will see in section 2.5 and chapter 4, those non-reductive physicalists rejecting the conclusion of Kim's causal exclusion argument, like Menzies (2008, 2013) or Raatikainen (2010), to be able to retain physicalism as a common ground are trying to find holes in other starting points of the exclusion argument.

2.4   Exclusion and no overdetermination

The no overdetermination premise of Kim's exclusion argument tells us at least the following: effects are never systematically overdetermined. Kim (1998, 2005) rejects systematic overdetermination as absurd. To investigate the issue, one first needs to ask what is meant by overdetermination. The motivation for the concept of overdetermination comes from some well-known causal scenarios. Imagine that two assassins are planning to kill a dictator. Both of them have a good vantage point for shooting at the target and when the time comes (events C1 and C2 on Figure 2.4-1[41]) they hit the target at the exact same moment. Both bullets go through the middle of the heart. The dictator dies immediately (event D on the same figure). The actions of the assassins are more than enough for the outcome. The dictator would have died even if there was only one assassin sending a bullet through his heart. The term "overdetermination" is used to refer to this excess of causes.[42]
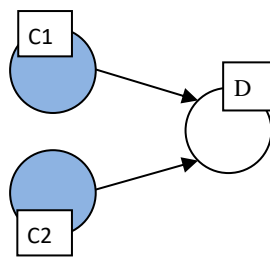
Figure 2.4-1

---

[41] On neuron diagrams circles signify events, full circles are events that took place, empty ones are events that did not. Normal arrows signify a manifested power producing another event, oval arrows signify prevention.

[42] Overdetermination presents a problem for many theories of causation (especially regularity theories and counterfactual theories), but the basic scenario itself was never denied as being non-causal or conceptually flawed. In the causation literature overdetermination is taken to be a fact of causal discourse, so all theories of causation are expected to provide some analysis for it. There are two basic approaches among theories of causation. Following Schaffer (2003) one could be called individualist, the other collectivist. The individualist thinks that both token overdeterminers are causes of the outcome, while the collectivist thinks that they cause the outcome only together. The basic counterfactual account conforms to the collectivist view, as the effect becomes absent only if both candidate causes are absent, while physical quantity transference views conform to the individualist view. David Lewis argued that we don't have clear intuitions about overdetermination in these respects, so it is natural that different theories of causation differ in their judgements.

If it is a fact of causal discourse that overdetermination exists, why do philosophers deny its existence in the context of the causal exclusion argument? In the mental causation literature people don't deny the mere possibility that certain physical effects can have both a physical and a mental cause simultaneously as overdetermining causes. What is denied is the systematic occurrence of the phenomenon (Kim 1998). In everyday life overdetermined effects are considered to be a rare phenomenon, and for good reason. It is not easy to set things up properly for a scenario like the one with the two assassins and many similar situations are a result of rare coincidences. In most causal scenarios there is only one cause that is sufficient for causing the relevant outcome. If overdetermination requires extraordinary circumstances it would be more than surprising if certain kinds of effects were overdetermined as a rule. Let us highlight the most important arguments against mental-physical overdetermination.

The first point against systematic inter-level causation is that it would be an arbitrary move to suppose that there is systematic overdetermination in the context of mental causation just to save autonomous mental causation in the context of physicalism. As far as we know overdetermination is a rare coincidence in nature. If there are no independent motivations for this solution, if we can't offer a mechanism that would produce systematic overdetermination it is better to avoid it.

Secondly, Ockham's razor might deliver a sufficient methodological motivation for the rejection. The use of Ockham's razor is nicely reflected in our everyday and also in professional practices. If police find the culprit for a certain homicide they don't continue to search for additional causes. The case gets closed. The police would require extra reasons for doing more. We implicitly assume that an effect usually only has one cause and that is sufficient for it.

Although only in rare cases, the police might be provided with good reasons to leave the murder case open. Imagine that the investigation into the death of the victim has shown that he/she was both poisoned and shot. In that case the police would check whether these two seemingly independent causes, lethal in themselves, had different origins or not. It might have been that someone intended to make sure that the victim really gets killed, so the person who ordered the killing hired two assassins, each relying on different methods. Or it could have been that two independent parties were interested in neutralizing the victim. In such a scenario the death of the victim is overdetermined, and this fact can be established by coroners. The effect – the state of the dead body – is quite different from cases where there is no overdetermination.

The exclusion principle summarizes the moral of what was said about overdetermination. There is no systematic overdetermination, therefore:

> **Exclusion principle**: In general, if an effect E is caused by C (where C is sufficient for
>
> E), then any other cause numerically distinct from C is excluded from the processes
>
> leading to E.

2.4.1 Non-coincidental overdetermination as a solution

Unfortunately, the motivation given for this version of the principle is not satisfactory to deny overdetermination in the case of mental causation. According to some philosophers (see e.g.: Bennett 2003), inter-level overdetermination cannot be modelled and rejected based on our knowledge of intra-level overdetermination, as there are important disanalogies. The supervenience relation that holds between the physical realizer and the realized mental property makes the occurrence of overdetermination systematic. If we accept the supervenience of the mental on the physical, then in all cases of mental causation we get two

causes for physical outcomes such as my typing of this text. A mental cause (my intention) and a physical cause (my brain state). These two always take place at the same time. This means that overdetermination might not only be a highly systematic phenomenon, but a universal and necessary one, at least in this special kind of case. This is the reason why I will call it non-coincidental[43]. The supervenience of the mental on the physical is taken to be at least nomologically necessary by non-reductive physicalists, but all supporters argue for stronger forms of necessity than that. This means that in worlds with the same laws as our world there is no way to separate the physical realizer and the realized mental property, but for many physicalists the separation is impossible even beyond this set of worlds. As a result, for supporters of non-reductive physicalism inter-level overdetermination is a strong necessity in the context of mental causation.

This allows for an argument against the principle of no overdetermination according to which such principles cannot be transferred from well-known cases of overdetermination to cases of overdetermination between levels. There is an important disanalogy between everyday or intra-level overdetermination scenarios and inter-level scenarios as in the case

---

[43] In ordinary causal discourse the term „non-coincidental" means at least the following: the occurence of two events is coordinated by a (probably not too distant) common cause. So, their co-occurence is not a coincidence, but it could be a coincidence, this is the reason why it is interesting to know which holds true in a particular case. In the case of inter-level overdetermination the connection between the mental and the physical is much stronger than that. Depending on how the supervenience relation is spelled out for mental properties the modal strength of the covariation might change, but any supervenience relation that is necessary for physicalism provide a stronger connection between the correlated causes then a common cause can. Redundant causes brought about by a common cause, as in the case of the two assassins and the dictator, are separable in our world. In some possible cases, the same kinds of causes occur in an unconnected manner. E.g. two assassins who don't know about each other and have different motivations for killing the dictator. But mental causes always go together with physical causes. So, their co-occurence is non-coincidental in the stronger sense that in our world it cannot be a coincidence.

of mental causation. Therefore, we can't simply model mental causation on intra-level causation.

Overdetermination might be rare in intra-level cases, but from this nothing straightforward follows for inter-level cases. Systematic overdetermination is weird in an everyday context as it involves surprising coincidences or very specific conditions that easily fail in practice. But inter-level overdetermination is a different breed. It involves no coincidences and the supervenience relation provides a neat explanation for the systematic occurrence of overdetermination (see Bennett (2003) or Sider (2003) for versions of this basic idea). As this approach provides us with a mechanism that produces systematic overdetermination it seems hard to deny the possibility.

A lot depends on the no overdetermination premise. If it proves to be a solid premise (together with causal closure and supervenience, premises never questioned by non-reductive physicalists), it is reasonable to turn the exclusion argument against the distinctness premise as suggested by Kim. If the above argument proves to be plausible it seems to be a promising way out for non-reductive physicalists. What can be put forward to answer the above argument? Why shouldn't we accept that overdetermination can be systematic in some cases even though it is not in everyday causal scenarios?

On the one hand, the inter-level dependence between the mental and physical explains systematic overdetermination. However, on the other hand, if all supervenient causes are redundant causes then we have endless amounts of excess causes in our universe and that sounds like a design failure[44]. Why would God or any other kind of designer need those redundant causes? Why would there be such an unparsimonious mechanism?

---

[44] If there are other layers of reality between the mental and the basic physical where autonomous causation should be taken seriously then this abundance of causes goes to an incredible extrem.

A possible, but dubious answer would be that these redundant overdeterminers are some kind of backup causes. This means that they are there to secure certain outcomes in our universe. However, redundant higher-level causes supervening on lower-level realizers are ill-suited for such purposes. To be proper backup causes their function should be something like to make sure that certain effects take place even when something goes wrong on some basic or default causal track.

In the assassin scenario the presence of more than one shooter aims to make sure of the death of the dictator. If there is a problem with a rifle or one of the shooters gets caught by security officers, there is still a sufficient backup option activated at the right time. Redundant higher-level causes can't serve as backup causes in this sense. A backup cause should be able to stay active when the default cause gets deactivated, otherwise it can't serve as a backup for the default cause. But, according to the supervenience based picture of the world, tokens of higher-level causes disappear whenever no token of their physical realizers is present. Therefore, higher-level causes can't provide backup for physical causes.

Can physical causes serve as backups for mental causes supervening on them? No, they can't and again it is the supervenience relation that makes this impossible. When a token of their physical realizers is present, a token of mental cause is present as well. If in our world the mental supervenes on the physical, at least with nomological necessity, then in this world the mental has no chance of disappearing and leaving its realizer bare. When it disappears, the realizer disappears as well. Therefore, in worlds like our mental causes cannot be backed up by physical causes.

Redundant causation as in the case of the two assassins is predicated on the premise that overdeterminer events one and two are separate, non-overlapping existences. This is what makes overdetermination scenarios either a rare coincidence or a carefully contrived,

but still fragile situation. But exactly because of their separateness they can work as backups when the situation is contrived so.

Redundant causes in the case of lower and higher-level causes are asymmetrically dependent. Higher-level causes depend on lower-level causes. For this reason, they can't work as backup causes for a single outcome. They are useless as redundant backup causes similarly to God's help according to the saying: "God helps those who help themselves". God (the mental) never causes physical outcomes directly only when some physical cause (the believer/human) causes it already. Why do we need God in that case?[45]

So, one has to pay a high price for accepting the idea that the non-coincidental relation between causes at different levels explains systematic overdetermination. It results in a highly unparsimonious picture of the world. In the absence of serious further positive reasons for it, we have a good reason to reject it. Simply insisting that non-coincidental overdetermination is a possibility makes the solution ad hoc as well. Nevertheless, there are further arguments for and against inter-level overdetermination.

2.4.2   Overdetermination all over the place. Why worry?

There are some authors such as Schaffer (2003) who argue that overdetermination is everywhere, even in everyday contexts. If he is right, then there is a stronger motivation for accepting non-coincidental overdetermination. If overdetermination is more common, then the systematic nature of inter-level overdetermination is less surprising. As Schaffer himself notes, supervenience based inter-level overdetermination does not require this extra

---

[45] Well, in case the saying intends to say that God provides extra help over and above what the self-helper is already doing, that sounds better. But that could be expected to make some measurable difference compared to cases where God provides no extra support as the saying gives no guarantee that God provides the extra help in all possible cases. More on issues like this in the next section.

motivation, supervenience provides a good enough background mechanism. In Schaffer's view cases involving lawfully correlated redundant causes are numerous. He quotes Mackie's example of the sledgehammer flattening a chestnut:

> "[T]he whole of the blow was not necessary for [the flattening of the chestnut], though it was more than sufficient: a somewhat lighter blow would have sufficed. Even if part of the hammer-head had been absent, this result would still have come about" (Mackie 1974:43).

Suppose that two non-overlapping parts of the sledgehammer would have been enough for the flattening effect. The two parts of the hammer are held together by basic physical forces, so they are lawfully correlated. It is easy to conjure up similar scenarios. Mackie calls this quantitative overdetermination. It occurs when a cause is decomposable into any number of distinct and independently sufficient parts. Because of lawful correlation between the parts there is nothing accidental or improbable concerning these cases. It does not involve accidents or difficult challenges for coordinated action.

It is unclear how common such scenarios are, but I will accept them to be common enough. However, there is still an important difference between these cases and non-coincidental inter-level overdetermination. The occurrence of the latter is not a contingent fact, it is as necessary because the supervenience relation between the mental and the physical dictates it to be. For the physicalist this requires something more than nomological necessity, whereas the "lawful" correlation in the case of quantitative overdetermination reflects a contingent fact of nature in our world. The parts of the sledgehammer are bound together by physical forces, but only when they are. This disanalogy can be transformed into an argument repeating what I said in the former section.

To motivate his point about frequent everyday overdetermination, Schaffer mentions that "engineers seek out quantitative overdetermination for safety" and so it must be a frequent phenomenon. Redundancies in the biological realm are also numerous for the exact similar, but evolutionary reasons. These redundancy-based strategies are effective and meaningful exactly because the correlation between the distinct causes is contingent in our world. But, as we saw in the last section, supervenience based correlation renders this function of redundant causes impossible. It doesn't matter whether overdetermination is more or less frequent in everyday life, supervenience based redundant causation still implies an unparsimonious setup.

2.4.3   Threshold effects, cumulated effects and overdetermination

To formulate a strong argument that works against inter-level, non-coincidental overdetermination, below I will consider how overdetermination scenarios are constructed conceptually. In doing this I am unpacking some underdeveloped comments Kim made (1998:45, 2005:46-48) according to which mental-physical overdetermination would violate the causal closure principle itself.

The first question we should pose is this: how is it possible that two independent, non-overlapping causes are not doing more than one of those causes alone? Only a few authors are cultivating this issue in the causation literature, let alone the mental causation debate, even though it is highly important to have an answer when it comes to supervenience based overdetermination. A useful and plausible suggestion to start with answering the question is that in overdetermination scenarios causal discourse is interested in threshold effects[46]. A

---

[46] This starting point of mine and the notion of a treshold effect originates from Hausman's work (1992: 162, especially footnote 5) in other respects the argument owes a lot to Bunzl (1979)

death can be achieved in many ways, can happen in many different ways, but what we humans are interested in is usually whether the threshold we call death was crossed or not.

There are cases where the existence of a threshold is even more straightforward. Imagine a small bowl receiving water from two independent sources. The threshold in this case is crossed when the bowl overflows, so bowl overflow is the threshold effect we are interested in. When there is enough water in the bowl overflow takes place. It is possible to achieve this threshold effect relying on only one source for water inflow, and it is also possible to do that relying on two sources where the two sources contain more water than the one source in the first scenario.

Imagine first that only I poured my full glass of water into the bowl and second that while I was doing the same someone else was pouring another glass that contained less than mine. Bowl overflow takes place in both cases, but there will be a difference that can be measured in the amount of water around the bowl on the table and maybe also on the floor. In the case where there are two sources of inflow more water flows over. This is what causal discourse dismisses as unimportant when the focus is on the crossing of the threshold. I will call it the cumulated effect.

Remember my earlier overdetermination scenario where police investigators and coroners gathered evidence pointing to the overdetermination of a killing. There is no difference between this scenario compared to a non-overdetermined version of the same scenario in terms of the threshold effect. However, there is in terms of the cumulated effect originating from two sources in one case, but from only one source in the other. There are many ways of reaching the threshold called death, or in other worlds the state when a living being becomes incapable of regaining its homeostatic equilibrium. In cases of overdetermined death much damage is caused to the body, which would have been enough

many times over. The damage done can be approached from the point of view of someone who is only interested in whether the body can still regain its homeostatic equilibrium, and also from the point of view of a coroner who tries to reconstruct how that particular effect came about.

It is important to highlight here that only threshold effects can be overdetermined, but there seems to be no way to overdetermine the cumulated effect produced by overdeterminer causes together. The cumulated effect always goes beyond the threshold effect[47]. The amount of water that flows over the bowl is determined by the sum of all inflow and it is impossible to have more or less overflow than what the sum of all inflow minus the volume of the bowl provides.

To sum up, on the one hand, in known overdetermination scenarios there is a threshold effect to attend to, on the other hand in all of these cases, there is also a cumulated effect to consider. Empirically speaking, this is an objective fact concerning such situations: whether we attend to this or that kind of effect does not make any of those effects less real or objective.

---

[47] The crossing of a threshold expresses a change in one discreet property of an object. E.g. that it reached its melting point and from a solid state it became liquid. But, as the heat transferred was more than necessary for phase transition there is a cumulated effect as well, the exact temperature of the resulting liquid. The difference between the threshold effect and the cumulated effect can probably be expressed in terms of the determinable-determinate distinction. The determinate state of matter would be the cumulated effect of the two causes, while the determinable state (being liquid) would be the threshold effect. In many cases we are only interested in manipulating the determinables, but that does not mean that there are no necessary side-effects to any such manipulation in terms of more determinate descriptions of that property.

2.4.4   Genuine inter-level overdetermination and different notions of causation

Even though mental and physical causes are linked up by supervenience, if mental causes are truly autonomous both causes should be sufficient for the physical effect in themselves. Remember two things. First, for Kim (see especially in his 2007) causation is some kind of production that can be spelled out, for example, in terms of energy transfer. Second, that in his view mental properties or higher-level properties in general, in case they are not epiphenomenal, but still irreducible have to "bring with them new causal powers, powers that no underlying physical-biological properties can deliver" (Kim 1993:350). Here he relies on Alexander's dictum, according to which something to exist has to have autonomous causal powers. This means that in inter-level overdetermination scenarios both mental and physical token causes have to have autonomous contributions to an effect.

On the one hand, this contribution results in what we called the threshold effect and on the other hand, as together the contributions produce some excess over and above the former, a cumulated effect. The latter point follows straightforwardly if causation is understood as some kind of energy transfer. If causation is energy transfer, then overdetermining causes must transfer more than the energy one of those causes would have transferred alone. Under this interpretation of causation this result is a necessary consequence of how causation is understood and what overdetermination by separate causes means.

However, I don't think that the argument requires us to rely on physical causation, be it energy transfer or any other version of physical causation. The result follows even if one relies on a counterfactual theory of causation. It is interesting that an important critic of Kim, Loewer (2007) agrees with Kim in finding overdetermination problematic if causation requires

something like the transfer of energy, but also thinks that if causation is counterfactual dependence the problem evaporates.

It might be that counterfactual causation allows more space for getting rid of the problem of overdetermination, what I am sure about is that it does not help if the overdeterminers are separate existences. There is robust empirical evidence that redundant causes, independently of the interpretation we choose, together bring about an excess effect over and above the threshold effect that shows itself in the same space-time region where the threshold effect takes place that any of the redundant causes would have caused alone. The excess effect grows or changes with the number of redundant causes. So, it seems that it does not matter whether we prefer the counterfactual analyses or a physical theory of causation, empirically speaking the cumulated effect is there in available examples even if causation is not reducible to energy transfer.

Bunzl (1979) expresses the same basic idea from a conceptual point of view. He thinks that overdeterminer causes and their effects should be understood as joint causes of an effect that is underdescribed. A description of the same effect in terms of fine grained, modally fragile events reveals that each seemingly redundant cause of this effect is as necessary for it as are other causes that we usually relegate to the set of background conditions. The pouring of the two glasses of water into the bowl are both necessary for the fine-grained outcome, the kind of bowl overflow that takes place, as much as the ambient temperature in the room is.

Schaffer (2003) has a reply to this argument, which he borrows from David Lewis. Going fragile concerning events should be rejected as it conjures up spurious causes. Any tiny difference in the past (every event in the back light-cone) of an effect event can contribute to differences in its manner and time of occurrence, but causal discourse seems to dismiss those

contributions. I agree with this observation, I think it is fair to say that the threshold effect, the coarse-grained effect event is not caused by those spurious causes and as McDermott (1995) tried to show on experimental grounds, the majority of naive students do share this judgement. I agree with that, if the task for a philosophical account of causation is pure conceptual analysis, then it should respect those intuitions. But I think the solution to the issue at hand lies elsewhere. It has more to do with the perspective from which we ask our questions concerning causes and effects.

We can ask at least two kinds of different, but equally objective questions concerning all causal situations. First, what contributed to this or that particular outcome, an event under such and such a description. This is what the idea of a threshold effect captures. Second, we can also ask what difference is made to the outcome by certain manipulations we make on its potential cause events. So, we can approach the causal relation from the front end as well. We can ask what difference is made to a body by two bullets instead of one.

We only lose sight of certain objective dependency relations redundant causes are involved in in the world as relevant factors when we attend to a particular coarse-grained threshold effect and approach the causal relation from the back end. Presenting the same causal scenario from the second perspective shows that what counts as a redundant, non-difference-maker cause from the first perspective counts as a difference-maker for something the first perspective dismisses. In all the overdetermination scenarios we have discussed, the extra difference is made to some disregarded aspect of the same effect event. What I would like to highlight here is that the difference-making power of redundant causes doesn't dissipate into nothingness just because they are irrelevant for an outcome under a particular description that is interesting for us for some arbitrary purpose. It just brings about something that is unimportant from that particular perspective.

Even if causes are conceptualized in counterfactual terms, we would say that a cause is a cause whenever it makes at least difference, not just when it makes a difference to something we highlight as important to us. Finer-grained descriptions of the world help us to discover objective causal dependencies that are invisible from a coarse-grained perspective, and as my examples have shown that is of high importance when it comes to investigations into the origins of certain effects, like deaths.

2.4.4.1 Overdetermination by negative causes

One might still argue that my argument piggybacks on a simple feature of the examples. They are examples of positive causation involving contiguity and some form of energy transfer. But difference-making theories allow for negative causation as well: cases where the absence of something makes a difference. Can't we construct cases where absences are causing something without having an excess effect? Absences don't transfer anything to their effects, so this sounds like a fair starting point.

Let's investigate a well-known scenario where the omission of an activity causes an outcome. A gardener fails to water flower, and it dies (McGrath 2005). Let's turn it into a case of overdetermination. Imagine that there is a couple who own the flower. Both of them are watering the flower regularly, giving a glass of water every day. The flower requires a lot of water, so one glass wouldn't be enough to keep it alive. Imagine further that one day both members of the couple forget to water the flower and it dies. This is a case of overdetermination as the absence of both acts of watering is more than what is required for the flower to die. The absence of one would have been enough. Absences are not connected to their effects by physical transfer. The interesting question for the present discussion is, can they make a difference to the way the flower dies?

They can. Compare and contrast a scenario where only one of them waters the flower with the one where neither of them. The flower dies in both cases, but in a different way. In the overdetermined case it has less water in its cells, so the so-called turgor pressure of the cells is lower and therefore plausibly it looks way more withered than in the other case.

Note, that this has nothing to do with the kind of spurious causation Schaffer (2003) talks about. This is not about small irrelevant traces of faraway events accumulated in irrelevant parts of the respective effect event. Despite the fact that there is no transfer of energy between the gardener and the flower, fine-grainers like Bunzl (1979) would be forced to accept that the specific state the dead flower arrived at because of the absence of water is partially caused by what each of the owners is doing instead of watering, such as lying on the beach in a different country. But I wasn't attending to such minuscule traces when arguing for the necessity of excess cumulated effects from overdeterminers. I compared the differences that overdeterminer events make together and alone to the outcome directly. So, there is a strong case for the importance of such cumulated effects even if causation is analysed in terms of counterfactuals and negative causation is accepted as genuine causation.

2.4.5 Genuine inter-level overdetermination and physicalism

Let's see what follows from these insights to the case of inter-level overdetermination in the context of physicalism. As we saw, sometimes we are only interested in the cumulated effects of certain causes. The effect of pain on our behaviour could be a relevant case. Based on experience, we usually think that the amount of pain experienced is roughly proportional to the amount of crying and yelling that results. The amount of pain received from two sources, must be greater than the pain received from only one of those sources and has to have an increased effect on our yelling and crying compared to receiving pain only from one of those

sources. If there was mental-physical overdetermination it would have to work similarly. Physical effects of mental pain like crying and yelling are caused both by the mental state and its physical realizer and more pain means more crying.

Now, imagine that a group of neuroscientists make experiments to figure out the level of certain neural activities (the realizers of pain) with the amount of yelling and crying. Their results show the same relation as the relation we can measure between the feeling of pain and yelling. In case neural and mental causes are both real causes of the outcome, in the sense required by everyday examples of overdetermination, we should suppose that the measured crying and yelling experienced results from the independent contributions of two causes, mental pain and neural activity.

Any of the two token causes could trigger a certain amount of crying and yelling alone, but the mental and the physical cause together should cause more of the crying and yelling than what one of those cases would trigger alone simply because under normal circumstances the effects of pain accumulate. From this it follows that the accumulated physical effect of a mental and a physical cause together is not sufficiently caused by the physical cause of the same effect alone.

Because of the supervenience relation between levels there is no way to check empirically whether mental causation really does involve overdetermination or not. In our world the realizer and realized causes always go together, so the same effect is measured both by neuroscientists using brain scans and traditional psychologists, probably using a survey-based qualitative approach to measure mental pain. Therefore, there are no empirical means to decide this issue. However, what I have shown above is that whatever we measure as the effect of inter-level redundant causes, the measured effect has to have two real

components that both make a difference to the physical goings on in terms of the cumulated effect.

This result is surely unacceptable for physicalists. It flies in the face of the causal closure principle which is at the core of physicalism. According to the causal closure principle all physical effects are sufficiently caused by other physical events alone. However, the cumulated effect brought about by inter-level overdetermination is not sufficiently caused by purely physical causes. Therefore, inter-level overdetermination cannot be reconciled with physicalism.

Is there a possible answer to this argument? Can't we simply utilize the strategy based on disanalogies between intra-level and inter-level overdetermination? It is not impossible that the difference between cumulated and threshold effects exists only in intra-level cases and there are no cumulated effects for inter-level overdetermination? This is more or less one of Schaffer's (2003) answers to arguments by Bunzl (1979). He insists that we should accept the option according to which excess effects of redundant causes can simply disappear into thin air. "As any child can attest, we have perfectly clear intuitions about causation involving spell castings, and other fairy tale devices" (Schaffer 2003:27). To say this, Schaffer has to presuppose that spells can overdetermine outcomes without producing excess effects, or, in other words, anything more than the exact threshold effect itself. Maybe this is not true in our world, only in some far away possible world with different laws. So, it may be a possibility. But there is a further price to pay for choosing this solution. It increases the level of ad hocness involved in the defence of systematic inter-level overdetermination.

First, it is an ad hoc move, as it is utilized only to save mental causation. Second, it is an unparsimonious solution, as it still results in endless dysfunctional excess causes for effects in our universe. Third, the solution is even more ad hoc, because such kind of systematic inter-

level overdetermination is too unlike other cases of overdetermination. They are different from other cases in terms of the relation between the redundant causes involved which was required for explaining the systematic nature of this kind of overdetermination, and also in not having cumulated effects. Combining this with the other reasons amounts to a strong reason to reject this deviation.

By the argument above I have provided strong grounds for rejecting inter-level overdetermination in the context of physicalism. The argument based on the distinction between threshold and cumulated effects provides motivation for rejecting any kind of real inter-level overdetermination, as it shows that such overdetermination would be in conflict with physicalism. I would summarize these insights in the following way:

No innocent overdetermination*: When a physical effect E is caused by autonomous overdeterminer causes C and D (where both are sufficient for bringing about E) together, they always create some unique difference in the physical world over and above bringing about E, that is dependent on their joint presence.

This new principle is different from the original idea relied on in other formulations of the exclusion argument. The no overdetermination principle says that overdetermination is rare, so in general the presence of one cause gives a prima facie reason to think that there is no competing cause. The no innocent overdetermination* principle tells us that overdetermination might be a frequent phenomenon, but if it takes place in an inter-level context it cannot be innocent, it would be incompatible with the causal closure of the physical. This provides perfect rationale for accepting the exclusion principle in Kim's rendition of the exclusion argument (IV in section 2.2.).

As we saw in the section on causal closure (2.3) overdetermination would not be a problem for the emergentist. If there were higher-level emergent properties with autonomous, novel causal powers, the behaviour of a system calculated using only basic physical laws would be different from its actual behaviour (see section 2.3.3.3). The difference between the results provided by such a model and one enriched with trans-ordinal and emergent causal laws is similar to the difference between a world without inter-level overdetermination and one where there is frequent overdetermination by independent causes at different levels.

2.5   A hidden premise in the exclusion argument

When Kim formulated the first versions of his exclusion argument, he assumed it to be neutral with respect to different accounts of causation. As I pointed out in section 2.2.1, in his 1998 book, Mind in a Physical World he argued that the exclusion of higher-level causes in favour of lower-level physical causes follows independently of whether one relies on the concept of a sufficient cause or on a counterfactual approach to causation. In retrospect it is easy to see that this is not true. I will get back to this issue later in this section and chapter 4 is devoted to the topic of exclusion in light of counterfactual theories of causation. In Kim's 2005 book and especially in his 2007 paper on "Causation and Mental Causation" he started to advocate physical theories of causation in contrast to "mere" counterfactual approaches and approaches based on lawful necessitation or regularities, even though he seemed to have held the latter kind of idea in mind back in the 1990s.

Kim's motivation to change his view concerning causation was at least twofold. First, some philosophers started to argue that difference-making causation is not only strong enough to be used in arguments for some kind of mental realism, difference-making notions are better than their competitors. Kim definitely wanted to answer that challenge. Second, he realized that regularity theories are not strong enough to bear the weight of the exclusion argument. In this respect, those who think that counterfactual causation is satisfactory and should be preferred are in agreement with him.

However, reading his texts it is hard to avoid the impression that Descartes' exchange with Princess Elizabeth is a defining story for Kim, as he recounts it whenever he discusses mental causation. The problem Elizabeth pointed to is that contact is a necessary condition for physical causation in Descartes' metaphysics, and meeting this condition is simply impossible for immaterial souls. Kim's preference for physical causation or, using his terms, a

"thick" notion of causation in contrast to a "thin" notion based on counterfactuals seems to go back to this story. As he says:

> "Such *contact, in more modern terms, represents the* imparting of energy, or *transfer of momentum, from one body to another*, and this fact constitutes the causal action." (Kim 2007:244, my italics)

When he says more about why we should prefer this "thick" notion over others, on the one hand he points to certain internal problems generated in the counterfactual framework. He points to problems of negative or absence causation and so-called late preemption scenarios. I will discuss these in some detail later in this section. On the other hand, he tries to convince us that accounting for human agency requires his preferred notion:

> "Why should we resort to this "thick" variety of causation in thinking about mental causation? My answer is pretty simple: We care about mental causation because we care about human agency, and *agency requires the productive/generative conception of causation*." (Kim 2007:257, my italics)

In this section, I will show that the later point is dubious, and also that there are strong reasons to choose counterfactual theories of causation over Kim's preferred concept of causation. The main argument for difference-making theories based on counterfactuals is that they can account for causal relevance whereas all physical theories of causation suffer from problems with respect to relevance. Moreover, most of the problems counterfactual theories face can be accounted for in a promising manner.

As far as I know, Peter Menzies (2003, 2007) was the first person to present a systematic reformulation of the exclusion argument based on an alternative theory of causation. The replacement of the notion of a sufficient cause in the exclusion argument by

another, supposedly more legitimate difference-making notion was a quite radical idea when it was proposed for the first time. There were earlier attempts to achieve something similar to what Menzies started, arguing for serious higher-level causal autonomy as a physicalist, but those attempts were based on different conceptualizations of the relation between realizer and realized properties.

In his classic paper, Yablo (1992) rejected the conclusions of the classic exclusion argument, but to achieve that result he relied heavily on the idea that the realizer-realized relationship is the determinable-determinate relationship. Even though he seems to have inspired a lot of people, probably Menzies among others, these days his view is considered to be problematic and, as we will see in chapter 5 section 5.4, there are convincing arguments against it.

To mention another important example, Jackson and Pettit (1990) had an interesting but probably unsuccessful idea to save mental causation. As Menzies explains in a critical assessment of the so-called program explanation project, according to that view, a higher-level mental state can only be causally relevant to an effect "through non-causally ensuring that there is some lower-level physical state … that is causally efficacious in producing" (Menzies 2007:29) the effect in question. So, higher-level causes and properties lack real causal efficacy and causal powers: only the physical realizer states have that. The metaphor of "program explanation" comes from computer science where higher-level functional roles spelled out in terms of higher-level program languages "program for", guarantee the presence of lower-level physical, electric circuit states that have the power to produce whatever relevant outcome. This picture fits the spirit of Kim's exclusion argument. It might be that the functionally defined role provides a high-level of modal resilience for functionally

defined mental states, but they have no autonomous causal efficacy, no causal powers of their own. Their powers are inherited from their realisers.

To endow mental causes with genuine causal powers a more radical shift is required, one that questions the priority of physical or realizer-level causation over everything else but lacks such questionable commitments as Yablo's solution. For the purposes of spelling out such a solution the most important premises of the exclusion argument are those that either implicitly or explicitly refer to causation, so no overdetermination, the exclusion principle and the causal closure of the physical. According to the exclusion principle when there is a competition between a physical cause and a mental cause we have to choose one among the two. Causal closure says that every physical event that has a cause at time t has a sufficient physical cause at t and because of that we are forced to choose physical causes.

The notion of a cause played a role in my discussion of the no overdetermination principle. There I argued that if overdeterminer causes are separate, autonomous existences, then it doesn't matter whether one chooses an account of causation based on the intuition that causes produce effects, like the energy transference theory or an account based on the difference-making intuition, like counterfactual theories of causation. The prohibition against overdetermination makes good sense in both cases. This is not the case when it comes to closure or to the exclusion principle.

Regarding the theory of causation relied on in the causal closure principle the landscape is different. Under some interpretations of the causal relation it might turn out that physical effects having a cause are caused by physical cause, but it is also possible that certain physical effects are not caused by physical properties or at least this is what people like Menzies (2013) or Woodward (2008) tried to show us. They questioned the legitimacy of using the notion of a "sufficient cause" to formulate the exclusion argument itself and claimed that

by substituting "sufficient cause" with a proper difference-making notion the argument will yield the opposite result to the one Kim intended. My investigation of Kim's texts has shown that one can find at least two different interpretations of the term "sufficient cause" in Kim's writings (1998, 2005, 2007) at places where he tries to make his notion of a cause more precise. What he presupposes in his later writings is that causation should be understood as some kind of production. So, it is fair to say that there is a hidden premise in the exclusion argument:

Causation is production*: causes bring effects about in a push-pull fashion. The relation involves some kind of contiguity and some kind of transfer between causes and effects.

In the following sections I will investigate the justification provided for this premise, by reconstructing the arguments Kim has offered in its favour and also considering received arguments against his conviction. My conclusion will be that an analysis based on counterfactuals provides a stronger account of causation, a result that delivers justification for those in favour of rejecting the hidden premise.

2.5.1 Nomological necessitation and sufficiency interpretations of causation

The first question on our path to understand the causal side of the exclusion problem is this: what is a *sufficient cause*? The first interpretation of a sufficient cause that can be found in Kim's early writings is based on the concept of nomological necessitation (Kim 1992), although he rejected this approach in later works (Kim 2005, 2007) it is interesting to see how that worked. First, it convinced a lot a people and interestingly enough it suffers from similar problems as those theories Kim started to favour in his later works.

According to the nomological necessitation interpretation C causes E in case there is a law of nature such that C necessitates the occurrence of E if C occurs. This is considered to be an old-fashioned concept, even in the late Kim's own view, that dates back to the heydays of the covering law model of explanation. On the one hand it requires too much, as law-like necessitation is not a characteristic of causation or at least causation at non-fundamental levels, not even at the level of chemical reactions. Any "law" that describes the behaviour of mereologically complex objects is hedged with ceteris paribus clauses as there are background conditions under which the relation would not apply. So, it can be argued that this concept is too strong: causes do not necessitate their effects nomologically, at most they necessitate their effects if certain background conditions hold. Before turning to further problems this interpretation faces it's better to go through a different possible interpretation because there is a common problem these interpretations share.

Even though Kim never refers to this theory, it seems natural to try to interpret the concept of a sufficient cause based on Mackie's INUS condition analysis of the concept of a cause. Mackie (1974) formulated a theory according to which C causes E if and only if C is an Insufficient but Necessary condition as part of an Unnecessary but Sufficient system of conditions for E. The whole analysis is based on the notions of necessary and sufficient conditions. Let see how this could provide us with an interpretation of the concept of a sufficient cause. According to Mackie a system of conditions (A & B & C) can be sufficient to bring an effect E about. C or any other member of this system of conditions is insufficient in itself, nut necessary in the sense that without it the remaining conditions are insufficient to bring E about. (A & B & C) as a system of conditions is unnecessary for E as a different system of conditions (X & Y & Z) could do the same work.

Under this interpretation a sufficient cause is the presence of a set of causal factors that together are sufficient to bring the effect about. But this is not exactly how Kim talks about a sufficient cause, as whenever he provides examples he talks about a single causal factor, e.g. the presence of the physical realizer of a mental property. But I think, charitably read, he simply highlights one factor from the set of conditions that is jointly sufficient. It is common practice to talk about "the" cause of an outcome bringing it into the foreground while pushing the other factors into the background. Mackie, Lewis and many other philosophers agreed that this practice is understandable as a pragmatic feature of causal talk. This practice is a product of our human need to highlight what is relevant, informative in the particular context of communication. If Kim's term of a "sufficient cause" is interpreted broadly, as an elliptic reference to a set of conditions together sufficient for an effect, it starts to make good sense.

The main problem with both the INUS-based and the nomological necessitation interpretation of a sufficient cause derives from the fact that we are talking about regularity theories of causation. In this respect the INUS theory is exactly like the concept based on nomological necessity: the criticisms put forward below apply to both theories. Regularity theories not only fail to handle problems concerning the direction of causation, but they also have a h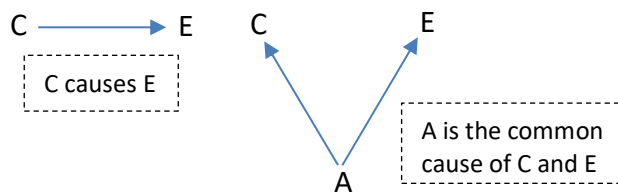ard time accounting for causal relevance. The problem of relevance can be formulated in the following way. It is possible that the effect E is always present when C, the cause is present, but it happens because E and C have a common cause, not because C causes E (see: Figure 2.5-1). When that is true, the presence of

Figure 2.5-1

C can still be interpreted as an INUS condition for E, also as being nomologically necessitated by C, but no one would say that it is a cause of E.

There are some other well-known cases in which there is a constant conjunction between C and E, but no causal connection between them. The classic example comes from Salmon (1984). A man takes birth control pills (C) to prevent pregnancy (E). The example is trivially non-causal, as a man is incapable of becoming pregnant, it is biologically not possible. Nonetheless, whenever C is present E is present, so, according to the regularity-based approach C should be a sufficient means to avoid pregnancy. The general problem is that an analysis based on regularities, cannot avoid counting non-causes as causes. And that is a problem for both the nomological necessitation view and the INUS analysis.

2.5.2 Counterfactual theories of causation to the rescue

Historically, the problems summarized above were exactly the kind that created the need for new, better accounts of causation. Reading Lewis' classic papers on causation, one might find that he talks too much about asymmetry and relevance, issues that are less interesting in the contemporary context. The reason why these discussions were more interesting in the 70s and 80s is that there were no viable solutions to these problems. The advent of the counterfactual analysis became an important paradigm shift exactly because it offered solutions for the pressing issues of the day. Even though, as today is apparent, it generated further problems, according to many contemporary philosophers the best candidate for analysing causation is still some version of the counterfactual theory. This is an important reason why Menzies and List decided to rely on a counterfactual theory in criticizing Kim's exclusion argument (Menzies &List 2009, 2010).

Counterfactual theories build on the so-called difference-making platitude of causation. Thinking of causes as difference-makers in terms of counterfactuals provides at least the following advantage. The approach avoids the problem of common causes (see: Figure 2.5-1) and the problem of empty conditions such as the case of the birth control pills used by a man. In counterfactual parlance the following expectation can be put forward: if one were to change C, E would change as well. Or to put it into more precise form (following Lewis 1973):

The **simple counterfactual theory** of causation[48]:

Assuming that C and E are two distinct, non-overlapping possible events, E is causally dependent on C if and only if it is true that if C were to occur E would occur and if C were not to occur E would not occur.

It is easy to see that if C and E are connected via a common cause, a change in C leaves E unaffected. By not taking birth control pills the man cannot raise the chances of his pregnancy.

Counterfactual theories also allow that causes necessitate their effects, but do so only under certain background conditions. In other words, the causal relations described in terms of counterfactuals are bound to certain specific contexts. This is in line with our everyday intuitions with respect to causal relations, and more than that it provides a better account of causal discourse in the special sciences.

As counterfactual theories provide a viable account of causal relevance the claim made by critics of the exclusion argument like Menzies that it would be better to reformulate the argument based on a difference-making theory of causation sounds justified. But in his

---

[48] For simplicity's sake, I omitted one basic but important constraint. The prohibition on backtracking counterfactuals. In Lewis' framework this ensures the time asymmetry of the causal relation.

later writings Kim turned to a more developed theory of productive causation. First, attacking the counterfactualist, he made use of Ned Hall's (2004b) analysis who have shown that many cases of causation cannot be handled using counterfactuals and also that this suggest the need for an alternative, productive notion of causation. However, Hall argued for a pluralistic view of causation because in his view cases of absence causation can only be handled in a counterfactual framework. Kim rejects this pluralism and argues that absence causation is not causation worth having. In the end, he prefers the so-called conserved quantity theory, a physical theory of causation that falls into the category of production views. Therefore, we need to check how a counterfactual theory compares with such a theory before a decision can be made about which theory of causation should be used in the exclusion argument.

2.5.3   Productive versus difference-making causation

Kim says little about the details of his productive view of causation, he mainly refers to theories already on offer. What he says explicitly is that production involves "real connectedness" in terms of a spatiotemporally continuous process. Investigating this idea, he starts out from an intuition, or platitude concerning causation that was identified by Anscombe:

> "There is something to observe here, that lies under our noses. It is little attended to, and yet still so obvious as to seem trite. It is this: causality consists in the derivativeness of an effect from its cause. This is the core, the common feature, of causality in its various kinds. Effects *derive from, arise out of, come of, their causes*." (Anscombe 1971:91, my italics)

> For Kim this platitude is more basic than the difference-making platitude. We should note that for Anscombe causation in the sense of production is taken to be a primitive and no

further explanation is provided for its content. This is definitely a legitimate move, but a richer theory, such as a counterfactual theory may have the advantage of providing us with deeper insights concerning the content of causal statements. Trying to add layers of content to his understanding of causation Kim moves on to search for some backing for the idea of causal connectedness. His first reference is to Hume's contiguity condition for causation which says that there has to be "contiguity in 'space and time' between cause and effect, either direct or mediated by a chain of contiguous cause-effect pairs" (Kim 2007, 243). Kim finds it important to highlight that even Hume's theory, which is usually taken to be a regularity theory, had this extra requirement.

To enrich his notion further he turns to Ned Hall's discussion of the classical counterfactual theory. It is worth reconstructing the main argument of Hall's article as a point of reference for later discussions on causation. First, he shows that the simple counterfactual theory faces difficulties when it comes to so-called pre-emption scenarios. Let us imagine that
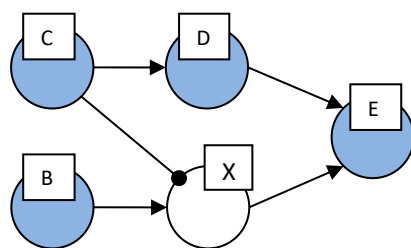
Figure 2.5-2

there are two assassins, a master (B) and an apprentice (C), planning to kill a dictator (E). They have an agreement. If the apprentice shoots and successfully kills the dictator the master remains in hiding, but in case the apprentice misses the target the master finishes the job. In our case, depicted on Figure 2.5-2[49], the apprentice does the job, so the master remains in hiding. This is called an early pre-emption scenario because B gets pre-empted by C early on in the process.

---

[49] On neuron diagrams circles signify events, full circles are events that took place, empty ones are events that did not. Normal arrows signify a manifested power producing another event, oval arrows signify prevention. Concentric circles signify stubborn neurons that require as many incoming signals as the number of concentric circles they have.

In the context of the classical theory, in a scenario like this event C only counts as a cause of E in case the chain of counterfactual dependences from C to E (C □-› D & D □-›E) is stipulated to be a transitive causal chain, since E is not dependent on C counterfactually. If the apprentice had failed the master assassin would have helped him out, so E would have happened even without the action of the apprentice. If one wants to incorporate such cases into the counterfactual analyses, modifications of the simple theory are required. Fortunately, it is true that E is dependent on D (the bullet flying from the apprentice's gun towards the dictator) and D is dependent on C. This observation suggested a solution to Lewis, defining causation as the ancestral of stepwise counterfactual dependency[50]. As there are numerous counterexamples to the transitivity of causation this stipulation created much tension in the classical Lewisian theory[51]. And this modification wasn't enough.

Another problem case reared its head, the kind they called a late pre-emption scenario. Imagine two youngsters, Billy (C) and Suzy (B), throwing rocks at a window, but only
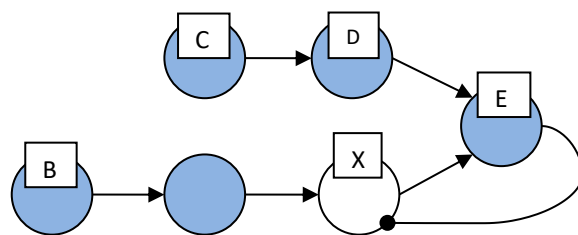
Figure 2.5-3

nearly at the same time. Billy throws first, and his rock reaches the window just right before the other rock. So, the window will be broken when Suzy's rock reaches the now empty location of the shattered window (see: Figure 2.5-3). According to Hall (2004b)

---

[50] Lewis had two reasons to define the causal relation this way and by doing that making transitivity part of the definition. Naturally, he wanted causation to be transitive, because that reflects our everyday conviction. But, counterfactuals seemed to be intransitive and early pre-emption scenarios required this stipulation as well.

[51] My chapter 3 provides a short summary of the problem of transitivity in counterfactual theories of causation. The problem of transitivity itself was among the main motivations for developing more complicated counterfactual theories. Today most theorists reject the transitivity of causation trying to find a solution that makes it is a safe assumption that at least certain causal chains conforms to transitivity. This is exactly what I aimed to do in section 3.6.

the explanation for our evaluation of this scenario, the reason why we think that C is a cause of E and not B, requires more than transitivity. For the purposes of the present discussion, it is enough to see that in a late pre-emption case not only is E not dependent on C counterfactually, it is not dependent on D (the rock flying towards the window) either. So, Hall proposes that a satisfactory solution requires further relations to hold between the cause and the effect over and above transitivity. He suggests locality (contiguity in space and time) and intrinsicness[52]. The three conditions together add up to what Hall calls a productive causal relation. But for some unknown reason Kim only picks up on the locality condition:

Locality condition: "causes are connected to their effects via spatiotemporally continuous sequences of causal intermediates" (Hall 2004b:225)

The mentioned limitations of the simple counterfactual analyses provide justification for Kim's attitude towards counterfactual theories of causation, but we shouldn't forget that in Hall's view the productive notion of a cause is far from being enough to cover all cases of causation. Cases of absence causation require an account purely based on counterfactual dependence. It is easy, and for present purposes it is enough, to see that the locality condition fails in cases that involve absence causation[53] as the absence of an event cannot be connected in space and time to a concrete event. If both the concept based on dependence and on

---

[52] This roughly means that by carving out a chain of events from this world and placing them into a different possible world with the same laws of nature in place we would not change what causes what. Note, that intrinsicness is not true in the case of negative causation (see: Hall 2004b for more).

[53] In Hall's rendition of the problem the same goes for intrinsicness. Transitivity also fails in some cases such as double prevention scenarios. I won't touch on these issues here. Chapter 3 discusses the transitivity problem for cases like double prevention in the contrastive theory of counterfactual causation. There, such scenarios are called short-circuits.

production is required to account for causation then this sounds like a stalemate, supporters of the production view can get the cake, but they can't eat it.

However, many philosophers, including Kim, reject the view that absence causation is real causation. It is true that, there is something metaphysically abhorrent in thinking that the absence of something can be causally efficient or, to use a formulation that emphasizes the absurdity of the thought even more, can manifest a power to bring about the presence of something else. But whatever we think about this issue the analyses of absence causation in terms of counterfactual dependence renders it meaningful, so some work is required to explain it away. Let's take a look at an example first.

To show that absence causation is not just a figment of everyday causal talk, and that it requires serious consideration friends of absence causation like to introduce it by examples from the natural sciences. The example I will introduce is a case of a so-called double-prevention scenario. It involves the prevention of a preventer event. Preventers cause the absence of something, so it is a perfect case of absence causation.

I will use the famous Jacob-Monod model of the mechanism by which E. Coli bacteria manage to break down and to digest lactose discussed by Woodward (2003). The function of the mechanism is to set in motion the production of certain enzymes, but only in the immediate presence of lactose. It would be wasteful for the cell to manufacture the enzymes that digest lactose when it isn't available. On the left-hand side of Figure 2.5-3[54], on the neuron diagram, O refers to the lac operon gene sequence that is responsible for the

---

[54] The picture on the right was developed based on an illustration from: Lodish H, Berk A, Zipursky SL, et al.: *Molecular Cell Biology*. 4th edition. New York: W. H. Freeman; 2000., figure 10-2.; It is a close replica of the upper left side of the original figure. Another useful description can be found here:
http://www.ebi.ac.uk/biomodels-main/static-pages.do?page=ModelMonth%2F2006-12

production of the enzymes capable of breaking down lactose. R refers to incoming lac repressor proteins produced by a different gene sequence to block lac operon in the absence of lactose. L is lactose gathered from the immediate environment. When there is no lactose in the environment, an E. Coli bacterium produces a lac repressor that attaches to the beginning of the lac operon sequence and blocks the translation of digestive enzymes. If there is lactose around the blocked operator region of the lac operon sequence it dissolves the block and other incoming lac repressors which results in the production of lac mRNA, the first step towards building the enzymes required. P is the event where the presence of lactose (L) prevents the repressor (R) to block the lac operon region by dissolving the repressor itself. This is a perfect case of double prevention. Lactose (L) prevents the repressors (P) to reach the operator region and by doing so prevents the reading of the lac operon sequence, which allows the reading of the lac operon sequence and initiates enzyme production (m). Lactose prevents the repressor, a would-be preventer, doing its job, so we get double prevention.
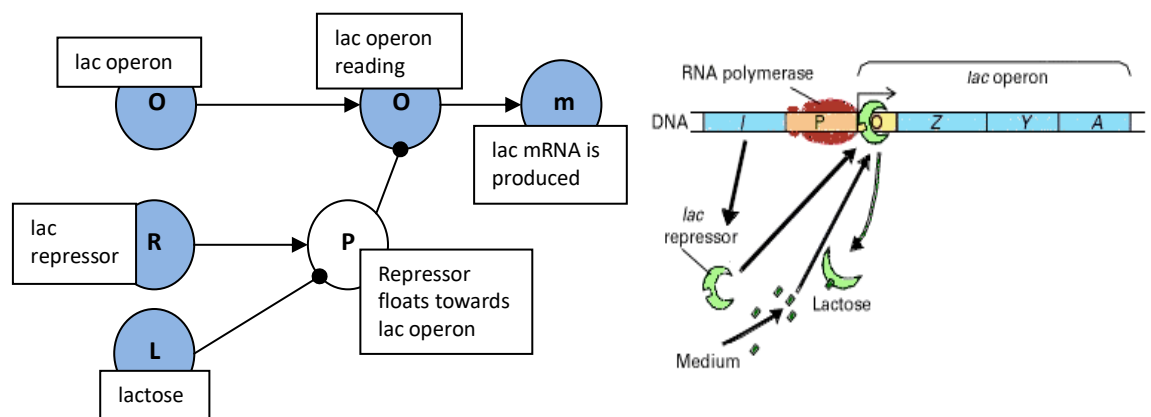

Figure 2.5-4

Friends of the realist interpretation for absence causation, such as Schaffer (2004), think that the use of absence causation talk by natural scientists provides an argument for taking it more seriously. I disagree: the absence of the Lac repressor as a cause is as weird as absences are in any such everyday scenario. Certainly, there are many options to explain what

we mean by absence causal talk. Here, I will choose an option that might seem to turn the tide in Kim's favour, but only to show that even if one rejects the realist interpretation of absence causal talk[55], the counterfactualist still has satisfactory arguments in favour of his view.

We might argue following Beebee (2004) that absence causation is not a case of ontologically serious causation as such, but only of causal explanation. Following Lewis, by explanation she means that mentioning a negative cause provides some information about the causal history of the effect indirectly, not by naming the concrete causes that brought the effect about. The information is delivered via a comparison suggested by the description of the absence. In our E. Coli example, we should compare cases where repressor molecules are not blocking the production of digestive enzymes because of the presence of lactose, and cases where they are blocking production. The information delivered by the comparison is this: as everything else is the same in the two cases, where there is no repressor blocking translation things must be brought about by what is present and what does manifest its powers under the circumstances[56]. So, the outcome is brought about by those factors that causal talk pushed to the background.

---

[55] Kim relied on at least one further argument, the argument from too many negative causes. It sounds natural to say that my not watering my flower caused it to wither. But if causation is counterfactual dependence then the same goes for Donald Trump not watering my flower and the UFOs not watering it. Had they watered it, it wouldn't have withered. Even if there is a method to single out the most relevant among these similarly as in the case of positive causation there is an important asymmetry between positive and negative cases of causation. Positive factors are concrete occurrences and that limits the circle of possible positive non-overlapping contributing factors. There is no such limitation on the side of negative "factors". On the dependency interpretation of causation any negative (absence) condition, however absurd or surreal the case might sound like, counts as a cause of the outcome.

[56] Theorists like Schaffer (2005) think that absence talk is referential. The description "not watering the flower" refers to the event that took place instead of watering the flower in the actual world, e.g. "napping in the back

2.5.4 The transference of conserved quantities and the problem of relevance

What was said above provides dual motivation for accepting the concept of causation Kim preferred in his later writings (2005, 2007), reacting to early criticism from Menzies (2003) and others, a version of process theories, the so-called conserved quantity theory that falls into the category of production theories. First, this theory stands a good chance of explaining our intuitions concerning both early and late pre-emption scenarios as the concept involves spatiotemporal contiguity between causes and effects. Second, it can easily agree with the diagnosis concerning the status of absence causal talk. Absences are not causes, but by talking about them we can deliver genuine causal information to others.

A common feature of theories of causation that build on the idea of a physical transference between causes and effects is that they don't aim for a reconstruction of causal talk. Their supporters, walking the revisionist path of the naturalist, base their notion of a cause on some interpretation of contemporary physics. The best interpretation of the later Kim's idea concerning causation is that in his conviction some physical transference view of causation provides the actual-world grounding of a more abstract production view of causation. In what follows, I will summarize the gist of the conserved quantity account that is still the most popular process interpretation of causation to be able to show that, when it comes to causal relevance, it suffers from the same kind of problem regularity accounts suffer from. After that, I will allude to some viable solutions for pre-emption scenarios in later versions of the counterfactual theory of causation to show that much of Kim's complaints against counterfactual theories lost motivation.

---

garden". But even if it is referential it is still true that there is no space-time contiguity or relevant physical transfer between this event and the "withering of the flower". A "nap in the back garden" has no potential power to make flowers to wither.

I will start by summarizing the essence of Dowe's physical theory of causation. Dowe (2000) defines causal processes and causal interactions in terms of conserved quantities. Conserved quantities are physical quantities under the governance of a conservation law of physics. Linear momentum or charge are straightforward examples. In Dowe's conceptualization a physical process is a world-line of an object possessing a non-zero quantity that is conserved. The concept of a world-line originates in relativity theory. It is the path an object traces in spacetime (in 4 dimensions). It is a sequence of events in spacetime corresponding to the history of the object. A causal interaction takes place when the world-lines of two objects cross each other and an exchange of conserved quantity takes place between them (e.g. of linear momentum). This account was set out to capture objective features of causal interactions in our world as they happen according to our best physical theories. It is not interested in causal mechanisms in faraway possible worlds nor in pure conceptual analyses.

Let's introduce a counterexample that is widely used in the physical causation literature. Imagine a man playing a game of billiards. He hits the cue ball with the cue and the cue ball projects a red ball into one of the corner pockets on the table. Earlier the player has chalked the end of the cue with chalk. When making the strike some of the chalk sticks to the cue ball. Some of the same chalk spot gets transferred to the surface of the red ball when the cue ball bumps into it. At the end the red ball lands in the corner pocket. The question is, what caused the red ball to land in that corner pocket? Intuitively the push by the cue ball that originated from the cue. A more scientific explanation would refer to the linear momentum of the cue ball. Can the conserved quantity theory deliver these results?

Both parts of the chalk that came from the cue and the linear momentum are getting transferred to the red ball in the interaction between the red ball and the cue ball. There are

two different conserved quantities travelling through the billiard table. According to the conserved quantity theory a causal interaction is an exchange of conserved quantities. Two exchanges took place, but the theory provides no resources to decide which is relevant for the outcome, they are indistinguishable in its terms. Woodward explains this nicely:

> "*the feature that makes a process causal* (transmission of some conserved quantity or other) *tells us nothing about which features of the process are causally or explanatory relevant to the outcome* we want to explain. For example, a moving billiard ball will transmit many conserved quantities (linear momentum, angular momentum, charge, etc.), and many of these may be exchanged during a collision with another ball. We still face the problem of singling out the linear momentum of the balls, rather than these other conserved quantities, as the property that is causally relevant…" (Woodward 2003:357, my italics)

What makes the situation even more difficult for believers of the conserved quantity theory is that there are cases of higher-level causation where there appears to be no conservation law governing the properties involved. Imagine that my anger makes my friend sad. On the face of it, the conserved quantity theory makes my anger causally impotent. One can always revert to the idea that there has to be an underlying physical process at least at the microphysical level realizing my anger that involves some conserved quantity, but that does not help to decide question of causal relevance.

The problem identified here is similar to the problem regularity accounts faced concerning relevance. A comparison with what the counterfactual analyses can deliver in such simple cases as the one involving billiard balls speaks for itself. The sentence "had the chalk not transferred to the surface of the red ball the ball would have ended up in the same corner

pocket"[57] is plausibly true, therefore the chalk was not relevant to the outcome. Whereas if the sentence "had the cue ball had just slightly different linear momentum it wouldn't have ended up in the corner pocket" is true it clearly shows the relevance of linear momentum for the outcome. So, the counterfactual account is still superior in terms of relevance.

2.5.5 Counterfactual theories save the day again

Nevertheless, the question remains, do we need something that can only be provided by physical theories of causation to solve the problems raised by pre-emption scenarios? Fortunately, two important shifts happened in the last two decades. First, philosophers, including the late Lewis (2000) himself, realizing the tension in the classical counterfactual framework, explained vividly by Hall (2004b), started to develop new solutions for pre-emption scenarios. Second, the availability of new solutions to pre-emption scenarios allowed believers of the counterfactual analysis to give up on the transitivity of causation as well which allowed a more consistent treatment of counterexamples to transitivity (see: Hitchcock 2001, Woodward 2003, my chapter 3). I will recount only one solution here close versions of which were developed independently by Yablo (2004) and interventionists like Woodward (2003). The solution is more straightforward in an interventionist, model-based rendering, but the formal apparatus required to introduce that is way more complicated. Here, I will restrict myself to making sense of the main idea.

---

[57] In a tiny percentage of possible cases the chalk dust could make a difference, but let's just suppose that in the example at hand the player had a firm hand and the chalk dust had no chance of making a difference with respect to the outcome.
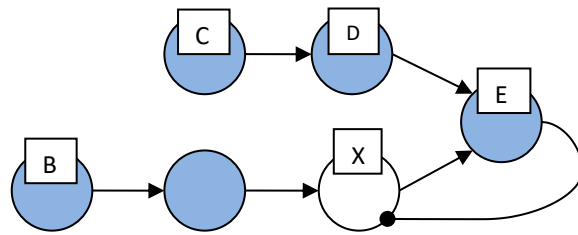
Let's go back to our example of late pre-emption with the two kids throwing rocks at a window (take a look at Figure 2.5-3 again: there is a replica to the left). I will demonstrate the new solution on this scenario, because this caused more sleepless nights for philosophers working on causation. Intuition tells us that Billy's throwing (C) was the cause of the shattering (E), but Suzy's throw (B) was not. The following procedure delivers this verdict. To test whether Billy's throwing a rock (C) caused the window to shatter, we should examine the behaviour of the process running from C through D (Billy's rock hitting the surface of the window) to E holding events extrinsic to this process fixed in their actual states. In our present case, we should "freeze" X (Suzy's rock hitting the surface of the window a tiny bit later then Suzy's). If this procedure allows C to make a difference to E then C is a cause of E. It is easy to see that the following counterfactual is true: if Billy hadn't thrown a rock and Suzy's rock hadn't hit the window (X froze), it would not have shattered.

We get the opposite result using the same procedure carried out on the process between B and E. To see if Suzy's throwing a rock (B) caused the breaking of the window, we need to "freeze" extrinsic processes like D in their actual states again. Than we need to evaluate the counterfactual that follows from the operation: if Suzy hadn't thrown his rock and Billy's rock had hit the window (D froze), it would not have shattered. This counterfactual is clearly false.

The asymmetry in the truth values of these counterfactuals explains both kinds of pre-emption scenarios we discussed. Accepting this solution allows us to question Kim's preference for productive causation.

Where are we now? On the one hand, counterfactual theories seem to have answers to questions concerning the causal status of pre-emption scenarios. At the same time, they outperform all known production theories when it comes to causal relevance. This is quite promising. On the other hand, counterfactual theories are seemingly compatible with the stance according to which absence causation is not ontologically serious causation, to use Dowe's term, it is only quasi-causation and only provides causal explanation. Therefore, it is justified to take counterfactual theories more seriously than they were taken by Kim. Whereas production views should be handled with more caution as they don't seem to perform well in some of the most important tasks for a theory of causation. If most problems of counterfactual theories have feasible solutions and they account for causal relevance, we are well justified to use them reinterpreting the exclusion argument.

2.5.6   Causation and agency

There is one remaining issue I should discuss. As I mentioned in the introduction of section 2.5, Kim contends that a viable interpretation of agency demands a productive view of causation. He says: "We care about mental causation because we care about human agency, and agency requires the productive/generative conception of causation" (Kim 2007:257). This might provide an argument for preferring productive causation at least in the context of the mental causation debate that seems to be independent of the general considerations I discussed in previous subsections.

I do think that there is something right about Kim's intuition, without being able to produce, or generate things we would be impotent as agents. However, the whole argument remains at the level of intuitions, Kim does not provide an account of agency and he seems to overlook important cases where agents achieve their goals exactly by means of omissions,

by not doing things or by not preventing certain things from happening and thereby allowing them to happen. Examples of such cases are easy to find in the context of moral action. An agent aims to control certain aspects of the environment and that control can be exercised by both doing things in productive manner and by not doing things and thereby allowing certain things to happen. Productive notions of caution can only account for the first of these options, whereas counterfactual theories can account for both. That is one important argument against Kim's idea and for the use of counterfactual theories in the more specific context of the mental causation debate.

More than that, as some philosophers argued in recent years (e.g. Gibb 2013, Russo 2016), mental causation involves double prevention, and it does so systematically. The idea is roughly this. Our will is a preventive factor in our mental life. In many cases, we do X because first we have removed, we prevented a conflicting motivation from manifesting itself and thereby preventing X to happen.

As we saw above, in double prevention scenarios causation takes place without the transference of a conserved quantity between the first cause and the final effect and this is the reason why some philosophers tried to explain away such scenarios. But this is not a good enough reason to reject such scenarios as examples of genuine causation, because it presupposes that causation is some kind of production, and so it begs the question against the proponents of counterfactual accounts. If Kim wants to maintain the view that agent causation should be productive causation in all cases, he has to show us that something about agency or the mental realm makes productive causation the only game in town. As I don't see such reasons on the horizon, I take it that at present it is well justified to reject production views of causation as accounts that are preferable when it comes to the description of mental goings-on and the control functions that agents carry out to survive.

2.6   Concluding remarks concerning Kim's exclusion argument

As we saw in section 2.2.1 initially Kim believed that a sufficiency account of causation and a counterfactual account of causation delivers the same results concerning causal exclusion. As section 2.5 shown, later, when pushed by others, he confessed his sympathies for productive, more precisely, conserved quantity theories causation, even though he himself recognized that the exclusion argument becomes an empty truism if one runs it based on that assumption as section 2.3.3.1 explained. The last section summarized the most general arguments that provide grounds for preferring a counterfactual, difference-making account of causation. We still need to explore the consequences of these insights to the exclusion argument. Chapter 4 takes up the job of exploring a new version of the exclusion argument that relies on a counterfactual theory of causation.

However, counterfactual theories are not without internal difficulties. One such issue concerns the transitivity of causation. Usually this feature of causal chains is presupposed by our discussions of causal processes. David Lewis' classical counterfactual account stipulated that causation is transitive, but the approach proved to create more problems than it solved. In the next chapter I will dive into this area of difficulties and I will try to suggest a solution to the problems created by counterexamples to transitivity.

Chapter 3: A problem for counterfactual theories of causation

3.1   The transitivity of causation in contrastive counterfactual theories

This chapter criticizes the way that counterexamples to the transitivity of causation are handled in counterfactual based contrastivist theories of causation and also provides a brief history of the importance of the topic. In particular it reconstructs and rejects the different ways that proponents of the theory have tried to handle so-called short-circuit scenarios, which are one kind of well-known counterexample to the transitivity of causation. The reason why I investigate the issue of transitivity in this context is that one important promise of contrastive theories was exactly to solve the tension created by transitivity in the classical counterfactual theory developed by Lewis.

Some adherents of contrastivism, like Cei Maslen, claim that they can provide sufficient conditions for demarcating non-transitive cases of causation from transitive ones, whilst others, like Jonathan Schaffer, argue that they can explain away all counterexamples. The chapter analyses the existing accounts of so-called short-circuit counterexamples in the contrastivist framework to show that both accounts fail. After that it goes on to argue that as there is no other viable method on the table to explain away these counterexamples they can only be demarcated from proper transitive chains on grounds of their structural features.

Cei Maslen [2004] utilizes a criterion worked out by David Lewis for the intransitivity of counterfactuals. Below, I show that the suggested application of this backwards counterfactual criterion disregards the context sensitivity of the causal links involved and that the criterion holds true not only in unproblematic cases of transitive causation but also for the discussed short-circuit counterexamples. As a result, the criterion fails to demarcate these counterexamples from proper transitive chains.

Jonathan Schaffer (2005) maintains that the basic contrastivist method to deal with counterexamples can be extended to short-circuit scenarios; we simply have to be more careful in identifying the ambiguities of the contrast event descriptions. I disagree with that view and so I will present a counterexample for Schaffer's account and shows that his solution leads to unwanted consequences in the analyses of simple transitive scenarios.

In the closing section, I survey possible reactions to the failure of the proposed solutions. As short-circuit scenarios involve negative causation one might think that the puzzle they pose amounts to a good argument against accepting negative causation as real causation. This road is not open to contrastivists committed to negative causation and would involve many other issues in metaphysics; instead I will argue that the best option for the contrastivist is to accept intransitivity and demarcate the counterexamples on grounds of structural features that are different from those defended by Maslen.

3.2 The importance of transitivity for counterfactual theories of causation

This chapter will discuss and criticize treatments of certain counterexamples to the transitivity of causation in the literature on contrastive causation[58]. Problems of transitivity in the wider context of counterfactual theories of causation have been the subject of important discussions in the last decade which initiated new trends including the invention of the contrastivist approach to causation itself. For the early David Lewis transitivity seemed to be an intuitive feature of causation and indispensable to solve the problem of causal pre-emption in the classical counterfactual framework (see Lewis 1973, 1986)[59], but the

---

[58] By the transitivity of causation we mean that for all (x,y,z) if x causes y and y causes z then x causes z.

[59] Pre-emption scenarios show that there is causation without counterfactual dependence. Imagine that there is a marksman shooting at a target, on a different roof there is a second marksman as a backup. The order says that the first marksman has to shoot down the target; in case he misses the second marksman has to do it. The

acceptance of transitivity as a universal feature of causation resulted in problems as there are some everyday causal scenarios that suggest the intransitivity of causation[60] (see Paul 2004, Hall 2004a, McDermott 1995).

Some philosophers proposed that the problems with transitivity are generated by the way causal relata are conceptualized in the Lewisian theory. According to these authors, fine-grained Lewisian events (exemplifications of properties in space-time regions) do not constitute fitting candidate relata for causation. E.g. Paul (2004) suggested that reliance on event aspects can explain most counterexamples away. However, other philosophers suggested the approach that is investigated in this chapter, according to which causation is a four place relation and both cause and effect events are accompanied by tacitly presupposed contrast events. These people think that making the contrasts explicit can help to disentangle problems generated by the counterexamples to transitivity (Maslen 2004). Later contributors, following the same line of thought, suggested that instead of solving only the particular problem of transitivity, one should utilize the contrastivist framework to solve other puzzling issues concerning causation (see Schaffer 2000; Northcott 2008; Reiss 2013a, 2013b).

Parallel to the contrastivist approach, working in a different framework, interventionist theorists, also following the tradition of counterfactual analyses, proposed the more radical idea that counterexamples to transitivity should be accepted as valid and

---

first marksman shoots and kills the target. In this case it is not true that if the first marksman had not shot the target than the target wouldn't have died. Therefore, the simple counterfactual analysis fails for the case.

[60] By the intransitivity of causation we mean that there are cases where it is not true that for all (x,y,z) if x causes y and y causes z then x causes z. By the non-transitivity of causation we would mean that there are no cases where it is true that for (x,y,z) if x causes y and y causes z then x causes z. However, in this paper I will only talk about particular examples in the case of which the causal relation is non-transitive, I won't talk about the non-transitivity of causation as a relation in general.

therefore we should accept that causation is an intransitive relation (Hitchcock 2001). This approach was made plausible in part by the development of new solutions for pre-emption scenarios not relying on transitivity. But giving up on transitivity as a universal feature is not an attractive option. As Maslen (2004) argued, it is possible to accept intransitivity, as long as one provides a good explanation of why transitivity fails in some cases but not in others. Without such an explanation we cannot rely on the notion of a cause to connect causes and effects through causal chains, as happens in certain cases of mechanistic explanations[61] or distant causation by intermediary causes. All of these cases require transitivity to hold or alternatively they require criteria that provide us with the means to differentiate between cases where transitivity can be relied on and cases where it cannot. If there are no such criteria, then causal chains are always in the danger of being broken even without interference from the outside. Advocates of the contrastive theory like Schaffer (2005) aim to explain all counterexamples away in order to preserve transitivity as a general feature of causation. This approach has obvious advantages considering the worries mentioned above.

Even though the contrastivist idea was extended to form a more general theory of causation, transitivity and some everyday counterexamples to it remained in the focus of discussion in the context of counterfactual theories of causation. Two basic approaches to the problem emerged in the literature:

- To save explanations based on transitivity, we have to explain away all the apparent counterexamples by showing that they don't involve proper causal chains.

---

[61] Here I refer to mechanistic explanations in a more traditional sense, not in the sense it is used in the contemporary literature on mechanistic explanations (e.g. Craver 2007). By a mechanism I simply mean a causal chain between a cause and effect event. For example there is a mechanism for delivering a letter to a different country. The first step is the event of sending the letter from the post office and the end effect is the arrival of the letter to the adressee. The processes that take place between the two can be described as a chain of events.

- To save explanations based on transitivity, we should be able to demarcate non-transitive cases and to show why transitivity fails in the problem cases, but not elsewhere.

The purpose of this chapter is to show that when it comes to transitivity the existing contrastivist theories of causation are fraught with problems and that the only viable solution is II., to accept intransitivity and to come up with a meaningful idea capable of explaining why transitivity fails in certain cases, but not in others. So, my stance on how to approach the problem is close to Maslen's, but as I will clarify below, I disagree with the particular solution she defends. However, in passing in brief footnote, she suggested an alternative approach to what she defended at length. In the closing section I will show that this suggestion that nobody took seriously if worked out properly provides us with the only viable demarcation criterion.

Even though in this introduction I tried to situate the problem of transitivity in a wider context, throughout the chapter I restrict myself to the discussion of contrastivist accounts and I only aim to show what is the best possible stance concerning transitivity for the contrastivist, I am not aiming to say anything about solutions worked out in other important frameworks, like the classical Lewisian or the more and more popular interventionist framework. That is a job for a different occasion.

The chapter proceeds as follows. In (section 3.2.1) and (section 3.2.2) I introduce the most important problem cases for the transitivity of causation. (Section 3.3) outlines the basic ideas of the contrastivist theory and shows its problem-solving power when dealing with easy counterexamples. The two sections following that (3.4-3.5) concentrate on two solutions provided for the more difficult counterexamples and show that neither of them is good

enough to provide satisfactory answers. (Section 3.6) considers what alternatives there are for handling the situation and settles on an idea that accepts intransitivity.

3.2.1 An easy counterexample to transitivity

Below I introduce two instructive counterexamples to the transitivity of causation. I will talk about two kinds of examples, so-called switches and short-circuits, because they are widely discussed and treated as the most important kinds in influential papers on the topic (Hall 2004a, 2004b; Schaffer 2005; Northcott 2008). The first one (Ia), was first used to show the flaws of the original Lewisian framework and it is usually treated as an intuitive counterexample to the transitivity of causation in the causation literature. I will categorize it as an easy counterexample as the basic contrastivist framework can handle it elegantly by showing that it is not a valid counterexample to transitivity and because there is no disagreement between contrastivists on how to handle it.

(Ia) **The runaway train**: A runaway train with broken brakes moves towards a station on a certain track. There is a pointer switch on this track that decides if the train travels along towards the station on track T1, or on track T2 . As a default the switch is set on T1. At a later point the two tracks converge into a single track again and after that point the train runs into the station building causing a massive accident (example from Hall 2004b). This is taken to be a counterexample, because among the following three statements we accept the first two as true, but, even if by transitivity it follows from 1. and 2., we reject the last one as false:

- The position of the pointer on T1 (C) causes the train to move on T1 (D).

- The train moving on T1 (D) causes the train to run into the station building (E)

- The position of the pointer on T1 (C) causes the train to run into the station building (E)

To make the case more comprehensible the two possible scenarios are depicted below using the neuron diagram convention. In neuron diagrams circles signify events (a space-time region exemplifying a property), full circles are events that took place, and empty ones are events that didn't (absences). Normal arrows signify the straightforward difference-making capacity of an event with respect to another event, oval arrows signify the prevention of an event. Concentric circles signify stubborn neurons that require as many incoming signals as the number of concentric circles they have.

Figure 3.2-2[62] gives us the above depicted version of the scenario. The train follows track T1 as the pointer is set on that track. On the diagram the pointer is depicted as a binary event the presence of which prevents the train from traveling on T2 and causes it to travel on T1. The meeting point of the train and the pointer is at (D) the event that takes place because of the presence of the train and the position of the pointer are making it happen together. The pointer also prevents the train from following track T2. As a result, the train runs into the building via track T1 (E). The alternative situation can be seen on Figure 3.2-1. There the train follows track T2 as the pointer is set that way. The presence of the train and the position of the pointer allow the train to follow track T2. At the end the train runs into the building.
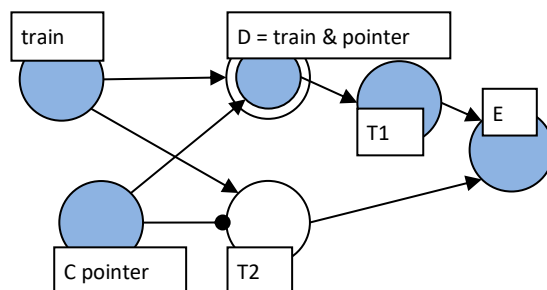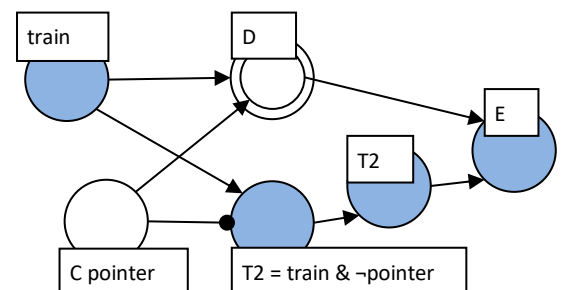

Figure 3.2-1

Figure 3.2-2

---

[62] On figures used in this and the following sections the first version of the causal scenario depicts the state of the system in question with the relevant first cause event in place, whereas the second version depicts the state of the system with the relevant cause event being absent.

3.2.2   A hard counterexample to transitivity

There is a widely quoted category of counterexamples which have been dubbed 'short-circuits' by Ned Hall (2004a). Cases of this kind form a more difficult set of counterexamples to transitivity, so I will call them hard counterexamples. As we will see authors who tackled the issue, either inside or outside the contrastivist framework, struggled with it and basically every one of them suggested a somewhat different treatment.

(Ib) **The assassin trainee**[63]: A master assassin escorts his trainee to his first real mission. Only the trainee has a rifle, but, he can shoot only if he is ordered to do so by his master. The action takes place on a marketplace. Suddenly, the master realizes that a group of agents are approaching to rescue the victim. He yells the order 'shoot', and the trainee does so, but the victim also hears the order and ducks avoiding the bullet and surviving. Again, among the following three statements we accept the first two as true, but, even if by transitivity it follows from 1. and 2., we reject the last one as false:

- The yelled order (C) caused the victim to duck (D).

- The ducking of the victim (D) caused the survival of the victim (E).

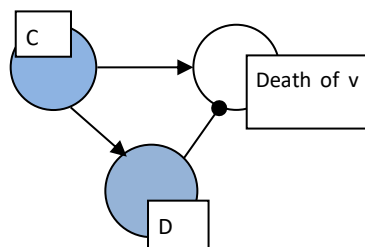- The yelled order (C) caused the survival of the victim (E).

Figure 3.2-3   **Figure 3.2-4**

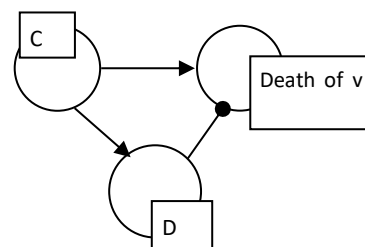[63] (Ib) appears in a paper from Northcott (2008), I have chosen this example, because I find it more apt to highlight important features of short-circuit scenarios than other well-known examples of short-circuits like the case of the falling boulder (see: Schaffer 2005).

3.3 The basic contrastive account of causation

Below I will show how the framework is applied to solve the easy counterexamples and, in the sections, following that I will discuss and criticize solutions to hard counterexamples. Contrastivist solutions to the problems of the classic counterfactual theory presuppose that the surface form of everyday causal statements is misleading. Even though on the surface causation seems to be a two-place relation, causal statements tacitly express a four-place relation. Causes and effects have contrast events attached to them. The basic scheme is as follows:

C rather than C1* causes E rather than E1*

Contrastivists maintain that by making contrasts explicit, in other words by making the worldly context of the causal statement explicit, many problems of the old counterfactual theory can be solved, the problems of transitivity, selection[64], negative causation, the graining of events and so on (see Schaffer 2005). The origin of these problems is identified in ambiguities that vanish if the proper contrast events, the context relevant for the statement is taken into consideration. It is easy to make any causal statement true or false by arbitrarily choosing contrasts for the surface statements, but it is also straightforward that in everyday practice we rely on a tacit system of rules to decide on the admissibility of contrasts relative to a particular context. Recently Reiss (2013a, 2013b) developed interesting ideas to account for this practice. Even though it is a received view in the relevant literature that the choice of contrasts is determined by the context a satisfactory account of the mechanism by which context determines the meaning of causal statements, the origin of the contrast classes

---

[64] Selection is the practice of choosing one cause as the cause among the many causal factors contributing to an effect.

remained elusive (see Schaffer 2013). This challenge, however, is not particularly relevant to the subject of this chapter and will not be considered here. Below, I presuppose that the basic contrastivist idea is sound enough and constrain the discussion to the solutions provided for the problem of transitivity.

In all of the proposed contrastive theories it is accepted that to test the truth of a causal statement one should test the truth of the counterfactual involving the contrasts events associated with the explicit cause and effect event (see: Maslen 2004; Schaffer 2005). It is the truth value of the counterfactual C1* □-› E1* relative to the actual world where both C and E occurs that decides the truth value of the causal statement 'C causes E' instead of C □-› E suggested by the surface form. These truth conditions were introduced to make the truth evaluation of causal statements relative to a specific context; the contextual features relevant for the truth evaluation are expressed by the contrasts.

The transitivity of causation should also be reformulated for this quaternary approach. It would mean that for all (x, x*,y, y*,z, z*) if x rather than x* causes y rather than y* and y rather than y* causes z rather than z* then x rather than x* causes z rather than z*. To demonstrate the power of the contrastive idea, it is useful to consider how the contrastive method works in the case of the runaway train (Ia). As we saw above, the contrastivist proposes that for a causal chain (meaning at least a chain of two causal links) to be a real causal chain it has to be continuous or linked up at the middle event (that is, the effect event in the first link and the cause event in the second link) not only on the explicit side but also on the implicit, contrast side. This is called 'differential transitivity' by Schaffer (2005:308). The contrasts can't simply be negations of the explicit side statements; the context provided by the particular causal scenario helps us and requires us to make the contrasts more precise

than that, but the contrast should be an event that is compatible with the negation of the explicit side event. Let us first analyse (Ia) using negations as default contrasts (Ia1*):

- The position of the pointer on T1 (C) rather than not on T1 (C*) causes the train to move towards the station on line T1 (D) rather than *to move not on T1* (D1*).

- The train moving towards the station on line T1 (D) rather than *moving not on T1* (D2*) causes the train to run into the station building (E) rather than not to run into the station building (E*).

- The position of the pointer on T1 (C) rather than not on T1 (C*) causes the train to run into the station building (E) rather than not to run into the station building (E*).

On grounds of the above formulation one can say that the last causal statement is clearly false, but both 1. and 2. seem fine and if this is all right then the counterexample holds. Fortunately, taking a closer look at the contrasts, one can show that 2. is false as well. Because of the apparent identity of the descriptions in D1* and D2*, we might come to believe that the scenario is a case of transitivity. But '*to move not on T1*' is equivocal. In the case of D1* the context of the scenario suggests that 'not on T1' has to be 'to move on line T2' but if one gives D2* the same interpretation it makes statement 2. false. To make 2. true 'not on T1' should refer to something like a derailment, where the train is not moving neither following track T1 nor following T2. But according to what we know about the situation no derailing is involved in the context of the scenario and furthermore if D1* would refer to a different event than D2* then the two statements would belong to different contexts, so we would have no reason to think that they belong to a single chain of events. Here we should remember that a valid transitive causal chain requires two properly connected, true causal statements. By disambiguating the middle contrast descriptions we might get two true causal statements, but in that case there is no continuity at the middle contrast. It is also clear that if D1* and

D2* are identical then we have no real causal chain either, because when they are (D1*=D2*='to move on line T2', or D1*=D2*='not to move on either line, derail') one of the two causal statements becomes false. The conclusion from all this is that in scenario (Ia1*) it would be wrong to accept 1. and 2. together as a proper causal chain as either 2. is not a valid causal statement or 1. and 2. aren't linked up, so it is no surprise that statement 3. is false. If there is no real causal chain involved in a scenario then the question of transitivity is irrelevant, so the counterexample is explained away.

Maslen and Schaffer agree that either a causal chain is properly linked up at the middle contrast and consists of valid causal statements or the presumed causal chain is not a causal chain at all. To have a proper causal chain the following statements should all be true about the contrasts: C*□-›D1*, D2*□-›E*, D1*=D2*. To sum up, the basic contrastivist understanding of the counterexamples is this: in cases of perceived non-transitivity we are misled by the ambiguities of the surface form of everyday causal statements. Intransitivity is an illusion created by sloppy descriptions.

We should now turn to the hard cases. People who tried to systematize the contrastive view recognized that these cases are different, but they disagreed on the required solution. Below I discuss the two known suggestions to argue that they require supplementation.

3.4   Maslen's account of the hard cases

According to Maslen, general criteria can be given for demarcating the problem cases, but causation is intransitive. The reason why we should have good general criteria differentiating transitive and non-transitive cases is that the validity of explanations by causal intermediaries depends on this. The lack of criteria to differentiate transitive and non-transitive cases would endanger the validity of these widely used causal explanations.

After she develops the basic contrastivist idea and applies it to the easy cases she turns to the hard cases and observes that the criterion already developed for easy cases can't help us with these (Maslen 2004: 351-354). Let us take a look at a contrastive reformulation of (Ib), I will call it (Ib1*):

- Yelling the order (C) rather than giving no order at all (C1*) causes the victim to duck (D) rather than not to duck (*to stay upright*) (D1*).

- The ducking of the victim (D) rather than not ducking (*his staying upright*) (D2*) causes the survival of the victim (E) rather than the death of the victim (E1*).

- Yelling the order (C) rather than giving no order at all (C1*) causes the survival of the victim (E) rather than the death of the victim (E1*).

The conclusion that follows from 1. and 2. by transitivity is obviously false, but there seems to be no way to interpret the contrasts of D as not being identical as it was possible in the case of the easy example. D1*=D2*=‘*staying upright*’ is a perfectly valid interpretation of the meaning of the default contrast ‘not to duck’ in the context provided. 1. and 2. are true independently of each other, so it seems that the first criterion is insufficient for demarcation.

To amend the situation Maslen suggests a further criterion. She claims that there is an interesting trait transitive causal chains have, but short-circuit scenarios seem to lack. It should be true about scenarios involving transitive chains that if the second contrast event had occurred then the first contrast event would have occurred as well (see: Maslen 2004:352). This amounts to a backtracking counterfactual criterion: $D^* \square \rightarrow C1^*$. Her claim is that if this criterion is violated, we have a case of non-transitive causation at hand.

Discussing this second criterion I will put aside methodological questions concerning the evaluation and interpretation of backwards counterfactuals[65]. I will concentrate on the problems of its application to the case in question. To check the soundness of the criterion let us apply it to a straightforwardly transitive example first, the case of the provoked bull (Ic1*):

- The bull being provoked by a stranger (C) rather than *having a peaceful environment* (C*) caused the bull to become angry (D) rather than *to remain tranquil* (D*)

- The angriness of the bull (D) rather than its being tranquil (D*) caused the bull to attack its owner (E) rather than to behave as usual (E*).

- The bull's being provoked (C) rather than having a peaceful environment (C*) caused the bull to attack its owner (E) rather than to behave as usual (E*).

D*□→C* seems to be true here. If the bull had been tranquil the bull would have had a tranquil environment earlier. So, it is somewhat plausible to think that the criterion identifies transitive cases in accordance with intuition.

Maslen maintains that her second criterion is violated in the cases of hard counterexamples. At first sight it seems to be true in (Ib1*) that 'for the victim to stay upright, the master assassin would have not to have given a command'. However, it is not simply the command that alarms the victim, it is the sound produced by the shouting. Concentrating on

---

[65] There is a general worry related to the use of backtracking counterfactuals. Counterfactuals are implemented by a miracle: introducing the antecedent by a miracle that happens later than the consequent one would break down the directed causal relationship between the antecedent and the consequent. So, it is not easy to see how to implement a backwards counterfactual properly. We should also mention that David Lewis himself was against the use of backtracking counterfactuals in all his versions of the counterfactual theory of causation, although to that worry one can reply that in this case it is not used to evaluate causal status, but to decide whether causal chaining is justified or not, which is a distinct issue. But the first worry concerning the implementation of such counterfactuals holds in any case.

this fact it becomes visible that there are other available contrast events. Let's call the following reformulation (Ib2*):

- Giving a yelled command (C) rather than *giving a silent command* (C2*) caused the victim to duck (D) rather than *to stay upright* (D*).

- The ducking of the victim (D) rather than his staying upright (D*) caused the survival of the victim (E) rather than his death (E*).

- Giving a yelled command (C) rather than giving a silent command (C2*) caused the survival of the victim (E) rather than his death (E*).

Maslen thinks that the availability of contrasts like (C2*) means that D*□→C1* is not true in (Ib1*) (see: Maslen 2004:352). She generalizes this observation to demarcate scenarios of the same type as failures of transitivity and suggests that there is an analogy with the intransitivity of counterfactuals as it was conceptualized by Lewis (1986). I won't discuss the question of the transitivity of counterfactuals here[66]; I only aim to show that the criterion suggested is insufficient for demarcation.

Let's unpack what the criterion has to say. It seems to be true that starting from a world with an upright victim a world with a silent command is closer than a world where there is no command given at all. In other words, it doesn't seem to be true that in the closest possible world with an upright victim there is no command either, because there seems to be a world that requires even less departure from actuality which has a silent command instead. If this were true, then according to Maslen the second criterion would demarcate (Ib1*) as a case of non-transitive causation (see Maslen 2004:352–353).

---

[66] In the literature on counterfactuals the issue of transitivity was thoroughly discussed quite early on and people like Lowe (1990) argued convincingly that all apparent counterexamples involve contextual equivocation, if he was right then the whole analogy between the intransitivity of causation and counterfactuals breaks down.

As far as I know, nobody has tried to criticize this solution directly, not even Jonathan Schaffer who proposed an alternative account which will be investigated in the following section. Schaffer has even said that in case his account fails Maslen's idea might provide us with a fall-back option, although it is less systematic in dealing with transitivity problems than the new framework he has developed (see Schaffer 2005:311, footnote 24). In what follows, I will show that Maslen's solution is not only less systematic, it has internal problems. To do so, I will point to the fact that in the context where 'giving no command' is the justified, admissible contrast 'giving a silent command' is simply not available and therefore the second criterion fails to deliver on its promise.

To spell out this criticism of mine I will draw on resources provided by Northcott (2008) and Reiss (2013a, 2013b) who developed the idea that the contrastivist framework is in need of criteria to decide the admissibility of contrast events. The motivation for this was the realization that contrasts were introduced to disambiguate causal statements. Unfortunately, if there is no principled way of fixing contrasts for particular causal statements than in many cases the method of contrasting does more harm than good as it provides means to change the truth values of causal statements arbitrarily (see Reiss 2013a:1067). The contrastive formulation is only helpful if there is a way to decide which contrast is acceptable for a particular causal statement. In my interpretation, attendance to admissibility criteria is attendance to how a causal statement should be interpreted and evaluated relative to its particular context. According to Reiss, the admissibility of contrasts in cases like (Ib) is determined by what we think to be a possible course of action under the circumstances (see Reiss (2013b:1079–82, 1088). The circumstances we are talking about are nothing else but the facts of the worldly context relevant to the causal statements in question.

Let us suppose that in the actual case the actions described in scenario (Ib) take place in a marketplace that is noisy, hence the need for a shouted command. Under the circumstances the master assassin has only two real choices, either to give a loud command and risk that by doing so he warns the victim as well, or to stay silent and wait for another opportunity. Non-verbal and other ways of signalling are excluded because of the positions the assassins took. The main thing is that we know that a silent command wouldn't be effective under the circumstances. In other words, the context of our causal statement in (Ib) is such that the contrast in (Ib2*), 'giving a silent command', goes against the logic of the situation, no master assassin would try that[67]. There might be a faraway world where the master assassin gives a silent command, but in that world many other things are different as well: the marketplace is peaceful, the assassins can easily hear each other, etc. That would be a different world that is far away from the actual where 'giving no command' is practically the only alternative for the master assassin.

One could also turn the example around adding contextual features to (Ib) where the marketplace is silent and it is feasible to think that the master assassin would give a silent command. This would select (Ib2*) as the right contrastive interpretation in which case (Ib) is not a counterexample to transitivity and the contrast in (Ib1*) 'giving no command at all' takes place in a faraway world.

Knowing the above, we can see that where the first contrastive formulation of the assassin trainee scenario (Ib1*) is justified there the backwards counterfactual criterion

---

[67] It is interesting to highlight here, that examples with a worldly contextual structure isomorphic to (Ib1*), where the only plausible alternative to a yelled command is no command given at all, are creating problems for transitivity, whereas (Ib2*) is a proper transitive case. So, in our example (Ib) the context of the surface level statements makes a huge difference with respect to the content of the causal statements involved.

required by Maslen is made true by the same contextual features that justify the choice of the first contrast event. If the admission of the contrast event 'giving no command' is justified in the actual world and the exclusion of 'giving silent command' is grounded in the same facts, then there is no possible way the criterion D*□->C1* could be false and demarcate the assassin trainee scenario as a case of non-transitive causation. The reader can easily verify that the same logic applies to some other widely discussed hard counterexamples to transitivity. Form this it follows that Maslen's second criterion cannot serve either as a necessary or as a sufficient condition for identifying hard scenarios as cases of non-transitivity.

3.5 Schaffer's account of the hard cases

Below, I reconstruct Schaffer's account and I provide arguments to show that it is not a viable alternative to Maslen's failed solution. When discussing the easy counterexamples to transitivity Schaffer (2005) follows the same general lines I described in (section 2.). According to him, even when contrasts are made explicit it is possible to formulate equivocal, in his words 'shifty', event nominals for the middle contrasts that can create the illusion of continuity. The logic should be clear from our easy example of the train from section 2., where using 'to move not on T1' as a contrast event description gave way to equivocation between 'to move on T2' and 'getting derailed'. But Schaffer pushes this idea further, he argues that the source of confusion is the same concerning both the easy and hard examples, implicit ambiguity in the middle contrast descriptions, and he attempts to show that the identity of middle contrasts in a case like the assassin trainee (Ib) is as illusory as in the case of the runaway train (Ia).

This means that it should be true of hard cases like (Ib1*) that D1* in the first link and D2* in the second have different referents against all appearances. Schaffer's analysis of the case would go as follows: 'not to duck' in the first causal statement of the scenario means that 'the victim is upright *when there is no command given*', whereas to make true the second statement of (Ib1*) D2* should mean something different like 'the victim is upright *when there is a yelled command given*', so D1* ≠ D2*. The contrasts are different, because in the first counterfactual link the contrast event on the effect side would take place in a world where there is no command given, whereas in the second link the contrast event on the cause side would take place in a world where the command is given already, because D2* □-› E* is true only in a world where the life of the victim is endangered by the shot, that is, where C occurs. If this is sound, we have a hidden contextual ambiguity in the middle contrasts and an explanation of why we don't have a real causal chain in (Ib1*).

My contention is that this solution has unwanted consequences; it misidentifies some normal causal chains as counterexamples to transitivity. To see this one should first scrutinize Schaffer's definition of events on grounds of which he justifies his solution. He, as all contrastivists, construes causation as a four-place relation where all the relata are events, to quote him "concrete, coarse-grained, worldbound occurrences" (Schaffer 2005:298). He explicitly commits himself to a view on the metaphysics of events that is, as he says "Davidsonian in spirit". This seems to imply that the space-time region occupied by an event is among its identity conditions, but it is only a necessary not a sufficient condition. The other necessary condition we should consider is worldboundedness. Schaffer acknowledges some minor differences between the Lewisian construal and his view, the most important being that while Davidson is noncommittal with respect to the interworld relations of events, the contrastivist view takes events to be Lewisian individuals with counterpart theoretical

profiles. This means that events occurring in different possible worlds cannot be the same event and therefore events in the middle contrast positions are the same only if they take place in the same possible world[68]. If they don't then even if they occupy the same space-time region as their counterparts in other close worlds they should be considered as non-identical events. I am not sure whether these criteria together should be considered sufficient for event identity but for our present investigation it is enough to know that we are provided with these two independent necessary conditions.

Let's check whether these conditions can deliver the right results for our hard transitivity counterexamples. If events were individuated only by the space-time regions they occupy then these conditional clauses ('*when there is no command given*', '*when a yelled command is given*') couldn't play any role in differentiating the contrast events, because the descriptions in the clauses refer to happenings in a space-time region that is distinct and distant from the core event. The yelling of the command and the absence of the command are events that cannot be part of the region that identifies 'the victim is upright' as a particular event, because the command is given (or not) outside and far away from the region that identifies it. Therefore, the space-time region criterion forces us to think that the two descriptions with different conditional clauses refer to the very same event. If the denoted event is the same event in both the first sentence of (Ib1*) and the second sentence of (Ib1*) and it makes both causal statements true, then (Ib1*) should be accepted as a proper causal chain.

---

[68]Trying to save the counterfactual theory other philosophers also subscribe to a counterpart theoretical approach to events. Their idea is to build the context sensitivity of causal statements into the counterpart relation (e.g. McDonnell 2016).

However, on grounds of the worldboundedness criterion, it is plausible to interpret the conditional clauses in the middle contrast descriptions as expressing information about the world belonging of the events referred to, as expressing the fact that the two events are taking place in different possible worlds. Another way of thinking about this is to say that as the middle contrast events occur in different worlds, they inherit some traces, at least some slight influences form those different worlds and the clauses inform us about the world belonging of the middle contrast events by highlighting exactly these traces. When the victim stays upright in a world without a yelled command the victim has no soundwaves in his ears originating from the master assassin, whereas when he stays upright in a world with the command yelled, he has traces of that sound in his ears. The application of this second criterion suggests that 'the victim stays upright *when there is no command given*' and 'the victim stays upright *when a yelled command is given*' refer to different events and they do so because the corresponding events occur in different possible worlds. If one goes with this interpretation for the clauses, then Schaffer's solution makes sense.

Now that Schaffer's solution for hard cases is made plausible, we should check whether it really provides us with a precise enough tool, one that identifies apparent causal chains in the problems cases as cases of causal disconnectedness but preserves our intuitions concerning ordinary transitive cases. As I will show, it is not precise enough, because the difference in terms of world belonging between middle contrast events, the supposed illicit shift identified by Schaffer in short-circuits appears even in some straightforwardly transitive causal chains.

In scenario (Ic) we had a bull provoked by a stranger C that made it angry D which in turn made it to attack its owner E. As we saw, this is a perfect causal chain without any known problem, where C causes, provides proper control over E. As the solution suggested by
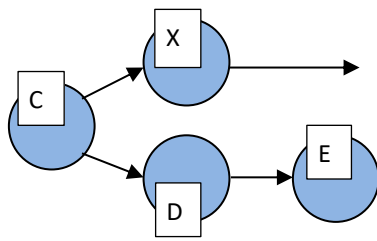
Figure 3.5-1

Schaffer aims to demarcate these normal cases from the non-causal, pseudo non-transitive cases, it should only apply to the short-circuit cases. But consider the following; let's add a few details to scenario (Ic), imagine that C has an effect other than D, X, where X has no effect on the occurrence of D or E. Let's call this version of the scenario (Ic1*). Its structure can be seen on Figure 3.5-1.

Adding such a detail is quite natural as most causes have multiple effects and this is all we need here, we need an alternative effect originating from C that leaves a trace on D. I would also risk the claim that such things occur quite frequently. Let's say, event X results from C in the following way. By making his provoking moves the stranger also disturbs a bird that flies over the bull casting a shadow over it for a second. Now if we accept Schaffer's criteria for event identity, this means that in the case of the first counterfactual the middle contrast D1* refers to the event where 'the bull is tranquil *when there is no shadow cast on it by a bird'*, but in the case of the second counterfactual D2* refers to 'the bull is tranquil *when there is a shadow cast on it by a bird*', as in this world C already took place causing X to happen, casting a shadow on the back of the bull.

You can see on Figure 3.5-1 that the arrow originating from X avoids E, making it clear that this is not a short-circuit scenario. According to intuition there should be a straightforward transitive causal chain between C and E. However, in case we accept Schaffer's criteria for event identity, we get a difference in the middle contrasts (D1* ≠ D2*) that amounts to an illicit shift, therefore, the chain between C and E is not a proper one.

To make things more transparent let us analyse scenario (Ic1*) in terms of the tests we should employ to test the truth of the two causal statements involved according to the contrastivist. First, we want to check whether C 'the stranger is provoking the bull' causes D

'the bull is getting angry with *a shadow cast on it'* is true. To do that, we should start from the actual world where C and D is in place, then fixing the past first, we move to the closest possible world where instead of C, we have contrast event C* 'the bull is left alone, the stranger is just passing by' and we check whether D1* 'the bull is tranquil *when there is no shadow cast on it by a bird'* is in place as well. To do that, we have to fix the past and change C to C*. The reason why in the first link of the chain D1* is 'the bull is tranquil *when there is no shadow cast on it by a bird'* can be understood from how the scenario is constructed. When C is in place the stranger disturbs the bird as well, not only the bull. So, C, as in the case of short-circuit scenarios, is a common cause of two events, D and X. But as C* does not excite the bird (X), in the counterfactual world where we should move to test the first causal link the shadow is not cast on the back of the tranquil bull either. Now, changing C to C* changes D to D1*, so the test gives the result that C causes D is true.

As a second step we should also check whether D 'the bull is getting angry with *a shadow cast on it'* causes E 'the bull is attacking its owner' is true. To do that, we should start from the actual world where D and E are in place, then fixing the past first, we move to the closest world where instead of D, we have contrast event D2* 'the bull is tranquil with *a shadow cast on it'* and we check whether E* 'the bull continues to sniff the flowers' is in place as well. The reason why D2* is 'the bull is tranquil with *a shadow cast on it'* can be understood from the details of the scenario. We know that the past of D contains C, the stranger had already disturbed the bird as well, not only the bull as C was a common cause of D and X. As a result, we change D to D2* in a context where event X remains in place. Now changing D to D2* changes E to E* so D causes E should be true.

Intuition tells us that C should also be the cause of E as it gives us perfect control over E. But Schaffer's differential transitivity does not apply here, because on grounds of his criteria

D1* is a different event from D2* as it occurs in a slightly different world from the world where D2* does. So, my scenario seems to be a good counterexample to Schaffer's solution for short-circuit scenarios. In other words, the application of Schaffer's solution for short-circuits to a normal transitive scenario resulted in the unwanted consequence that we should analyse it as a deviant case. I take this to be a good enough reductio against the solution suggested by Schaffer.

To sum up, Schaffer relying on his special view concerning event identity tried to extend the contrastive solution for easy counterexamples to hard counterexamples. The idea was that in the case of both kinds of counterexamples non-transitivity is an illusion generated by the ambiguous nature of implicit contrast event nominals and by disambiguating them we can uncover the hidden crack in the causal chain. I have shown that in short-circuit cases this strategy fails, because the equivocation uncovered by Schaffer is such that it can be found in supposedly normal transitive cases as well.

3.6 How to handle the unexplained counterexamples?

If there is no better solution from the failure of the proposed solutions it follows that if we would like to maintain the contrastivist theory, we should accept the intransitivity of causation[69]. The problem with this stance is how to answer the questions put forward by Maslen. Even though we can't maintain the view that intransitivity is only an illusion created by sloppy descriptions, we still have to demarcate the non-transitive cases by some reliable

---

[69] A similar idea is promoted by Hitchcock (2001) in the interventionist theory of causation which is quite close to the contrastivist. According to Woodward (2005) the interventionist theory is a special case of the contrastivist theory. The main difference between their view on transitivity and Maslen's is that these authors accept all kind of counterexamples to transitivity as valid, not only the short-circuit cases.

means. This is the only way to preserve the explanatory power derived from transitivity where it is required.

Some people working on different theories of causation suggest that the only reason we are still struggling with the hard counterexamples is that we mistakenly take negative causation to be real causation (e.g.: Moore 2009). Short-circuits involve preventions of other events in the second link and prevention is a case of negative causation, but if negative causal talk is not ontologically serious then these counterexamples are only relevant for causal explanation, not for causation itself. Most philosophers working on process theories or power theories of causation agree with this image. (Dowe 2001, 2004 Mumford and Anjum 2011). Some argued for the same idea even in the context of counterfactual theories (Beebee 2004) This solution has its advantages, but most supporters of the contrastivist framework and many believers of counterfactual accounts in general are explicitly committed to ontologically serious negative causation for good enough reasons (see Schaffer 2004, 2005; Lewis 2004; Menzies 2006), therefore pursuing this line of thought requires a thorough discussion of the issue which exceeds the limits of this discussion. I will only consider answers that are directly available for the contrastivist.

As there are no known hard problem cases other than the short-circuit scenarios the demarcation might be done quite trivially. One might say that whenever we have a proper short-circuit scenario, we have a case of non-transitivity at hand, and there are no other cases. The question is how to identify short-circuits properly? The solution I will argue for originates from a sentence-long footnote in Maslen's paper. She mentioned the idea as a possible alternative to her backwards counterfactual criterion. I believe it is the only working solution. The statement below should be true in cases of proper transitive chains and it should be false

when and only when it comes to short-circuit type scenarios (adapted from Maslen 2004:357, footnote 35):

(C* & D*) □→ E*

Think about a cascade of dominos as a straightforward case of a transitive causal chain. It is true that if neither the first nor the second domino had fallen then the third domino wouldn't have fallen either. This works in the case of the angry bull (Ib) as well. If the stranger had walked along and the bull had continued grazing than the bull wouldn't have attacked its owner. However, the criterion is false for short-circuit scenarios like the assassin trainee. It is false to say that if the master assassin had stayed silent and the victim had stayed upright then the victim would have died. This sounds promising. It is even surprising that Maslen chose to work out the more problematic backwards counterfactual criterion in her paper. This suggestion seems to work and it is devoid of the theoretical difficulties of her main proposal. The only problem one can raise is that the criterion is unmotivated; we get no explanation of why it is capable of doing the job of demarcation.

To give an interesting explanation one should reformulate what the criterion has to say. In short-circuit scenarios the occurrence of the first event of the causal chain in focus (C-D-E) and the alternative chain of events the first event initiates (e.g. the trainee pulling the trigger in (Ib1*)) is a precondition for the truth of the second causal statement, D causes E., in the (C-D-E) chain. Or to put it differently C serves a dual role here, by transitivity it is the cause of E and it is also among the background conditions of the causal relation between D and E[70]. The counterfactual if C hadn't occurred then D rather than D* wouldn't have caused

---

[70] The occurrence of effects always depends on a number of causal conditions even though in everyday practice we select one as the cause of the effect. Most counterfactual accounts are indifferent with respect to causal selection they don't differentiate between causes and background conditions. From the perspective of those

E rather than E* holds true for these cases. In (Ib1*) if there is no loud command given then there is no shooting, but if there is no shooting then 'ducking' (D) rather than 'being upright' (D*) does not make a difference to the occurrence of the survival of the victim (E).

In normal transitive scenarios like domino cascades or in cases like (Ic1*) this is not true. The bull getting angry (D) rather than remaining calm (D*) does make a difference to whether it attacks its owner or not no matter the stranger was there to excite it (C) or not (C*). For the fall of the third domino in a queue of dominos the presence of Earth's gravity is a background condition and we would probably say it is caused by the fall of the second domino. The fall of the third domino counterfactually depends on both of these causal conditions and there are other events on which it does not depend, e.g. the lighting conditions of the space where the third domino falls. Interestingly, it is true about the fall of first domino as well that it is not among the background conditions of the fall of the third domino. The fall of the second domino can cause the fall of third in the absence of the fall of the first domino. At the same time, one could argue that the first domino can cause the third to fall via causing the second to fall. But even if it does, the capacity of the second domino to cause the fall of the third is not dependent on the fall of the first domino. As we saw, this is not true of short-circuit cases and that is what leads to the existence of non-transitive causal chains.

Unfortunately, I can't prove that the identified structural feature only appears in short-circuits, but on grounds of our limited induction base, what we know about examples of transitivity and counterexamples to it, the criterion that the occurrence of an earlier event

---

frameworks any event counts as a causal condition of an effect if the occurrence of the effect depends on that event counterfactually and, naturally, not all past events are among the causal conditions of an outcome. Here, I rely on the everyday distinction between the cause and other causal conditions and will call the latter background conditions because the distinction is helpful in highlighting the dual role the first event plays in short-circuit scenarios.

in a causal chain is a precondition for the truth of a later causal statement in the same chain seems to do the job of demarcation, it identifies short-circuits and only them. At present I don't see any good reason to think that the above structural characterisation is not good enough to identify all problem cases relevant for transitivity, so I suggest it as a temporary solution for the problem of demarcation.

However, there is an important issue one should consider, short-circuits can come in many shapes and colours they might involve more links, more causal lines cancelling each-other out or they might have other less predictable variations. If they share those basic structural features I have highlighted above, we might be able to identify all of them. Proving this properly is a big enough project in itself and I won't pursue to do that here. And there is even more to this problem, it is well known from the literature on actual causation that the sheer number of possible causal structures is beyond human control (see: Glymour et al. 2010) and if it is, we simply can't be sure that there are no further counterexamples beyond those we know about at present. So, the tentative solution I suggested is only a solution until we encounter some further tricky counterexamples, but at least till that time it could be a useful suggestion.

3.7   Concluding remarks

This chapter has shown that contrastivists cannot demarcate or explain away short-circuit type counterexamples to the transitivity of causation. Maslen was unsuccessful in formulating a working demarcation condition for non-transitive cases. Schaffer tried to extend the solution developed for easy counterexamples to cover the hard cases and explain them away, but it turned out that if one treats the middle contrast events of short-circuits only as close counterparts in different worlds to resolve the issue unacceptable consequences follow. The counterpart events story led us to the problem that some straightforward causal chains do not conform to the suggested criterion. From all this I drew the conclusion that there are no good enough methods for the contrastivist to explain away the hard counterexamples, but I have suggested that it is possible to demarcate them based on certain structural features and thereby to save those theoretically important explanations that rely on transitivity.

So, it seems that counterfactual theories of causation have their difficulties as well, even if they perform better on a lot of tasks than productive theories of causation, but the research program built around the basic counterfactual notion of causation still seems to be a progressive one.

Chapter 4: Menzies' reformulated exclusion argument

4.1   Turning the tables on Kim: reformulating the exclusion argument

As we have seen in chapter 2, it is plausible enough that Kim was wrong about the role different notions of causation can play in the exclusion problem. First, there are good reasons to rely on a difference-making theory of causation based on some version of the counterfactual analysis when thinking about exclusion and mental causation. In spite of Kim's claim to the opposite, in many respects production is a weaker notion of causation then counterfactual dependence and difference-making (see section 2.5). Second, it is not straightforward that Kim's exclusion arguments are neutral with respect to the theory of causation presupposed. As we will see in this chapter they are not.

Moreover, as we saw in section 2.3, reliance on a transference theory of causation makes the original exclusion argument empty as it makes the exclusion principle redundant in the argument. In that construal the closure principle says that "physical effects only have physical causes". This is what Kim (2005:51) himself called strong closure and advised against exactly because it makes the exclusion argument empty. If causation is fundamental level physical energy transference, then there is no use in running an exclusion argument as the theory of causation employed renders the starting supposition of the argument that mental properties can cause physical properties meaningless. We also saw that reliance on such a restrictive notion of causation creates problems for the empirical backing to the causal closure principle. The empirical support people mustered for the principle is based on the conservation of energy principle and on further support for the belief that energy in physical systems can only be redistributed by transference of energy from the outside. However, if

causation is reduced to the transference of physical energy these premises don't do real work anymore. Strong closure short-circuits the whole exclusion argument project.

So, there are good reasons for changing the game and turning the table on Kim and his version of the exclusion argument. To do that, one can accept most premises of Kim's exclusion argument with the exception of the hidden premise presupposed by some of them concerning the interpretation of causation. For the later Kim causation is some kind of production, in his later writings some form of physical transference of a conserved physical quantity between causes and effects, but as it was shown in section 2.5, there are serious reasons to prefer a counterfactual theory over physical theories. So, the reformulated version of the argument is run based on a counterfactual theory of causation. In effect this means that one has to reformulate the exclusion argument based on some form of difference-making causation. The relevant changes are italicized for easier visibility:

I.   **Supervenience**: All mental properties supervene on basic physical properties.

II. **Causal closure***: Every physical event that has a cause at time t, has a *difference-making* physical cause at time t.

III. **No overdetermination***: Mental *difference-making* causes do not overdetermine their effects.

IV. **The Exclusion Principle***: If an event e has a *difference-making cause* c then no event distinct from c is a *difference-making cause* of e.

V.   **Mental causation**: mental properties are real and causally efficacious.

VI. **Distinctness**: mental properties are numerically distinct from physical properties.

Menzies and some others argue that causation is proportionate counterfactual dependence and if mental properties are multiply realized mental properties might be the

proper difference-makers of certain physical outcomes. In certain cases, the realized property excludes the realizer property from causal efficacy, this is what he calls downwards exclusion. In other cases, the physical realizer is the difference-maker cause, that is called upwards exclusion. The result is that the direction of exclusion, as we will see later, is claimed to depend on the empirical details of the particular scenario.

Interestingly enough the Causal closure principle (at least a version of it) is enlisted and accepted by Menzies. This is an obvious requirement for someone who claims to be a non-reductive physicalist. But from the point of view of the reformulated version of the exclusion argument it is redundant as the new exclusion principle is derived directly from the difference-making interpretation of causation. So, the only function it can serve is to highlight basic physicalist commitments. No overdetermination is also accepted, but nothing really depends on this premise either[71] as in the reformulated argument the proportionate difference-making notion of causation bears the weight of the argument alone. As Phil Dowe says about Yablo's different, but closely related defence of higher-level causation: „proportionality is itself an exclusion principle: only one in a set of properties related by the determinate/determinable relation can be proportional to a specific effect." (Dowe 2010:446) So, in certain respects, Menzies's new exclusion argument is a huge shift from Kim's.

However, the reformulated exclusion argument accepts the basic spirit of the old argument. We are still working on the assumption of causal exclusion; however, the principle is based on an alternative theory of causation. On this interpretation there are two competing candidate causes and the possibility of overdetermination (two competing causes of equal

---

[71] Menzies only alludes to this issue in footnotes (see: Menzies and List 2009, footnote 4; Menzies 2013, footnote 2), and he does claim to think that the mental does not work as an overdeterminer cause in the everyday sense, so overdetermination is not an issue.

standing) is avoided on grounds of the assumption that only one of the competing causes can be a proper, proportionate difference-making cause for the outcome in question. These are prima facie exclusive options. Therefore, we can call both the old and the reformulated exclusion arguments incompatibilist approaches to the exclusion problem[72]. Kim's approach to exclusion is also a version of incompatibilism, although in his conceptualisation the lower-level always wins out.

To sum up, for Menzies, causal exclusion can happen both upwards[73] and downwards where the higher-level cause, the mental property wins out, which was outright impossible in the context of Kim's version of the exclusion argument. This means that, contrary to Kim's version, the exclusion of higher-level causes is far from an analytic truth, it is a contingent matter.

In their joint papers, Menzies and List (2009, 2010) went further than providing a reformulated exclusion principle that allows downwards exclusion as well as upwards exclusion. Their joint work on the topic resulted in what they call a compatibility result. That amounts to the possibility of special cases where neither downwards, nor upwards exclusion takes place. In cases of compatibility of higher and lower-level causes both the mental cause and the physical realizer comes out as a proper, proportionate difference-maker.

This extends beyond the insight that the direction of exclusion is a contingent matter, it tells us further that it is a contingent matter whether exclusion takes place at all. The most straightforward and the only case of compatibility Menzies mentions is the identity of mental

---

[72] Woodward (2008, 2015) developed a compatibilist solution from quite similar starting points, where higher-level and lower-level causes can be causes of the same outcome, but I won't discuss his approch in this thesis.

[73] In the case of Kim's argument this was always the case. Closure forced the choice between higher and lower-level candidate causes in favour of the realizer, so the realizer excluded the realized from causal efficacy in all possible cases.

and physical properties. In such a case the difference-making powers of the mental and physical causes are identical as well. He allows for other possible interpretations but cannot provide examples. Later, I will show that more interesting cases can be constructed as viable interpretations of the compatibility option, cases the existence of which have further interesting consequences with respect to the status of the new exclusion argument.

Making sense of this latter option Menzies and List introduced a distinction that seems to have created a bit of confusion among some critics (such as Shapiro 2012), so it is important to get things straight. I am talking about the distinction between realization-sensitive causal relations and realization-insensitive causal relations. Causal relations between typical multiply realized mental causes and their effects (where downwards exclusion takes place) are characterized as being insensitive to the manner of realization. Whereas in cases where mental and physical causes are causally compatible the relation between the higher-level cause and its effect is characterized to be realization sensitive. It is instructive to start with explaining the latter case. Realization sensitivity means that the higher-level causal relation is modally fragile. It may break down with even small changes in the realizer state. In contrast to this, realization insensitivity means that the higher-level causal relation remains stable under a wide range of changes in the realizer state, in the base properties on which the mental property supervenes.

It's worth summarizing the different options in terms of exclusion and the ways these are related to the stability or sensitivity of higher-level causal relations. Menzies and List highlighted some connections between the two issues, but they never clarified the relation

between upwards exclusion and causal sensitivity explicitly[74]. Table 4.1-1 aims to make the connections more transparent.

Table 4.1-1

The relation between candidate causes at different levels:	The stability of the higher-level causal relation (mental to behaviour/physical):
Upwards exclusion: physical realizer properties are the proportionate difference-makers of behavioural/physical outcomes.	***Stability is irrelevant*** in these cases, as there is no higher-level causal relation to characterize under those terms.
Downwards exclusion: realized mental properties are the proportionate difference-makers of behavioural/physical outcomes.	**Realization insensitive** higher-level causal relation. Stable over a wide range of changes in the realizer state.
Compatibility of levels (no exclusion): realized and realizer properties have identical difference-making powers.	**Realization sensitive** higher-level causal relation. It breaks down even under small changes in the realizer state.

With this overview at hand it will be easier to process the issues raised by the new exclusion argument. In the first major part of this chapter and also in parts of the following chapter (sections 4.2-4.3 and 5.5-5.6) I will argue that the new exclusion argument formulated by Menzies and List against Kim's original exclusion argument based on the counterfactual theory of causation suffers from some internal problems as an account of causal relations at lower levels. The next chapter will give more attention to realization insensitivity and its uses in understanding higher-level causation and the issue of reduction.

---

[74] Menzies (2013) haven't even mentioned upwards exclusion explicitly. That paper only talks about downwards exclusion (as a case of realization insensitivity) and compatibility (as a case of realization sensitivity). This seems to highlight the importance of the connection between those concepts but leaves the evaluation of upwards exclusion to the reader's judgement.

4.2 Counterfactuals, proportionate difference-making and exclusion

4.2.1 The basics of counterfactual causation

To decide whether downwards exclusion from above and upwards exclusion from below are justified ideas, first we have to pin down what we mean by causation for the purposes of this investigation. The classical counterfactual theory of causation considers token causation, a relation that holds either between particular events, or particular events exemplifying properties. Lewis understood those properties as the essences of the events in question, but for the discussion below the precise interpretation of events is not decisive. Any interpretation that takes events to be space-time regions instantiating properties is compatible with the approach Menzies and List took, and also with the argument I will put forward.

Applying this loose framework to the discussion of the causal exclusion argument, we can say that in the case of lower-level realizers we talk about a token event that exemplifies lower-level physical realizer property P, presumably a complex property that is the supervenience base for a higher-level realized property, a mental property M. In the case of the realized property we are talking about a token event exemplifying M. The two events occupy the same space-time region but instantiate different properties. These two events wouldn't count to be distinct if one were to identify events only in terms of the space-time region occupied, but neither Lewis nor Kim or Menzies advocate such a view. So, I think, for the discussion that follows, it is safe to suppose that the relevant higher and lower-level events are distinct. The question we will track below is whether the higher or lower-level property has causal prominence and why.

Lewis (1973, 1986) defines the causal relation as follows in terms of possible worlds semantics[75]: C causes E if and only if C and E both 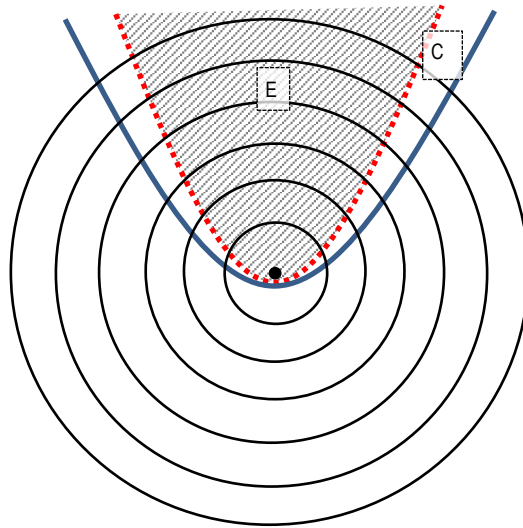occur in the actual world and in the closest possible world without C there occurs no E either (or alternatively if C and E are both absent in the actual world and in the closest possible world in which C occurs E occurs as well).

Figure 4.2-1

To guarantee the transitivity of the causal relation Lewis had to amend the simple definition, as counterfactual dependence is not transitive[76]. The amended definition says: event C causes event E if and only if there is a chain of events C, D1, …, Dn, E such that members in the chain are counterfactually dependent on the events preceding them. So, the causal relation is a descendent of counterfactual dependence.

It is important to highlight some basic details of this counterfactual test for causal claims. First, it requires us to fix the present and the past of C, the cause, than to move to the closest possible world that differs from the actual world with respect to C. If this alternative ¬C world is a ¬E world as well, then C is a cause of E. Figure 4.2-1 describes this situation using

---

[75] Obviously, some further basic requirements should also be met. C and E should be non-overlapping events and excluding the possibility of time-travel C should happen before E. Also backtracking counterfactuals are prohibited.

[76] I say more about this in the chapter on transitivity and causation. There were two main motivations for this move. First, intuitively causation is transitive. Distant causation where the cause and the effect are connected via intermediaries requires this assumption but also cases of early preemption that can only be handled by an account that stipulates that causation is transitive.

a Lewisian similarity space with the actual world in the middle. The circles around the actual world consist of points representing worlds which are equally similar worlds to the actual world. The closer the points (possible worlds) of a circle are to the actual world the more similar they are to the actual world. Distance is measured in terms of overall similarity. The farer a world is from the actual world the smaller is the overall similarity between the two. Overall similarity allows many different ways of being different for the same level of similarity.

The diagram does not work with a continuous similarity space, there are no worlds between the circles, only on the circles. This allows me to draw boundaries between relevant regions in a more transparent manner. Using this convention to be able to separate relevant regions from each other boundary lines do not need to overlap. On Figure 4.2-1 worlds in the cross-hatched region with a dashed border are all E worlds, the region encompassed by the continuous convex line contains worlds allowing C to occur. It is easy to see that if one moves to the closest world without C on the innermost circle it is a ¬E world as well as the region of C worlds contains the region of E worlds. So, according to the classical analysis C is a token cause of E in the actual world as both C and E occur in it.

This simple account was never used by Menzies, not even in his first (2003) formulation where he tried to spell out his ideas concerning mental causation for the first time. In papers written together with List (2009, 2010) they relied on an interestingly modified version of the classical counterfactual theory, therefore before developing my arguments I will go through the details of that modified analysis. To see the reason why this modification was important I will reiterate the basics of the Menzies account first.

4.2.2 Proportionate causation and the semantics of counterfactuals

According to the argument put forward by Menzies and List a multiply realized mental property (M) is a better candidate cause for some behavioural outcome (B) than any of its possible realizers P11 or P12 because only M counts as a proper difference-maker cause for B and none of those realizers (see Figure 4.2-2 that follows the structure of Kim's depiction). They arrive at this conclusion by introducing the constraint that causes should be proportionate to their effects. In contemporary literature proportionality is taken to be a vexed principle. Quite a few philosophers have recently debated whether this is a legitimate constraint on causation at all (e.g.: Bontly 2005; Dowe 2010; Shapiro and Sober 2011; McDonnell 2017b). One prominent way of criticizing the reformulated exclusion argument is this 'causation first strategy' as McDonnell calls it (see: Christensen 2018:8). However, for the sake of the present inquiry I will accept it as a sound constraint because my aim is to show that
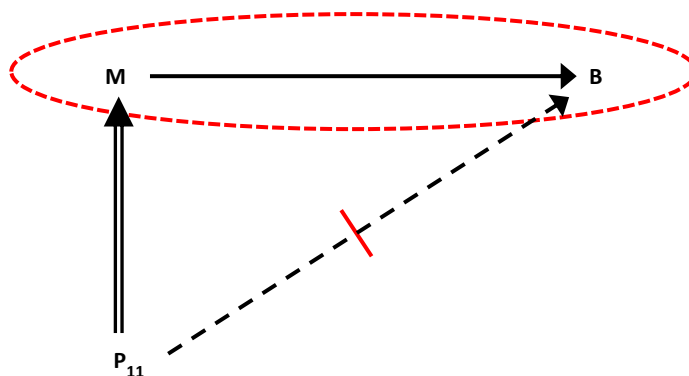
Figure 4.2-2

even if one accepts it, proportionate causation delivers somewhat different goods than Menzies and others hoped it would.

The proportionality of causation means that the cause should be neither overly specific nor insufficiently specific to the effect. If something is insufficiently specific, then its presence does not guarantee the presence of the effect. Let's suppose that it is true to say that seeing something of red colour (C) would make a bull angry (E). In some cases, it is also true to say that something coloured (C1) would make a bull angry, but this is true only when "being coloured" refers to something that is red. If the object is green, yellow, brown or any

other non-red basic colour it might count as coloured as well but according to the popular folk causal hypothesis it would not excite a bull. Therefore, highlighting that the object shown to the bull was coloured pointing to the cause of the angry response from the bull that ensued would be incorrect; it is insufficiently specific as it does not guarantee the occurrence of the effect. So, the following counterfactual statements are true:

i. $\neg(C1 \,\square\!\!\rightarrow E)$

ii. $\neg C1 \,\square\!\!\rightarrow \neg E$

Point i. tells us that C1 (coloured) is an insufficiently specific cause and therefore it is not a proper difference-maker for E. ii. tells us that C1 is not overly specific, so its absence guarantees the absence of the effect.

If something is an overly specific cause then its presence does guarantee the presence of the effect, but its absence does not guarantee the absence of the effect. Suppose again that seeing something red (C) makes a bull angry (E). It is also true that seeing something scarlet (C2) would make it angry as scarlet is a shade of red, but scarlet is overly specific as a cause because in a world where the object in front of the bull is not scarlet, but another shade of red E still occurs. For instance, candy apple is still a shade of red therefore it makes the bull angry, so scarlet is not a proportionate cause, as its absence does not guarantee that the bull won't be angry, holding other things fixed. Therefore, the following statements are true with respect to C2:

iii. $C2 \,\square\!\!\rightarrow E$

iv. $\neg(\neg C2 \,\square\!\!\rightarrow \neg E)$

Point iv. tells us that scarlet is an overly specific cause and therefore it is not a proper difference-maker for E. According to the proportionality principle utilized by Menzies and List

(2009, 2010) only seeing the colour red (C) is a proper cause for the angriness of the bull (E), and this is expressed by the following pair of counterfactual statements:

v. $C \; \square\!\!\rightarrow E$

vi. $\neg C \; \square\!\!\rightarrow \neg E$

This picture of causation is fairly intuitive but deviates from Lewis' account with respect to the evaluation of statements i., iii. and v. above. For Lewis there is only one world that is the most similar to the actual world and it is itself. Even though this is an intuitive starting point, according to Menzies and List (2009), this makes Lewis' analysis one-sided, incapable of identifying insufficiently specific causes. If Lewis' assumption is accepted, then if the actual world has C and E in place $C \; \square\!\!\rightarrow E$ is trivially true and to know whether C causes E or not we should only check what resides in the closest possible ¬C world.

This is the reason why in Lewis' framework preconditions for the cause are also counted as causes of the effect. In the context of Lewis' account being coloured (C1) (not being black[77]) would also count as the cause of the bull's attack. For something to be red it has to have a colour, if it has no colour (it is black) it isn't red either. If in the actual world we have C and E in place, then we also have C1 and E in place. $C1 \; \square\!\!\rightarrow E$ is trivially true and checking what resides in the closest ¬C1 world we get ¬E as an answer, which means that on that account C1 causes E. From this it follows that Lewis' framework provides no means to identify cases where one relies on a too loose description for the cause.

This leaves the door open for insufficiently specific causes in Lewis' account. The problem with too lose candidate causes like C1 is that they might not be sufficient for the occurrence of the effect. To get rid of them Menzies allows that there are more worlds equally

---

[77] According to at least some theories of colour black is the absence of colour as it is the absence of light.

similar to the actual world instantiating the cause[78]. This is what he calls weak centring of worlds in the system spheres around the actual world. It replaces Lewis' strong centring where no world can be as similar to the actual world as itself, so the smallest sphere around the actual world contains only the actual world.

In our example, equally close worlds would be worlds with coloured (non-black) C1 objects in front of the bull, some with an angry bull, but some others without it. This conceptual framework allows us to check whether or not the presence of a coloured object is sufficient for the effect to occur as we can get a more informative answer to the question whether $C1 \,\square\!\!\rightarrow E$ is true than we had in Lewis' framework. Checking all the closest C1 worlds we see that there are worlds with C1 without E and this means that C1 is an insufficiently specific cause. While checking all the closest C (red) worlds we see that there are no worlds with C without E which means that C is a sufficiently specific cause.

Adapting how Menzies and List redefined the similarity relations in the system of spheres we get the following criterion for causation. C causes E iff both $C \,\square\!\!\rightarrow E$ and $\neg C \,\square\!\!\rightarrow \neg E$ are true in the actual world w (presupposing that it is a C and E world):

A. $C \,\square\!\!\rightarrow E$ is true in world w if and only if E is true in all the C-worlds within the smallest C-permitting sphere of worlds around w. (*C is sufficient for E*)

B. $\neg C \,\square\!\!\rightarrow \neg E$ is true in world w if and only if ¬E is true in all the ¬C-worlds within the smallest ¬C-permitting sphere of worlds around w. (*C is required for E*)

These two criteria tell us that deciding the truth value of a causal statement relative to the actual world requires us to move to the closest worlds with C and also to the closest

---

[78] According to Menzies there are further, independent motivations for this move. Lewis' interpretation allows for the inference rule strenghtening the antecedent for counterfactuals. So, if $C \,\square\!\!\rightarrow E$ is true, then $(C \,\&\, A) \,\square\!\!\rightarrow E$ should be true as well. But this should not be true for counterfactuals, at least not in all cases.

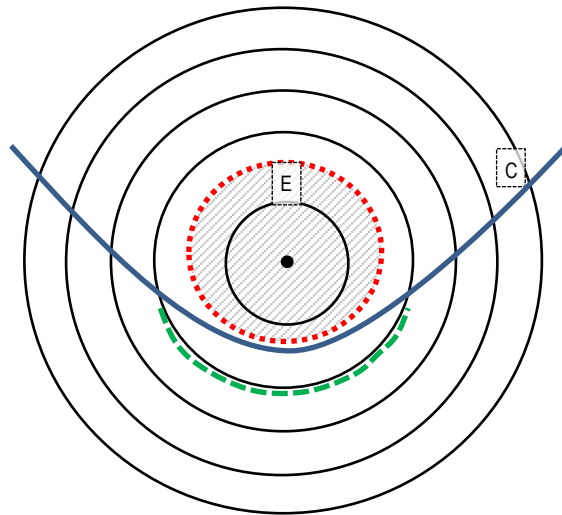worlds with ¬C and to check what else do we find there. This formulation of the semantics for counterfactuals and causal statements was built to include proportionality into the counterfactual account.

Figure 4.2-3

On Figure 4.2-3 one can follow the changes compared to Lewis' system. There is a black circle called the smallest C-permitting circle around the actual world. This includes worlds that are equally similar to the actual world as it is to itself under the event descriptions used. They are all C worlds, but as this description allows more determinate cases of C in some of these worlds we have scarlet objects, in others crimson, etc. According to Figure 4.2-3 all of these worlds also have E in place, therefore C is a sufficiently specific cause of E. Criterion 1. rules out insufficiently specific causes (like being coloured) of E efficiently.

There is also a black circle around the latter circle called the smallest ¬C-permitting circle around the actual world. Below the blue line on Figure 4.2-3 it includes worlds that are equally dissimilar to the actual world and which are all ¬C worlds. E is absent from all of these worlds; therefore, C is required for the occurrence of E (see the green dashed curve). Criterion 2. rules out overly specific causes (like being scarlet) of E efficiently. As Figure 4.2-4 makes it even more visible, scarlet (C2) would be an overly specific
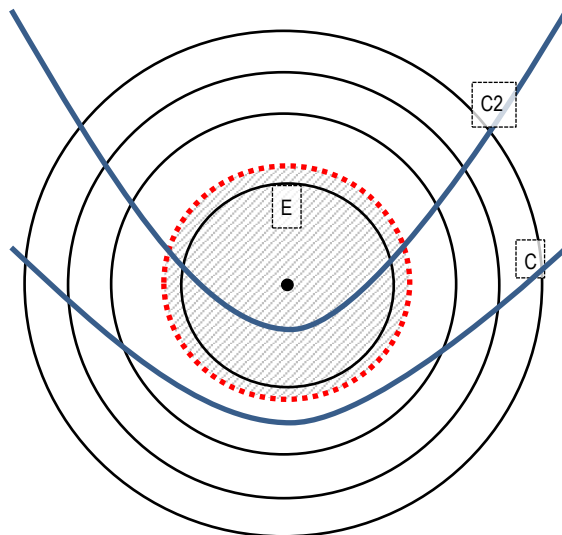
Figure 4.2-4

cause of E because all ¬C2 worlds on the smallest C2 permitting sphere (which is the 1st circle) are E worlds as well. This is because they are C (red) worlds.

Now, that the semantic theory for proportionate causation is clarified, accepting that this is the right semantics for counterfactuals the road is open to follow Menzies in turning the exclusion argument upside-down.

4.3   The reformulated exclusion principle in detail

Let see how Menzies and his allies turn the exclusion argument on its head! First, using the theory of causation we just described they formulated a version of the exclusion principle that says something fairly close to Kim's principle.

Revised exclusion principle: For all distinct properties M and P such that M supervenes on P, M and P do not both cause a property B.

The principle has two possible applications, the second of which was ruled out by physical causal closure in Kim's rendition. The first is a close counterpart of the application Kim favoured. He assumed that the upwards formulation always wins out because of physical closure, and because physical causes are always sufficient for their effects.

Upwards exclusion principle: If a property P causes a property B, then no distinct property M that supervenes on P causes B.

The second application is only available if causation is interpreted as difference-making, and there are empirical cases where higher-level causes win out as difference-makers.

Downwards exclusion principle: If a property M causes a property B, then no distinct property P that subvenes or realizes M causes B.

Before we move on, we should note that in this new regimentation of the exclusion problem the closure principle does not play the role it played in Kim's version. The new exclusion argument might go like this. There is a Mental event M that is supposed to cause behavioural outcome B. There is a subvening event P below M. M and P are numerically distinct candidate causes of B. If M is multiply realized, then P cannot be a difference-maker for B. But M is a difference-maker for B, so downwards exclusion should be applied. Empirical scenarios that

involve details fitting this description confirm our conviction that mental properties can be real and efficacious causes. For Menzies neuroscience proves this about e.g. ordinary intentions, but probably the same will be true in other cases as well. Now we see, that the decision between competing candidate causes is not made based on the causal completeness of the physical realm, but instead is based on what causes the outcome in question according to the difference-making theory of causation utilized.

As already mentioned, a radical implication of the work Menzies did together with List is the contingency of the revised exclusion principle itself. Menzies and List proved that the compatibility of causes at different levels for the same outcome is a live option. As we will see, such inter-level causal compatibility was advertised as a purely theoretical result and nothing was said about the plausibility such scenarios. By compatibility and incompatibility (the latter includes downwards and upwards exclusion) between candidate causes Menzies means a relation between the following two pairs of counterfactuals:

Realized property (M) instance causes some behaviour (B) in the actual world iff:

  i.   in the closest worlds where there is M there is B as well and

  ii.   in the closest worlds where there is ¬M there is ¬B as well.

Realizer property (P) instance causes some behaviour (B) in the actual world iff:

  iii. in the closest worlds where there is P there is B as well

  iv.   in the closest worlds where there is ¬P there is ¬B as well.

Menzies and List came up with precise conditions for the compatibility and the incompatibility of these two causal statements, two pairs of counterfactuals. In the following subsections I will analyse the available options step by step following their lead however introducing small modifications where I see room and motivation for that.

4.3.1   Conditions for upwards exclusion

Let us take a closer look at how Menzies and List spell out the conditions for upwards exclusion. Figure 4.3-1 depicts possible examples in the abstract. It is an enriched version of the diagram in Menzies and List (2009:494). The convex region encompasses all M worlds, the second convex region inside the former encompasses all P worlds, where P is one possible realizer of M. The dotted line encompasses the region where B is present. It is easy to see that M is not proportionate to B, but P is. So, M is excluded from causal efficacy. This how the general condition for upwards exclusion is formulated[79]:

> Necessary and sufficient conditions for upwards exclusion: An instance of upwards exclusion occurs if and only if (i)[80] P is a difference-making cause of B[81] and either (ii) B is absent in some closest M-worlds that are ¬P worlds or (iii) B is present in some closest ¬M worlds outside the smallest ¬P permitting sphere.
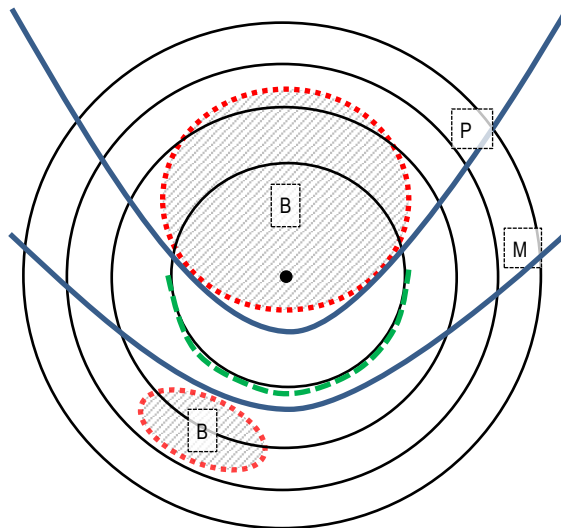
Figure 4.3-1

This is all fair, but I think everybody would want to see a concrete example. Woodward (2008), arguing for a parallel but different solution for the exclusion problem, provides a fitting example. This example satisfies both condition (i) and (ii) that is sufficient for upwards exclusion. Imagine a

---

[79] Stating the condition, I follow the original Menzies and List (2009:493) formulation, I changed is the notation used for properties to match mine and also the numbering of the subconditions (see: footnote 80).

[80] In the original text the first subcondition lacked a number for some reason whereas in the condition for compatibility this wasn't so. Therefore, for consistency's sake, I changed the numbering of the subconditions.

[81] This means that *B is absent in all closest ¬P worlds and B is present in all closest P worlds*.

croupier who can use a variety of hand motions to spin a roulette wheel. A kind of hand motion is multiply realized by micro muscle movements (P, R, S etc.), and the micro movements determine whether the ball ends up in a certain slot. However, the kind of hand motion the croupier chooses (M) makes no difference to where the ball ends up (B). This is why the croupier can't cheat, she has no control over the outcome. So, we determined the explanandum, the outcome, we wanted to know why the ball ends up in this slot rather than in another and we found that this difference can only be accounted for in terms of fine-grained physiology the croupier has no access to and no control over. On the diagram the dashed green curve signifies the closest M-worlds that are ¬P worlds. These are worlds where the croupier utilizes some hand movement (M) that is not realized by P on the physiological level. So, a concrete position of the roulette ball is caused by a particular realizer of some kind of hand movement done by the croupier, but not by the macro-level hand movement itself.

Now, I think condition (ii) requires further clarifications never explicitly provided by Menzies or List. If we want P to exclude M, then according Menzies' account P should be a proportionate cause of B without M being a proportionate cause. Condition (ii) requires that there are closest M worlds that are ¬P worlds without B in place. The existence of such worlds ensures that M is not proportionate to B. These worlds are situated somewhere on the green dashed curve on Figure 4.3-1. Note further that condition (i) cannot be met if condition (ii) is not met. However, the opposite doesn't seem to be true.

Imagine that M has three possible realizers P, R, S. R is situated in the set of closest ¬P worlds that are M worlds (somewhere on the dashed green bottom of the innermost sphere on the diagram) where B is in place (the diagram does not show such option). S is situated in the set of closest ¬P worlds that are M worlds where B is absent. Assume that P is closer to R

in the similarity space than P is to S. This scenario also ensures that M isn't proportionate to B as some closest M worlds are ¬B worlds.

But this scenario is incompatible with P being proportionate to B. If P and R are neighbours in the similarity space, while P and S are not, so some of the closest ¬P worlds are R worlds, but no closest ¬P worlds are S worlds, then only the disjunction P v R is proportionate to B. This is something the upwards exclusion criterion was designed to preclude as condition (i) rules out such scenarios. However, if the closest ¬P worlds were all S worlds than both (i) and (ii) would be satisfied.

This is an interesting issue, let us stop here for a while. So, for the first sight, it seems that Menzies' approach to inter-level causal competition prohibits disjunctive properties as lower-level candidate causes. But if some of the closest ¬P worlds are R worlds, but no closest ¬P worlds are S worlds and P and R worlds are B worlds then nothing substantial prohibits P v R from being the proportionate cause of B if the closest S worlds are ¬B worlds. We are simply confused by the notation used. The only thing required to ensure that a disjunctive property is a cause that excludes M is that there should be at least one realizing state for M, like S, that is not accompanied with B. We simply have to substitute P for P v R in criterion (i) to get a scenario like that.

I don't want to commit myself to anything considering the status of disjunctive properties here. The only thing I would like to point out, is that if there are many possible realizers for M and among them there is more than one that would be sufficient for the occurrence of B then the choice forced by the new exclusion principle is not necessarily between one particular lower-level candidate cause P and higher-level candidate cause M, but between a disjunctive lower-level cause and M. If we accept disjunctive properties as causes, it seems to be true that the upwards exclusion condition pins down all possible

scenarios where M would be excluded from causal efficacy by a disjunctive lower-level property. If we are motivated to avoid disjunctive properties as competitors for M, we have to modify the whole upwards exclusion condition adding the proviso that P is not a disjunctive property.

Menzies never provided an example for condition (iii), but its function is clear. This condition could be satisfied by a property that is not a realizer of M (in the last example a hand movement the croupier can control), if it is also capable of causing B. Figure 4.3-1 has a smaller B region at the bottom on the left, on the second sphere around the innermost sphere below the convex curve encompassing M worlds that refers to such possibilities. Those B worlds are closest ¬M worlds outside the smallest ¬P permitting sphere, exactly what condition (iii) required. The function of (iii) is to ensure that M comes out as an insufficiently specific cause even if (ii) is not satisfied under the circumstances because some closest ¬M worlds have B.

4.3.2 Conditions for downwards exclusion

Let's now turn to the case of downwards exclusion. Figure 4.3-2 depicts a possible example in the abstract. It is a close replica of the diagram in Menzies and List (2009:496). The lower convex region encompasses all M worlds, the other convex region inside the former encompasses all P worlds, where P is one possible realizer of M. The dotted red line encompasses the region where B is present. It is easy to see that P is not proportionate to B, but M is. So, P is excluded from causal efficacy. The general condition for downwards exclusion is as follows[82]:

> <u>Necessary and sufficient conditions for downwards exclusion</u>: An instance of downwards exclusion occurs if and only if (i) M is a difference-making cause of B[83] and (ii) B is present in some closest ¬P worlds that are M worlds.
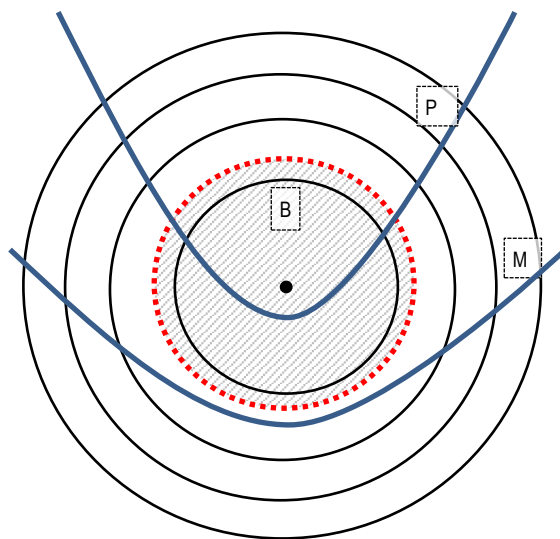
Figure 4.3-2

For believers of multiple realizability examples of downwards exclusion are easy to find. Imagine that you are thirsty, and you believe that there is a bottle of soda in the fridge (M). The belief will move you to go to the fridge, to open it and to drink some soda (B). The belief that there are no drinks at home would have the effect of moving you to go shopping (supposing that you are thirsty enough). The realizer of M, some brain state

---

[82] Stating the condition, I follow the original Menzies and List (2009:495) formulation. I changed the notation used for properties to match mine and included numbers for the subconditions as, for some reason, in the original text the condition for downwards exclusion came without such numbers while the others had them.

[83] Which means that B is absent in all closest ¬M worlds and B is present in all closest M worlds.

(P), has no such difference-making power. As it can be seen on Figure 4.3-2 the closest possible ¬P worlds with M in place are all B worlds. In those worlds a different state realizes M, so P is not a proportionate difference-maker for B.

The function of subcondition (ii) is to guarantee that P is not a proportionate cause of B. If B is present in at least some closest ¬P worlds, then P cannot be a difference-maker for B according to the proportionality constraint as in that case P is not required for B. However, "at least some" in (ii) seems to be a too weak constraint when (ii) is paired with (i). So, let us check whether (ii) is consistent with (i). Is it consistent with (i) that some closest ¬P worlds that are in the set of M worlds are B worlds[84]? There are two available interpretations of (ii) to consider. According to the first all closest ¬P worlds that are also in the set of M worlds are B worlds. This is surely consistent with (i) as we saw in the previous paragraph.

According to the second interpretation of (ii), some, but not all closest ¬P worlds that are also in the set of M worlds are B worlds at the same time. This interpretation is problematic. It implies that there are some closest ¬P worlds that are M worlds where B is absent. Note however, that these worlds cannot be among the closest M worlds (or ¬M worlds). That would contradict (i), because M is only a proportionate difference-maker for B if M is sufficient for B. If there are closest M worlds without B, then it cannot be sufficient.

If B is present only in some, but not all closest ¬P worlds that are M worlds then P cannot be excluded by M, but it is possible that it is excluded by a disjunction of lower-level realizer properties for M. If M has realizers P, R, S, assuming that P and R are closer in the

---

[84] Such worlds are between the convex curves on Figure 4.3-2. The top of the innermost sphere contains P worlds, but as the convention used on this diagram dictates it is the closest ¬P permitting sphere as well the bottom part of which below the convex line encompassing P worlds has only ¬P worlds that are M worlds as well.

similarity space than P and S, and in the closest ¬P worlds that are M worlds as well B is always accompanied with R but never with S, then P is excluded by P v R but not by M. Again, such scenario would be compatible with condition (ii), but incompatible with condition (i).

Therefore, closest ¬P worlds that are also M worlds and where B is absent could only be somewhere far away in the similarity space, maybe on the third or fourth (or on any outer) circle between the convex curves on Figure 4.3-2. But such worlds to be able to conform to the criterion should be quite different in some respect other than what is highlighted by the property referred to in the antecedent of the relevant counterfactual. They are faraway M worlds without B (farer from actuality than closest ¬M worlds) which implies that they should be different from the actual world in terms of the background conditions as in those worlds something has to block M from being able to bring about B. But in that case, they cannot be closest ¬P worlds that only differ from P worlds with respect to their P-ness. So, the second interpretation of subcondition (ii) is inconsistent with (i)[85].

Obviously, this is not a problem for the downwards exclusion condition as a whole in a purely logical sense because the truth of (i) and (ii) is still compossible and we can say (i) excludes the problematic interpretation of (ii) explained in the last paragraphs. In the special case when B is present in all closest ¬P worlds that are M worlds both condition (i) and (ii) are satisfied. But only this second interpretation of (ii) counts as a meaningful addition to the content of the condition for downwards exclusion.

Therefore, condition (ii) in its present form is a misleading formulation. So, I think it is useful to formulate a modified version that provides a more apt explanation of what the

---

[85] One might also ask the following question: is it possible that a closest ¬P world is farer away than a closest ¬M world? Well closest ¬P worlds are either closest M worlds or closest ¬M worlds at the same time. So, this is not possible.

downwards condition as a whole requires in lower-level terms, what has to be true about realizer states for downwards exclusion to take place:

Necessary and sufficient conditions for downwards exclusion*: An instance of downwards exclusion occurs if and only if (i) M is a difference-making cause of B and (ii) *B is present in all closest ¬P worlds that are M worlds as well*.

I need to highlight one non-trivial feature of the presentation used by Menzies and List (see Figure 4.3-1 and Figure 4.3-2 again, or later the diagrams in this section). Realizer and realized properties are placed on the same diagram. The presentation creates the impression that the worlds containing the different realizer states are close neighbours. Now, this is reasonable if we approach the situation from the point of view of the realized property. Worlds containing different versions of M should be close neighbours if we structure the similarity space based on some description in terms of M. This perspective is less natural if we approach from the point of view of realizer states. Would they occupy such close positions in the similarity space on their own terms? Should they be neighbours just because they are realizers of the same higher-level property? It surely depends on the similarity relations between the different realizers. At this point it would be difficult to come up with a general answer. But note that there is an issue to think through. I will get back to this issue in the next chapter in section 5.2 and the sections following that section.

4.3.3 Inter-level causal compatibility

The last case I discuss is inter-level causal compatibility, in other words the general condition

for the compatibility of the four counterfactuals (i.-iv.) formulated at the end of the intro

section to the discussion of the reformulated exclusion principle (4.3)[86]:

<u>Necessary and sufficient conditions for compatibility</u>: Lower and higher-level

causes are compatible (don't exclude each other) if and only if (i) B is present in all

closest M worlds; (ii) B is absent in all closest ¬M worlds[87]; and (iii) B is absent in all

closest ¬P worlds that are M-worlds.
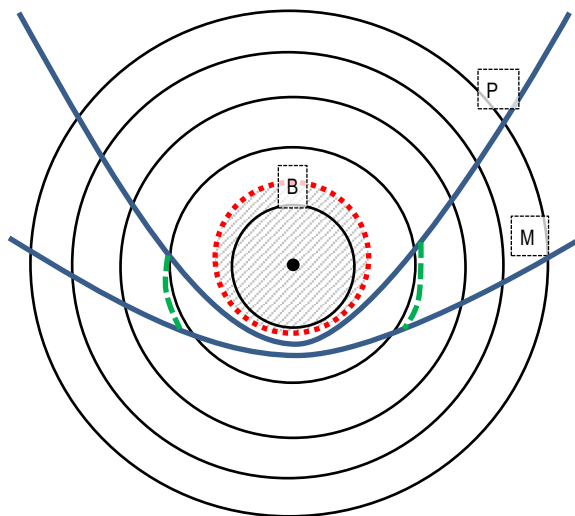
Figure 4.3-3

Conditions (i) and (ii) are trivial, together they state that M is a proportionate difference maker for B, while (iii) is way less non-trivial, so I will get back to it after providing some introductory clarifications. What could serve as a good practical example for the compatibility condition? Menzies (2013:79-80) says that if M and P were

identical the compatibility condition would be satisfied, but he is also explicit in allowing for

other possible interpretations. I should mention here, that judging from the limited discussion

---

[86] Stating the condition, I follow the original Menzies and List (2009:491) formulation. I changed two things. First, the notation used for properties to match mine. Second, before "if and only if" I talk about inter-level causal compatibility instead of the compatibility of the four counterfactuals (i.-iv.) you find at the end of the intro of section 4.3. But these two formulations refer to the same thing.

[87] For some reason in the original text by Menzies and List splitted the condition that M should be a difference-maker cause of B into (i) and (ii) in the case of compatibility, but not elsewhere.

Figure 4.3-4

Menzies gives to multiple realization he would treat the identity of M and P to be a highly unlikely scenario.

Note that if the identity of P and M was the only possible interpretation of compatibility, we would be better off with a diagram where the convex curve encompassing P worlds is m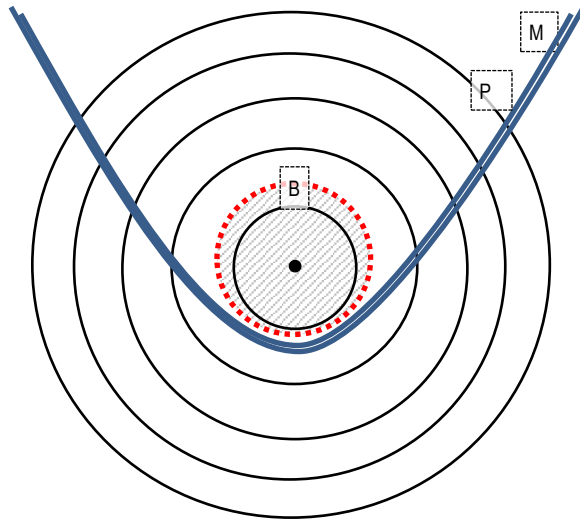atched on the convex curve encompassing M worlds (see: Figure 4.3-4). Note also that if compatibility was restricted to the case of identity between M and P then subcondition (iii) would be vacuously true because in that case the closest ¬M worlds would simply be the closest ¬P worlds, so there would be no closest ¬P worlds that are M worlds at the same time. However, we are provided with a different diagram (see: Figure 4.3-3) and condition and these seem to allow for more. As Menzies allowed for cases of compatibility beyond identity, I think that we are invited to think about the motivation behind or the possible uses of subcondition (iii).

(iii) says that B is absent in all closest ¬P worlds that are M-worlds. ¬P worlds are below the curve encompassing P worlds. As the innermost sphere only contains P worlds the closest ¬P worlds are on the second sphere around the actual world. Some of them are M worlds as well. The relevant worlds are highlighted by dashed green curved lines on Figure 4.3-3 between the curve encompassing P worlds and the curve encompassing M worlds. Let us focus on this gap between the curves.

The existence of this region seems to mean that there are closest possible ¬P worlds where M is not realized by P (but by R, S, etc.) and these are ¬B worlds. As the first sphere contains closest P worlds, the second sphere is the closest ¬P permitting sphere. The green

dashed curves signify those regions on the second sphere (closest ¬P worlds) where M is in place and realized by other realizers of M like R, S. But if these worlds are among the closest possible M worlds as they should be then it follows that M is not a cause of B, which is in contradiction with (i) and (ii). Therefore, for (iii) to be consistent with (i) and (ii) the closest ¬P worlds in question should be M worlds that are not among the closest possible M worlds. But that is not possible.

From this it seems to follow that subcondition (iii) requires something that is only consistent with (i) and (ii) when (iii) is vacuously true as there are no closest ¬P worlds that are M worlds without B. If properties M and P are identical this is definitely true (Figure 4.3-4) as all ¬P are ¬M worlds. If all this stands true, then the only meaningful scenario is identity.

However, I think to find a more interesting interpretation for (iii) we should attend to
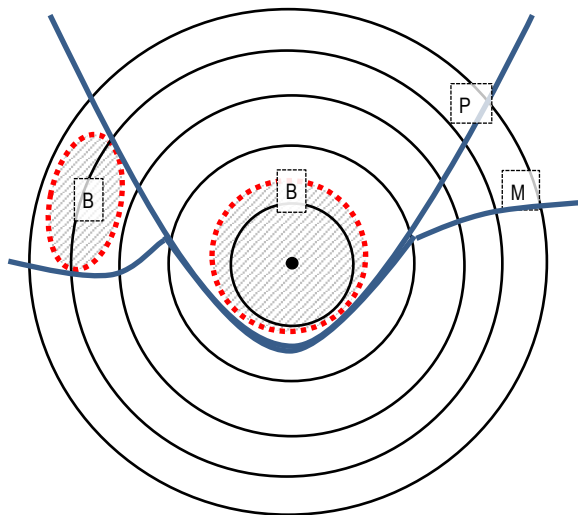
Figure 4.3-5

what the subcondition allows. Read charitably, and taking into consideration Menzies' suggestion that the compatibility condition should allow for cases beyond identity, it seems plausible that (iii) is there to make sure something I would call the *local compatibility* of M and P as causes of B while at the same time it allows for the distinctness of M and P. In modal terms this would amount to the possibility of far-away worlds without P where M exists accompanied by B (see the left-hand side of Figure 4.3-5 between the convex curve encompassing P worlds and the concave curve encompassing M worlds). Both Lewis' account of counterfactuals and Menzies' modified version allows for worlds that are different from the actual world not only in terms of the property referred to

by the antecedent of the counterfactual in question but also in other respects, in terms of what in most theories of causation one would call the background conditions. So, my proposal is this. M and P can be compatible causes of B even if properties M and P are not identical if, assuming that locally M and P are both proportionate causes of B, there are worlds that are not closest P or ¬P worlds and not closest M worlds and where M exists accompanied by B and where M is realized by a ¬P realizer S, R, etc. I will add more meat to this option later in section 5.6, for now suffice it to say that identity is definitely not the only viable interpretation of inter-level causal compatibility. As I will show later this is a result that could have been a surprise for Menzies both in a positive and in a negative sense.

So, we are left with three distinct cases for the application of the new exclusion principle. There are cases of upwards exclusion, when lower-level realizers exclude mental properties from causal efficacy. Downwards exclusion, when higher-level mental properties exclude their realizers from causal efficacy. And finally, there is the case when exclusion fails as both the realizer and the realized are proportionate causes of the same outcome.

4.4   Realization sensitivity in light of the exclusion principle

As I already mentioned there is another dimension Menzies and List utilized to characterize causal relations of interest. This is what they called the realization sensitivity of causal relations. Here the interest focuses on higher-level causal relations like M causing B. Intuitively sensitivity to realization means the following. To paraphrase Menzies, causal relations can be characterized as fragile or robust. A fragile causal relation is sensitive to how M, the cause is realized. A robust causal relation holds however the cause is realized.

More precisely a causal relation is robust if changes in the realization base of the cause do not disrupt the relation. Menzies calls this realization insensitivity. A higher-level causal relation has this quality if fixing other features of the actual situation and moving around to worlds with different realizers of M (P, R, S) the outcome B remains in place. As we saw in section 4.3.2 in such cases downwards causation takes place.

A causal relation is fragile if changes in the realization base of the cause do disrupt the relation. Menzies calls this trait realization sensitivity. A higher-level causal relation has this quality if fixing other features of the actual situation and moving to the closest world without the actual realizer of M, we get to a world where no realizer of M is in place and B does not occur either. As we saw in 4.3.3 these are the cases of inter-level causal compatibility.

Menzies and List are explicit in expecting that in cases of mental causation and higher-level causation in general realization insensitivity holds because of the multiply realized nature of the properties involved. This is accompanied by downwards exclusion as a parallel feature that is also based on the multiple realization of the cause. Whereas realization sensitivity takes place when higher-level properties are not realized multiply. The latter option is accompanied by inter-level causal compatibility. However, in light of the results of

the investigations in previous sections it is fair to say that this dichotomy leaves some issues

uncovered. Let me turn to these issues as a conclusion for this chapter.

4.5   Partial conclusions concerning the reformulated exclusion argument

First, as they never talked about the issue explicitly, one might wonder why cases of upwards

exclusion aren't characterized in terms of sensitivity to realization by Menzies and List (see:

Table 4.1-1)? The example they relied on for upwards scenarios does not fulfil the condition

for compatibility, but for the first sight it seems to fulfil the requirement of being realization

sensitive (see: Figure 4.3-1) as it is true that changes in how the putative cause is realized

effect the causal relation. In the closest worlds where the actual realizer of M is absent (a

particular uncontrollable hand movement the croupier performed) the effect B is absent as

well. The simple answer is that in the case of compatibility M is a proportionate cause of B

while in cases of upwards exclusion only P is. When the proportionality condition is not

fulfilled for M and B it is meaningless to talk about the realization sensitivity of the relation

between M and B.

So far so good, but there is another and more intriguing puzzle I would like to draw

attention to. Are there any interesting consequences of the fact that the compatibility

condition allows for more than cases of property identity? When in Menzies (2013) he says

that a natural interpretation of compatibility would be the identity of M and its realizer P, he

explicitly adds that he is open to other interpretations of the compatibility condition. The

abstract case I identified in section 4.3.3 as a new interpretation, that is consistent with the

compatibility condition allows that M can be distinct from P and multiply realized without

downwardly excluding its realizer from causal efficacy (see Figure 4.3-5 again). If that is a

viable interpretation, then it follows that the multiple realization of a proportionate cause is

not sufficient for (1) downwards exclusion and neither for (2) the realization insensitivity of the relevant higher-level causal relation. This is a significant insight as Menzies argues both that multiply realized higher-level causes downwards exclude their realizers from causal efficacy and that the relevant higher-level causal relations are realization insensitive. He never states this explicitly, but this implies that the multiple realization of proportionate higher-level causes is a sufficient condition for downwards exclusion and for the realization insensitivity of causal relations. I disagree with that view, so I will get back to this point later, when I have more conceptual resources to work out my interpretation of difference-making at different levels and after I clarified further issues concerning the realization relation that are relevant for the evaluation of Menzies' project.

Chapter 5: Exclusion and the content of closest possible worlds

The aim of this chapter is not to criticise the proportionality interpretation of counterfactual causation directly, but to highlight some plausibly unwanted consequences of it. I contend that lower-level realizers of multiply realized higher-level causes for a certain outcome are much less hopeless as candidate causes for the same outcome than Menzies et al. claimed. To see this, one has to attend to the nitty-gritty of how the relevant lower-level counterfactuals are evaluated on their own terms.

There is a worry concerning the reformulated exclusion principle that was raised by Shapiro (2012:15) discussing Menzies' project. In short, the question is this: what resides in the closest possible world without the actual putative lower-level cause property? A clear answer to it is fundamental to the reformulated exclusion argument. If that closest world contains another realizer property downwards exclusion takes place, if not then it is possible that a multiply realized property does not exclude its realizers causally. In what follows, I would like to develop on this worry in more detail than Shapiro did. His main idea was to show that his theory of multiple realization[88], combined with Menzies' exclusion argument shows that the problem of causal exclusion is a non-issue. I disagree with this diagnosis, later I will explain my reasons, but I do think that Shapiro raised an important question concerning causal exclusion and realization.

---

[88] A view that barely allows for multiple realization. Shapiro is basically a careful advocate of the identity theory. He considers himself a naturalist, and in most of his work on realization his aim is to show that the empirical evidence available in the sciences favors the identity theory over the non-reductivist alternative.

5.1   What determines similarity between realizers?

In this section, I will explain the importance of Shapiro's question concerning the content of the closest possible world where a putative cause is absent and will show how the presumed relation between realized and realizer properties influences possible answers to it. This investigation helps us to bring to light tacit background assumptions Menzies and List worked with developing the new exclusion argument.

Let's start with a quick recap of the reformulated exclusion principle in token causal terms. Let M be a multiply realized mental property (the belief that I have a beer in the fridge) and B an action that results from it (assuming that I am thirsty, I open the fridge and grab the beer). Let P11 and P12 be particular realizers of M. In the case of a higher-level realized property exemplified by an event (M) it causes the behavioural outcome (B) in the actual world iff:

  i.   in the closest worlds where there is M there is B as well and

  ii.   in the closest worlds where there is ¬M there is ¬B as well.

In the case of a lower-level realizer property exemplified by an event (P11) it causes the behavioural outcome (B) in the actual world iff:

  iii.   in the closest worlds where there is P11 there is B as well

  iv.   in the closest worlds where there is ¬P11 there is ¬B as well.

As we saw the question one has to decide to be able to evaluate the reformulated exclusion principle is which of these statements are true, when and for what reason? Menzies claims that in the case of multiply realized higher-level mental properties where M is a proportionate cause of B, P11 is excluded as a cause of B. The truth value of iv. depends on what resides in the ¬P11-worlds of the smallest ¬P11-permitting circle around the actual world. If Menzies and co. are right, then at least some closest ¬P11 worlds should be P12

worlds or worlds with any of the different possible realizers of M (P13, P14, etc.). As we saw, if they are, downwards exclusion applies, iv. is false and P11 is not the cause of B.

This is what we see on Figure 5.1-1. Note one difference compared to earlier diagrams. On this figure and on similar ones used later the P11 region might seem to include the P12 region. But this cannot be assumed so. Different realizers of the same higher-level entity and property are not compossible. If P11 is in place, P12 is absent and vice versa. So, the closest possible P12 worlds are situated between the two convex curves on the 2nd sphere around the actual world in the middle.
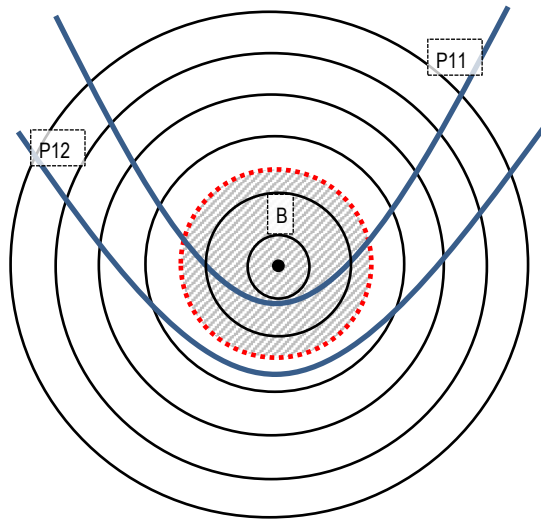
Figure 5.1-1

In this toy example the closest ¬P11 worlds are considered to be P12 worlds populating the smallest ¬P11 permitting sphere. The cross-hatched region in the middle has B in place. The actual world is inside that region and region P11. Moving to the closest possible ¬P11 worlds on the second sphere does not bring us outside of the B region. It brings us to P12 worlds that are B worlds as well. To reach a region in the possibility space without B in place we would need to reach at least the points of the third sphere around the point signifying the actual world in the middle.

But how do we know whether the closest ¬P11 world is a P12 world? If property P11 is closely similar to property P12 then this is quite plausible as in counterfactual tests we are asked to move the closest possible world(s) without the antecedent in place. The question is whether P11 and P12 are really similar or not.

When motivating the proportionality constraint and the modifications suggested in the semantics of counterfactuals Menzies illustrated the point with a classic example from Yablo (1992) in which a pigeon is trained to peck when it is shown something red. The example works like the example with the bull used earlier in section 4.2.2. Red is the true cause of the pecking and none of the particular shades of red like scarlet or crimson fit that role.

i.   Target is red □→ pigeon pecks

ii.   Target is not red □→ pigeon does not peck

iii.   Target is scarlet □→ pigeon pecks

iv.   Target is not scarlet □→ pigeon does not peck

The closest possible non-scarlet world is a world which matches the facts of the scarlet world except the scarlet colour itself, which should be replaced by the next closest shade on
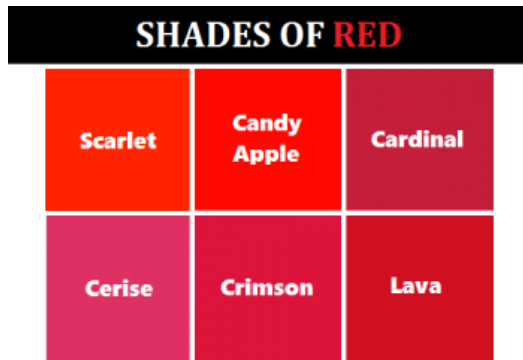
Figure 5.1-2

the red spectrum. Let's suppose now that it is candy apple (see: Figure 5.1-2). This is because in terms of colour that is among the closest colours to the actual colour. So, if P11 functions in the same way as scarlet and P12 functions in the same way as candy apple then P11 is not a token difference-maker for A1 as the closest ¬P11 worlds are P12 worlds. Notice that the similarity measure is dependent on the property descriptions or, to follow Lewis' way of thinking concerning events, on what we take to be the essence of that event. Similarity to the property highlighted by the cause event description used provides the baseline for deciding what is close to and what is far from the actual world[89].

---

[89] Lewis' essentialism is only one possible take on the issue. Alternatively, one might follow people who think that Lewis should have opted for an anti-essentialist, counterpart-theoretic view for which events are coarse-

In simple counterfactual accounts like Lewis' the standard used to decide what resides in the closest possible world in the absence of a cause is what McDonnell (2017a) calls the excision standard. Asking what caused Socrates' death (E) the usual answer is that it was Socrates' drinking hemlock (C). How do we know this? Fixing the past, we move to the closest world where the cause event ceases to exist. It is important that we go just as far as it is necessary to reach the border between C and ¬C, but we have to go at least that far.

If C is defined as "Socrates' drinking hemlock" we need to go to the closest world in which he does not drink the hemlock and, at the same time, to a world which is the closest to the world where he is drinking the hemlock in all other respects. This standard prevents far-fetched replacements of the antecedent in the relevant counterfactual. We can't just move to a world where Socrates instead of drinking hemlock fell off a cliff. And it also prevents alterations that are too small. Moving to a world where Socrates sips hemlock is not enough. We need to reach a world where the drinking of the hemlock completely ceases to exist. It is definitely underdetermined what exactly takes place in that world as it is plausible that the actual world has more than one immediate neighbour where C is completely absent, but where everything else is as close to actuality as it can be[90].

---

grained, world-bound occurrences as it fits his general metaphysical outlook better than his original essentialist construct (see: McDonnell 2016, Kaiserman 2017). According to this view the context sensitivity of counterfactuals and causal statements is built into the counterpart relation that tells us which otherworldly events should be taken as counterparts of an event in the actual world. The main advantage of this view over Lewis' original picture is that it avoids the double counting of events. We can talk about one event under many different descriptions. We can talk about 'my writing of this paper' and also 'my motivated writing of this paper' as being the same event, whereas for Lewis these would be different events with different essences. For the anti-essentialist the descriptions we use determine the counterpart relation. For my purposes in this chapter these differences are not particularly important as similarity relations are equally important for all of these accounts and the similarity relation is determined by the descriptions of the events involved.

[90] This might result in interesting problems; however, this is not an issue to tackle here.

Now one might think that there is something unrealistic in thinking that when Sophie the pigeon is not presented with scarlet pieces then it is presented with something crimson. This intuition stems from certain presuppositions concerning the situation. Under normal circumstances my room is not populated with objects of different shades of red. On the contrary, my present room only has one candy apple object and that's it. If I were to present a pigeon with the one red object sitting in my room, it would be straightforward that in the closest possible world where I present the pigeon with a non-candy apple object it wouldn't be red, at least according to the excision standard applied to token events. Here it is important to highlight that Menzies and List (2009) spelled out their account of causation and causal exclusion in terms of property instances, not in terms of events (see: Menzies and List 2009:478, footnote 9). This modification makes it far more natural to think that in the closest world where scarlet is absent another shade of red replaces it. And the answer to the question what makes it natural is that those properties are quite similar to each other. If the properties in question were less similar, then it would be less natural or even implausible to think that they are neighbours in the similarity space.

Arguing that proportionality is the right idea to solve the exclusion problem and arguing for the disproportionateness of lower-level candidate causes for higher-level effects Menzies says the following about counterfactuals involving lower-level realizers:

"It is natural to interpret these counterfactuals in terms of a similarity relation *that makes the closest-worlds in which the target is not crimson ones where it is some other shade of red*." (Menzies and List 2009: 488, my emphasis).

The statement of the quote above sounds plausible, but what makes it so? Even though Menzies is explicit in rejecting Yablo's interpretation of the realization relation (see:

Menzies 2008), the motivating concept behind all the examples provided is the determinable-determinate relationship between different colours. That relation makes it plausible that the closest possible non-crimson region in the similarity space utilized by Menzies:

- firstly, is confined to a different region of the red spectrum, and

- secondly, that in the closest possible worlds to the region where crimson occurs are worlds where colours like cerise or lava occur instead (see: Figure 5.1-2). The example of red versus crimson conforms to proportionality exactly because the determinate-determinable relation backs it up.

This way of approaching the non-occurrence of property instances is not only intuitive, but plausible in the case of colours because colours stand in relatively easily definable similarity relations to each other, they share their so-called determination dimensions (see: section 5.3), they can be placed in a parameter space that might as well make their distance quantifiable.

However, the crucial question to ask at this point is whether the realization and realized relationship works in the same way as the relationship between determinable and determinate colours. If it doesn't, then the assumption that realizer states like P11 and P12 are close neighbours in the similarity space seems to be groundless.

Yablo's work on mental causation had a different focus then Menzies', but his examples and approach to exclusion probably was an important influence for Menzies (see: Menzies 2003, 2007, 2008). Arguing for higher-level autonomy Yablo (1992) proposed that the reason why we see those counterfactual patterns that Menzies identified on independent grounds is the determinable-determinate relation itself. His aim was to bring more clarity into the discussion concerning realization and causal exclusion at the same time and came up with the influential idea according to which the realization-realized relation is a case of the

determinable-determinate relation. So, for Yablo, P11 and M stand in the exact same kind of relation as scarlet and red, the determinable-determinate relation that also provides the explanatory reason behind the counterfactual patterns we observe in the case of multiply realized higher-level causes.

Menzies et al. aimed to disconnect themselves from Yablo's approach, but they wanted to retain the essence of Yablo's project at the same time (see: Menzies & List 2009:480, Menzies 2008). Instead of the determinable-determinate relationship as the grounding idea, they suggested the proportionality of causation as an autonomous principle that can perform all the useful things Yablo could achieve in terms of higher-level causal autonomy. The advantage of this new approach seemed to be that it delivers everything necessary, but without the metaphysical baggage that comes with relying on the determinable-determinate relationship as an interpretation of the realization relation.

Now, in itself it is fair that Menzies rejects Yablo's interpretation of the realization relation. If the properties involved in the causal statements relevant for the reformulated exclusion argument are all determinables and determinates as red and its different shades, then the same similarity relations apply to them even if it is not true in general that the realization relation is a case of the determinable-determinate relation. However, this is highly unlikely as realized and realizer properties fall to different property kinds and so realizer properties cannot be the determinates of the properties they realize (for further discussion see section 5.4). If this option is not available, we have to think through the consequences that follow with respect to the evaluation of lower-level counterfactuals (iii. and especially iv. above).

What I would like to show in the following section is that, even though he rejected Yablo's approach, Menzies was not successful in divorcing his reformulated exclusion

principle from intuitions suggested by the determinable-determinate relation. The best explanation of the way he and List handles inter-level relations between properties is that tacitly they are motivated by an analogy between the determinable-determinate relation and the realization relation. Reliance on this analogy is required for the introduction of the guiding examples that involve determinable-determinate colours. However, applying the logic distilled from and motivated by these examples to cases where the determinable-determinate relation does not hold between the relevant higher and lower-level properties is unwarranted. To evaluate the relevant lower-level counterfactuals, one needs to decide what resides in the closest possible world without the actual realizer and as we saw that depends on how similar one realizer property is to another. Let me highlight, that without the help of the determinable-determinate relation there is no a priori guarantee that the realizer properties are closely similar in the same way as in the case of the shades of a colour.

The project I defined above is motivated by the question Shapiro (2012:15) posed, however for him this remained a side-issue, my aim is to dig far deeper in that direction. Shapiro's strategy was to question the most important starting point of Menzies' exclusion argument, multiple realizability itself. He focused on interpretational problems of popular examples of realization from neuroscience research used by many advocates of causal autonomy (Menzies and List 2009, 2010; Menzies 2013; Raatikainen 2010; Woodward 2008, 2015)[91]. He tried to show that the way these authors tacitly conceptualized realization drawing on those examples is implausible and led them to the false belief that they have a

---

[91] There is an important difference between Woodward and the others in the list. Woodward does not argue for causal exclusion. Rather, his approach shows which causes (higher or lower-level) are more informative. His position is consistent with both higher and lower-level properties being causes.

proper example of multiple realization at hand[92], while in fact the realized and realizer properties are identical. Denying the starting points of an argument, in this case the Distinctness premise in the exclusion argument is a way to go, but not the strongest possible approach. In the present context this is especially true because Kim's exclusion argument also accepted the basic non-reductivist starting points. If those are rejected the debate about higher-level causal autonomy becomes obsolete.

Regardless of what one thinks about Shapiro's approach to the exclusion problem, the nature of the realization relation definitely plays a role in this controversy, so I will devote a lot of attention to it in the rest of this chapter. For now, let me explain how Menzies might have been led astray by his reliance on the old examples from Yablo and by some assumptions concerning the similarity of properties in the evaluation of counterfactuals.

---

[92] This result only follows if one accepts Shapiro's theory of multiple realization (Shapiro 2000, 2004; Polger & Shapiro 2016).

5.2   Downwards exclusion and the independence of lower-level causation

Let's take a closer look at an example from neuroscience preferred by Menzies and List (2009) and Woodward (2008). They like to refer to research done on behaviour prediction based on modern brain imaging methods. Musallam el al. (2004) searched for predictors of the reaching behaviour of monkeys based on neural firing patterns in their brains and they found some reliable ones for certain behavioural outcomes. According to their paper, they could identify an important part of the neural basis of the intention of a monkey to reach out for an object. They established a correlation between the occurrences of a certain neural firing pattern in some brain region and the behavioural outcome which amounted to reaching for objects in front of the monkey. The animals could choose from eight different objects. The correlation was strong enough to form a base for probabilistic predictions under this very narrow range of circumstances. The accuracy was 67.5%. So, neural level observations proved to be an effective basis for prediction.

According to the interpretation Menzies & List (2010) provided for the example just introduced, there are many different possible brain states or particular neural firing cascades (P11, P12, … P1n) realizing the same intention of the monkey (M). By their lights it is fair to say that this is a good example of multiple realization, and we can run the reformulated exclusion argument to check which is the right level of causation. Applying the already familiar scheme the result should be the exclusion of the lower-level realizer:

i.   Monkey has intention M □→ monkey performs B

ii.   Monkey doesn't have intention M □→ monkey doesn't perform B

iii.   Monkey has P11 □→ monkey performs B

iv.   Monkey doesn't have P11 □→ monkey doesn't perform B

Here we see the already familiar pattern. Because the last counterfactual, where P11 is absent, is false we should conclude that the cause of the behaviour is not the realizer, but the intention of the monkey. But do we have good grounds to assume that the closest possible world without P11 has some other realizer in place? Menzies & List seem to find this assumption as natural as it was in the case of the colours shown to Sophie the pigeon:

"*Assuming that the closest worlds in which the monkey doesn't have neural property P11 are ones in which it has another neural property realizing the intention M*, one can see that the second counterfactual is false..." (Menzies & List 2009:488, *my emphasis; I changed the notation for properties to match mine*)

But why should we believe this assumption? Shapiro (2012:15) asks basically the same question, and he thinks that there is no answer to it in Menzies' papers on the topic (Menzies 2008, Menzies and List 2009, 2010). However, there is one explicit answer in the last section of Menzies & List (2009). This answer has nothing to do with the excision standard I talked about in the last section. It is based on the concept of realization sensitivity. The next two subsections reconstruct this answer and find it to be unsatisfactory. In my view, if this answer is not viable, then Menzies' project faces formidable difficulties when it comes to the exclusion of lower-level causes by higher-level causes. If this answer is not available, the only guiding principle one can rely on to argue for the closeness of different realizer properties in the similarity space is the determinable-determinate relation which Menzies rejected form the get-go.

5.2.1 The sensitivity of causal relations

To provide an understanding of why Menzies thinks that it is safe to assume that closest worlds where one realizer of a higher-level property is absent there is another realizer in place, I should clarify the notion of realization sensitivity. Let us put some more meat on the idea by tracing it back to its origin in Woodward's discussion of causal invariance (Woodward 2003) and sensitivity of causal relations (Woodward 2006, 2010)[93]. Woodward's interventionist theory of causation differs from the counterfactual theory utilized by Menzies in many interesting respects. I will only highlight one relevant aspect here and leave out many of the technical details that would be required for a deeper understanding of the account.

For the interventionist, causation is a relation between variables. Variables in a special case can have binary values, like the presence of an event (1) and the non-occurrence of the same event (0). A causal relation holds between these binary variables if introducing a change into the value of the putative cause variable (from the presence of M (1) to its absence M (0), in our case) by an intervention, one can change the value of the putative effect variable (from the presence of B (1) to its absence B (0), in this case). There is a causal relation between the variables (or the events referred to by the values) if such a change related correlation holds true[94].

---

[93] Woodward and other authors run the same basic idea under different names such as: stability, non-contingency of association, insensitivity and invariance.

[94] I presented an oversimplified version of the theory similar to the ones that were utilized by Menzies on several occasions (see: Menzies 2007, 2008). This simplification brings it so close to the Menzies and List account that at this limit the latter theory can be treated as a special case of Woodward's theory. However, Woodward's theory of causation is sufficiently different from the account used by Menzies to deliver a different result. Applied to the exclusion problem it calls into question the whole idea of causal competition between levels, as for Woodward proportionality is not a constraint on causation, therefore any kind of influence between causal variables is deemed to be sufficient for a causal relation to hold. In his view, it might be that higher-level causal

But there is more to the relation than this. A causal relation is characterized by an
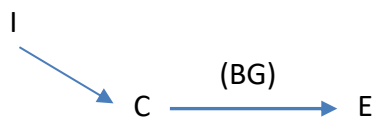

Figure 5.2-1

intervention variable I, a cause variable C (where the relation between I and E is assumed to be causal), an effect variable E, and last but not least a variable BG representing the set of background conditions under which the causal relation holds (see Figure 5.2-1). Interventionism provides a good general depiction of experimental practices in the sciences. Most causal relations are established in specific experimental conditions, in isolated systems, in a carefully controlled environment. On the one hand, this aims to ensure that the interventions on the cause variable don't influence the putative effect variable via any other routes than the direct route between I and C. On the other hand, finding a causal relation under such special circumstances does not ensure that it holds under other circumstances as well, so it is always a further task for research to discover and to measure the scope of the causal relation: to determine the conditions under which it holds outside the laboratory. BG expresses the full set of empirical conditions under which the causal relation, the change related covariation between C and E holds invariably.

So, the invariance of a causal relation is bound to a set of background conditions. This dependence on a particular set of conditions explains the sensitivity of causal relations. A causal relation is considered to be sensitive if it holds only under a narrow range of background circumstances. It is considered to be insensitive if it holds under a broad range of background circumstances. Woodward claims that sensitivity is an important factor when it comes to evaluating causal statements. It might be that a pair of counterfactuals are true:

1a. $C \square \rightarrow E$

---

statements provide us with better explanations, but that is not a reason to exclude their lower-level pairs from causal efficacy (see: Woodward 2008).

1b. ¬C □→ ¬E

However, we might still be reluctant to accept the claim that C causes E. This might be because the causal relation is highly sensitive to background conditions: holds only under a narrow BG. Menzies, following Woodward, emphasizes the importance of sensitivity in the case of counterfactual 1a: the presence condition. If the presence of the cause is in a fragile relationship with the presence of the effect, ordinary speakers are more reluctant to accept a causal statement to be true. This shows that sensitivity plays a role in our causal judgements. Menzies & List borrowed an example for sensitive causation from Woodward:

> "Lewis writes a strong letter of recommendation that causes X to get a job he would not have got otherwise, which in turn causes Y, who would have gotten the job in the absence of Lewis's letter, to take a job in a distant city, where she meets and marries a person; she has children with this person, and these children in turn have children, and so on. Call one of Y's descendants N. […] the causal statement 'Writing his letter of recommendation caused N's death' is very sensitive." (Menzies & List 2009:498)

There is no question whether counterfactual 1a. is true. If Lewis hadn't written that letter of recommendation, Y would not have married her husband and her descendants would not have died later. However, the cause and the effect are connected via a long chain of events and there are a great many not too far-fetched possible changes to the actual circumstances that would disrupt the causal relation between the strong letter of recommendation and the death of the late descendant N.

As an example for insensitive causation they describe the following scenario:

"[…] shooting someone at point blank range with a large caliber bullet: as we vary the background circumstances in ways that are not too unlikely or far-fetched, it continues to be true shooter had fired the bullet, the victim would have died." (Menzies & List 2009:497)

The presence counterfactual (1a.) is true for this case as well. But here, in a typical case, it would take serious departure from actual the situation where someone shot at a close range to get to a situation where the bullet does not kill the victim. This is why we don't hesitate to accept the statement that the shooting caused the death of the victim.

5.2.2   Realization insensitivity and the content of closest worlds

Now that the stage is set, and we understand the basics of causal sensitivity we can turn to the notion of realization sensitivity. Remember, Menzies claims that realization sensitivity provides an answer to the question: what resides in the closest possible world where the actual realizer state is absent. This is where things start to get interesting.

Menzies (along with Woodward 2008) understands realization sensitivity to be a case of causal sensitivity on the assumption that changes in the realization of a property are like changes in the background circumstances[95]. Furthermore, he presupposes that in the case of higher-level sciences we only accept causal relations that are robust and realization insensitive. So, the starting point in answering our question is not that there is a similarity measure[96] that makes it natural to suppose that closest non-actual-realizer worlds contain other realizers. It is that special-science generalizations wouldn't be accepted if they weren't

---

[95] However, it is worth nothing here that insensitivity to background conditions and realization insensitivity might occur more or less independently of each other. It might be that the causal relation between C and E is highly sensitive to background conditions but insensitive to changes in the realization of C.

[96] Such measure could be based on the determinable-determinate relation or some other relation available.

realization insensitive and it is this latter feature that tells us that realizers are immediate neighbours in the similarity space.

According to the definition of downward exclusion the closest possible realized property (M) permitting sphere is populated with relevantly similar worlds, with worlds where the background circumstances are the same as in the actual world. So, the variation among the worlds on that sphere are only with respect to realizer properties (P11, P12, …, P1n). If under fixed background conditions it is true that whichever realizer of M occurs the effect (B) also occurs, then the relation between M and B is realization insensitive.

So, the idea seems to be this. We know what resides in the closest possible world where a particular realizer property is absent because the relevant higher-level causal relation is realization insensitive. To see that this is really how Menzies and List argue and how this argument is supposed to work let us take a look at two passages from their text:

> "[…] in arguing that the target's being crimson is not a difference-making cause of the pigeon's pecking, we claimed that the counterfactual 'The target is not crimson □→ the pigeon does not peck' is false because if the target had not been crimson it would have been red, […]. *But this is analogous to assuming a similarity relation according to which some of the closest ¬P11-worlds are M-worlds* in which B is present." (Menzies & List 2009:496, *my emphasis; I changed the notation to match mine*)

The first thing to note is that the quote admits taking the case of determinable-determinate colours and the realization of mental states by brain states to be analogous. As I already said earlier, if the analogy holds it provides a good reason to think that all closest possible ¬P11 worlds are worlds with some other realizer of M. But to maintain the analogy

Menzies needs further non-trivial arguments for it as he rejects the view the M and P11 stand in the determinable-determinate relation. In the paragraph immediately following the last quote we get the argument:

"Why is this assumption so easy to make? Why is it so natural to assume that when there is a higher-level causal relation, it is realization-insensitive? One reason is that we intuitively require difference-making causal relations to hold in various possible situations, not just in the actual one." (Menzies and List 2009:497)

The second sentence implies that the reason why different realizers are immediate neighbours in the similarity space is the realization insensitivity of that higher-level causal statement. This argument is a non-sequitur. The statement it tries to prove is that it is natural to handle cases of realization analogously to the case of determinable-determinate colours. The first premise it relies on is that the mentioned cases can be handled analogously because higher-level causal relations are realization-insensitive. From that premise it does follow that higher-level relations hold in various possible situations and this is what should prove the point. However, this is not enough for that purpose. Higher-level causal relations might hold in various possible situations without it being true that some closest possible non-actual realizer worlds contain other possible realizers. To convince us that realizers should be closely similar to each other further argument would be required.

In the case of determinable-determinate colours we are provided with a well-structured similarity space for e.g. shades of red or any other colour[97]. In that context it is straightforward that certain determinate colours falling under a determinable are close neighbours in the similarity space. Consequently, if a determinable-level causal relation is

---

[97] We will see this in more detail later in section 5.3.

realization-insensitive, and holds in various possible situations in that sense, the determinates (realizers) of that determinable (realized) will be close neighbours. However, nothing guarantees this if realizer properties don't fall under a common determinable.

In simple counterfactual theories, and Menzies' theory is not an exception in this respect, the similarity space is always structured around what we take to be a cause under some description in the actual world, under the actual circumstances. My *pushing the "on" button* turns on the water boiler. My pushing the "on" button *slowly* (rather then quickly) does not affect that. The closest world that is similar in other relevant respects, but where the button is pushed with a normal speed, or even more slowly than it was pushed in the actual world, still has the water boiler turned on. The second description picked out the wrong property, but the similarity space was structured by the description. And in both cases, I relied on two things to evaluate the relevant counterfactuals. First, the property description highlighted and second, the excision standard to decide what is in the closest non-occurrence world.

What makes the Menzies' theory peculiar is that it allows for more than one equally close world with the presence of the putative cause where some variation in how the cause is present is acceptable (see: section 4.2.2). That is the assumption that allows higher-level causal statements to gather together all realizer properties on the closest higher-level property permitting sphere around the actual world. But the fact that they are brought together by this formal procedure to check whether the putative higher-level cause is sufficient for its effect or not has nothing to do with the closeness of realizer properties to each other on their own terms. Therefore, the same semantics might surprize us when applied to lower-level statements independently of higher-level causal statements.

So, Menzies might be right to:

"suggest that in the special sciences higher-level causal relations are typically required to be invariant under changes to the way in which higher level properties are physically realized." (Menzies & List 2009:499 my emphasis)

However, this is not a reason to suppose that in the closest worlds without one realizer property another realizer property is present. At least not on grounds of the semantics Menzies relies on. In other words, higher-level causal statements might well be typically robust in the sense that they are realization insensitive, but from that nothing straightforward follows for the causal status of the relevant lower-level causal statements. Their causal status requires independent evaluation. So, Menzies and List does not have a reason to continue like this:

"[…] If it is correct that realization-insensitivity is a general requirement in higher-level causal claims, then it follows that the conditions for downwards exclusion are generally satisfied." (Menzies & List 2009:499)

This quote implies that for Menzies it should be true that, what resides in the closest worlds with the non-occurrence of the actual realizer is determined by how the relevant higher-level causal statement[98] structures the similarity space for itself. This assumption is surely false. The realizer states brought together on the smallest M permitting sphere can be far away from each other in the similarity space if it is structured on their own terms. What is more, it is plausible that they are far away from each other as multiple realization standardly conceived implies that the different realizers fall under different kinds in lower-level science terms. If they are like Fodor's example of money which can be realized by tokens that have

---

[98] The antecedent of which refers to a putative higher-level cause realized by the putative cause property in the relevant lower-level causal statement.

nothing interesting in common in a physical sense, then clearly those kinds don't have much in common, so there is no reason to think that some commonality would hold them close to each other in the similarity space.

This section has shown that Menzies and List have no viable argument for maintaining that the relationship between realized and realizer properties in general can be handled analogously to the determinable-determinate relation. If that is so, then to decide whether downwards exclusion takes place in any particular case one has to attend directly to the similarity relations between the different possible realisers.

5.3   Proportionate causation and determinable vs. determinate properties

Let us review the argument so far. The last section shown that the only available measure for Menzies to decide what resides in the closest possible world where a putative lower-level cause does not occur is the excision standard. If that is the case, the downwards exclusion argument requires further backing to persuade us that the different possible realizer states for a higher-level property are surrounded by other possible realizer states as immediate neighbours in the similarity space. However, if it turns out that the determinable-determinate relation holds between relevant lower and higher-level properties that might be enough to secure that the similarity space is structured in this manner.

The question I will investigate in the following two sections is whether the determinable-determinate relation would really be enough to secure exclusion in the way expected. Or in other words, is it true in all possible cases that the determinable-determinate relation delivers the kind of similarity measures the truth of lower and higher-level counterfactuals relevant for downward exclusion scenarios require. My answer is that the determinable-determinate relation delivers on its promise.

The next question that I will address in section 5.4 is whether the realizer-realized relation is a version of determinable-determinate relation. Agreeing with Funkhouser (2006, 2014), Menzies (2008) and providing further arguments, I will reject the analogy. But rejecting the analogy has a price one has to pay. A price Menzies didn't realize, when he rejected the analogy. Without the determinable determinate relation to back up the similarity measures for counterfactuals one needs to find different ways of measuring similarity and that leads to unexpected complications and unwanted consequences.

5.3.1 Determinables, determinates and the determination space model

Let's dive deeper into how the determinable-determinate relation works using the traditionally favoured example of colours (see: Yablo 1992; Funkhouser 2006, 2014; Wilson 2009). In most theories of colour colours have a small group of common features or as in the philosophical literature Funkhouser (2006, 2014) started to call them, determination dimensions. The models of colour built around these features serve important practical purposes. One simple purpose is to identify specific colours in a precise manner by the help of a few parameters. This has many applications e.g. in printing, in the construction of different kinds of screens, monitors or TVs. Such models are also capable of providing guidelines to measure the closeness of different colours to each other and therefore can help us to find the practically less, but philosophically more important closest possible non-crimson or non-scarlet worlds.

The notion of a determination dimension was first introduced by Funkhouser (2006). Since it allows for more clarity than previous approaches to the determinable-determinate
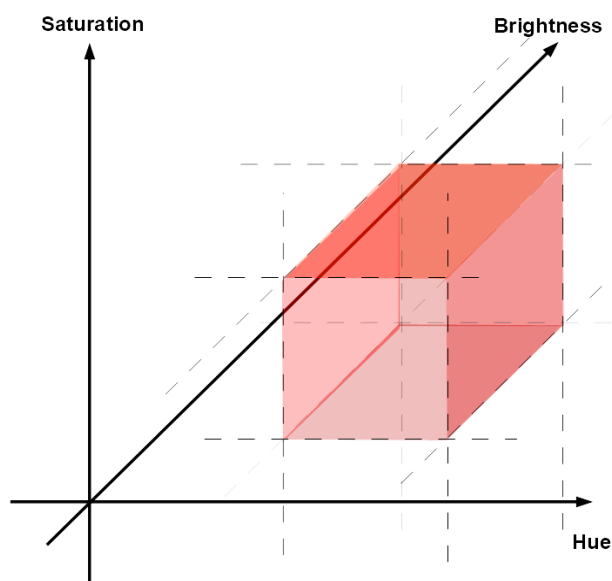
Figure 5.3-1

relation and was accepted by many recent participants in the debate concerning the uses of this relation in analysing realization (e.g.: Wilson 2009) in what follows, I will rely on this model. Let me spell out the example of colours in more detail in terms of this model. One prominent theory of colour, the so-called HSL model, utilizes three determination dimensions: hue, saturation and brightness/lightness. As we have three dimensions in which values can be assigned, any colour can be described as a well-defined

region in a 3-dimenional coordinate system (see Figure 5.3-1). Ideally such a region would have a cuboid shape. This is not necessarily the case, but for simplicity's sake I will proceed as if it were. All points of one such cuboid belong to the category of red. A region as broad as red is called a determinable property. Nonetheless, it is possible to formulate even broader characterizations, like having a bright colour, that would also deserve the name determinable property.

Narrower regions inside a determinable property region, shades of red like scarlet, are usually called determinates of red even though these are properties that can be determined further. Compared to red, scarlet is a more determinate property and it is a determinate of it. However, in an absolute sense it is still a determinable. The reason why these relatively determinate colours are usually called determinate is that these categories are the finest we have for colours in natural languages.

We can also talk about fully determined colour instances. According to many theorists' determinable properties are abstractions, what really exist in the world, or in my visual field where I see something as red is a particular colour falling under the abstract category of red (see e.g.: Gillett & Rivers 2005, Crane 2008). That property corresponds to a point in the space defined by the determination dimensions for colours. A point in the space is a superdeterminate property that cannot be determined further.

On Funkhouser's construal an "empty" determination space defines a property kind like colour. Anything defined by a restriction in that space is a property. The determination dimensions that characterize a property also provide the definition of the essence of the property kind in question. So, the essence of colour is that it has hue, saturation and brightness. If two properties have different determination dimensions, then they have different essences and therefore they fall info different kinds.

So, properties can be represented and are determined in determination dimensions. It is easy to see that the relation of determination dimensions to properties is asymmetric, properties can be determined in the dimensions but not vice versa, and irreflexive, dimensions determine properties but not themselves.

In accordance with this concept, e.g. triangles are determined in determination dimensions such as the three dimensions for side lengths. However, certain features of triangles in the physical world, such as their mass, transparency, constitution or colour are not their determination dimensions because, as Funkhouser argues, we do not distinguish triangles along these features. The same applies to colours as well.

There are five criteria Funkhouser provides to identify determination dimensions. Each of these serve as independent necessary conditions to distinguish determination dimensions from properties.

The first (1) criterion is that we should be able to provide a unique, unambiguous identification of properties that fall under a property kind in terms of the parameter values in the determination dimensions (Funkhouser 2006, 553). Let's unpack this last criterion, because it is non-trivial.

Apparently, the features that distinguish different triangles or colours according to certain standards count as determination dimensions. However, some components in terms of which certain properties can be analysed do not count as determination dimensions for those properties, even though they might seem to fulfil the other criteria with the possible exception of the fourth one which, as we will see, is only a half-hearted suggestion. Funkhouser's example is momentum which might seem to be determined by two dimensions:

velocity and mass[99]. Contrary to first appearance, velocity and mass are not determination dimensions of momentum. The reason is that objects with different velocity and mass can have the same momentum (Funkhouser 2006:553). This creates a problem as one cannot provide unique identification for a particular momentum property in terms of mass and velocity. Infinitely many different mass and velocity pairs might result in the same momentum value. Consequently, the momentum of an object as a physical quantity results from its mass and velocity, however, is not determined by mass and velocity as a determinable. Momentum has component properties, but those are not its determination dimensions. It has its own independent determination dimensions[100], its magnitude and direction.

According to the second criterion (2), every property should have a limited number of determination dimensions. Third (3), the dimensions should be independent (Funkhouser 2014:23) which implies that their values are uncorrelated. For some reason criteria (2) and (3) are not put to serious theoretical work neither in Funkhouser's 2006 article nor in his 2014 book. Here I would like to emphasize the importance of criterion (3), going back to the example of colours again, as it will become important for my argument later.

In the case of the HSL colour model, the three parameters in the colour determination space are, as in geometry they would say, orthogonal to each other. This means that the

---

[99] Here I am simply following Funkhouser. However, it is important to note that momentum is a vector quantity, so even though in many standard textbook calculations the direction of momenta is disregarded for certain purposes, it does have a direction in 3-dimensional space.

[100] In Funkhouser's rendition (see previous note) the determination space along which momentum is determined has only one constituent (e.g.: a meters per second scale), therefore it counts as simple determinable. One should note here that in his discussion Funkhouser disregards the direction of momentum. So, even though I accept his argument against taking mass and velocity as determination dimensions, I think, he is not justified in saying that momentum is a single determinable. Momentum should be taken to have two determination dimensions: its magnitude and its direction.

parameters are independent[101]. This implies that one can freely change the parameter value in one dimension without changing the values of the remaining parameters. Any combination of possible values in the different determination dimensions is possible and results in the proper description of a particular property. Combined with criterion (1) this results in a requirement that makes the determination space model efficient and useful. It does not contain excess or redundant information, so it provides satisfactory identification and representation of all possible colour properties[102].

We should note however, that the independence of determination dimensions has certain interesting limits. Funkhouser accepts shape as a property with different determinations dimensions for any particular kind of shape property. Consider triangularity. Triangles have three sides that are closed. The length of those sides determines the exact shape of the triangle, so the three length parameters provide the determination space for triangular shapes. Now, there is a well-known theorem in Euclidean geometry called triangle inequality. It says that the sum of any two sides of a triangle must be greater than the length of the third side. If side lengths are the determination dimensions of triangularity then according to this theorem the dimensions cannot be independent of each other.

---

[101] This implies e.g. uncorrelatedness, but correlation tests only for linear relationships, which leaves space for certain other kind of dependencies. So, independence is a stronger notion, at least in mathematics. The independence of the determination dimensions can be expressed in terms of the orthogonality of any one of the three features to the others: Hue $\perp$ Brightness & Saturation, Hue & Brightness $\perp$ Saturation, etc.

[102] The requirement of independence or orthogonality of parameters is highly important for scientific model building in general. Funkhouser gives it less emphasis than I do here in my discussion, he only mentions it in passing (Funkhouser 2006:551), but as he refers to scientific practice as a justification for his model of determination spaces and determination dimensions, I think it is fair and important to focus more light on this requirement.

Here it is important to recognize the importance of a distinction Funkhouser makes. To individuate properties, one needs two components: determination dimensions and non-determinable necessities (Funkhouser 2014:38). In the case of triangularity, both points and subregions in the determination space, so both particular triangles and determinable types must conform to the requirement of having three sides that form a closed plane figure. This requirement expresses the relevant non-determinable necessities. These are traits that each instance of a property kind must have. They are called necessities because they do not allow for variation within that property kind (see: Funkhouser 2014:37). In other words, determinate properties under a certain determinable cannot differ with respect to non-determinable necessities. Determinates under a determinable have exact similarity in those terms. So, triangles cannot differ with respect to triangle inequality as the theorem follows from the non-determinable necessities for triangles.

Even though Funkhouser nowhere endeavours to do this, it seems to me that he could easily make a useful distinction differentiating dependency relations between the determination dimensions explained by non-determinable necessities and dependency relations that are not explained by non-determinable necessities. I think, by the help of this distinction, we can make the independence criterion more precise. The determination space of a property might be restricted by non-determinable necessities if the definition of a property kind requires them. In such cases, usually there is some kind of a limited dependence between the determination dimensions. Dependencies between the dimensions are only acceptable if the attached non-determinable necessities are providing a clear explanation for it. Besides such kind of dependencies, the determination dimensions should be independent.

The fourth (4) criterion says that determination dimensions need to be inseparable in the sense that they cannot be manifested in separation (Funkhouser 2006:552). For example,

it is impossible to manifest hue, without manifesting saturation as well or to manifest pitch and timbre without some level of loudness at the same time[103]. There are also features that can go together, but don't need to go together. Like shape and rebound hardness. E.g. a body of liquid water might have a certain shape without manifesting rebound hardness as the latter kind of property is only manifested by solids. To sum up, it is a necessary condition for being a determination dimension that the features in question are inseparable, cannot be manifested in the absence of the others.

According to the fifth (5) criterion, determination dimensions in themselves cannot figure in, cannot make unique contributions to causal laws. At the same time, for properties determined in terms of such dimensions this is a matter of necessity as this is what legitimates a property as a property[104]. On the basis of criterion (5) Funkhouser provides general advice concerning the epistemology of property kinds. The advice is to look for terms that are necessary to formulate successful laws and explanations as that should lead to the discovery of proper property kinds.

---

[103] I have doubts concerning the efficacy of this criterion. The theory provided by Funkhouser aims to differentiate properties from determination dimensions and arbitrary constructs like the property of colour-shape. Colour and shape seem to be as inseparable as hue and saturation. Anything that has a colour, has a shape as well. The opposite is questionable, it depends on how one defines colour. If black is not considered to be a colour, we might not see the colour of a thing under certain circumstances, but if the function of colour perception is to track the surface reflectance traits of objects, and colour itself is defined as a surface reflectance property then it is fair to say that black is a colour. If this is true, then shape and colour are inseparable. Criteria (1-3) are also applicable to shape-colour. Shape and colour provide unique identification for shape-colour. There are a limited number of determination dimensions. Shape and colour are independent in the required sense. So, I am not sure that this criterion can do real work in differentiating properties from determination dimensions even when combined with other criteria.

[104] At least according to certain sparse theories of properties Funkhouser seems to favor (see his: 2006:552 and footnote 17).

The fifth criterion is only formulated as a suggestion on Funkhouser's side, he recognizes it as being problematic (Funkhouser 2006:553). E.g. Dretske have provided a plausible counterexample to this idea (see: Funkhouser 2014:21). Pitch, loudness and timbre seem to be good candidate determination dimensions for sound. However, in Dretske's interpretation, according to material science the pitch of a sound alone can bring about the shattering of a glass. So, in light of our criterion pitch should be treated as a separate property instead of a determination dimension.

I think, the example of pitch described above is slightly misrepresented. Without the sound being loud enough even the right pitch is insufficient to trigger the effect. In the case of a standard wine glass, if loudness is below around a 100 dB, the pitch has no effect. So, there are at least two conditions that need to be satisfied to break a glass: one has to find the right pitch and the sound should be sufficiently loud. Therefore, pitch has no contribution to the effect without loudness, but it still seems to have a unique contribution to a causal law.

Funkhouser relegates final judgement with respect to the difference between determination dimensions and properties to science[105]. He claims that the methods of philosophy are ill-suited for making such decisions, so all property descriptions in terms of determination dimensions are tentative and can be modified or falsified in the future. This means that the legitimate application of criterion (5) always requires reference to scientific authority.

Now that we have a solid understanding of determination dimensions, we can go back to the example of colours to analyse the determinable-determination relation further. Even

---

[105] As he says: "Perhaps the scientist can explain why the determination dimensions of certain complex determinables must co-occur, but this seems to be a problem ill-suited for the methods of philosophy." (Funkhouser 2014:22)

though there is no conceptual reason why regions corresponding to different colours could not overlap in the determination space, standardized colour names refer to regions mutually excluding each other. However, superdeterminate colours belonging to the same property kind necessarily exclude each other, as concrete colours occupy one point in the determination space. A point in the space might belong to the cuboid region of crimson and that cuboid is contained by – is a proper subset of - the cuboid region of red.

Basically, everybody agrees on one thing: it is always true that determinate properties are more specific versions of their own determinable properties. There is a ladder of increased specificity between e.g. warm colours and tomato red. This observation forms the basis for the best general and widely used characterization of the determinable-determinate relation:

Increased specificity: Property P determines property Q iff for something to be P is

for it to be Q, in a specific way.[106]

What Funkhouser's model adds to this definition is that the specificity of properties always increases in respect of a particular determinable, because those determinables, as in the case of colours, have common determination dimensions with their determinates. Based on the concept of determination dimensions and on Increased specificity a further principle can be formulated that makes the idea of increased specificity more precise. From this principle everything we already said about determinable and determinate colours follows, and also anything that can be said about determinable-determinate property pairs in general:

Qua principle: a determinate property P specifies a determinable property Q only

along the determination dimensions of Q.[107]

---

[106] For further discussion of the principle see Wilson (2009), I have adapted the formulation developed by her.

[107] For further discussion of this principle see: Wilson 2009; Funkhouser 2006, 2014

Funkhouser (2006:554, cf. 2014:40) defines the determination relation along three criteria. Property P determines property Q if (i) property P and property Q have the exact same determination dimensions, (ii) property P has the same non-determinable necessities as property Q, (iii) the property space of P is a proper subset of the property space of Q.

Funkhouser (2006) distilled one further principle from the Qua principle emphasizing that different same-level determinates of a determinable property Q cannot be the same with respect to their Q-ness:

Difference principle: distinct same-level determinates of a determinable Q differ along at least one determination dimension of Q.

Defined in these terms the determination relation has the potential to provide us with measures of the distance or closeness of properties that belong to the same property kind or fall under the same determinable. In a space defined by determination dimensions satisfactory measures of the distance or closeness of different properties can be defined as there is a coordinate system representing the relevant properties that provides proper guidance for decisions concerning distance (see: Figure 5.3-1 and Figure 5.3-2). The coordinate system defining the property kind might use different kinds of coordinates corresponding to different sets of possible states. If the sets are ordered[108] then the distance between the points, i.e. super-determinate properties, can be measured as Euclidean distance in the coordinate system of the determination-space. This is a standard measure in colour science. So, we can decide which of two colours is closer to the actual colour of

---

[108] They definitely are in the case of the determination dimensions of colours. However, this isn't clear in the case of mental states. It is reasonable to represent the intensity feature of mental states on a scale, but this is less straightforward when it comes to attitudes. If there is a natural ordering for attitudes then we can set up a scale from e.g. believing x to hating x. But, as far as I know, no such scale has been suggested by science yet.

something. Measuring the distance between determinable regions such as red and blue would require more complicated mathematical apparatus, I won't try to present a method, but it can plausibly be worked out as a similar measure in terms of something like the overall distance between the points of two contiguous regions. What is important, we can easily find the immediate neighbours of regions or points in the determination space. If the red region is divided into non-overlapping regions corresponding to shades of red it is easy to answer the question, what resides in the closest possible non-crimson world?

So, if the determinable-determinate relation is a good model for the relation between realized and realizer properties, then there seems to be no problem with answering our pressing question: what resides in the closest possible world where the actual realizer of a higher-level property is absent (¬P11)? (the iv. counterfactuals in the intro to section 4.3). As we saw the coordinate system of the determination space provides satisfactory measuring procedures for distance, so we should be able to determine which are the closest property regions to the region occupied by the property that is considered to be the cause in the actual world. Let us see how this works.

5.3.2 Determinables, determinates and lower-level causation

The first goal of this section is to check whether similarity measures based on the determinable-determinate relation really provide the expected results to the question what resides in the closest possible world without some putative cause property in all possible cases. As we saw, this is what is required to run the argument according to which multiply realized higher-level properties downwards exclude their realizers from causal efficacy. Below, tentatively, I will talk as if the realization relation would be strictly analogous to the determination relation, but only for the sake of the argument. The criterion Menzies and List put forward for downwards exclusion is reasonable accepting these assumptions[109]. Exclusion in the downwards direction is straightforward if the relevant counterfactuals are interpreted based on determinable-determinate property pairs. I will strengthen this statement by checking the status of some special, borderline cases of determinate-level counterfactuals[110] (versions of iv. in section 5.1-5.2), cases where some immediate neighbours of lower-level realizers are not realizers of the relevant higher-level property.

Let's go back to the determination space that defines colour properties. On Figure 5.3-2 next to the cuboid containing shades of red I have highlighted a neighbouring cuboid that contains shades of yellow. Moving along only on the hue axis this is the next region our everyday colour concepts single out as an interesting unit. There has to be a border between any two regions. It is signified as a plane that is cross-hatched on Figure 5.3-2 with two points on it signifying two maximally determinate properties. The plane itself and the points on it either belong to the red region or to the yellow region as there can be no points that don't correspond to a colour. Wherever it belongs, the border plane is a plane the points of which

---

[109] However, as I already said in section 4.3.2, the second subcondition expects less than what is really necessary.

[110] I am talking about versions of Yablo's classic example: target is not scarlet □→ pigeon does not peck

are such that some of their closest neighbours belong to the yellow region, some others belong to the red region[111]. If the point is situated somewhere inside, not on a surface of the red cuboid then any closest non-actual world has a different shade, or "realizer" of red in place. When a point, a particular "realizer" is situated on any surface of the cuboid, for example on the cross-hatched plane on Figure 5.3-2, then it has some neighbours belonging to a non-red region. On Figure 5.3-2 these are "realizers" of yellow. The same can be argued for cuboid regions inside the red cuboid sharing the same border with the red cuboid. Those regions also have yellow regions in their immediate neighbourhood. One can find some even stronger special cases in this model. If a point or a sub-cuboid is laid on one of the point corners of the red cuboid then there are multiple non-red neighbours available, probably more than red neighbours.
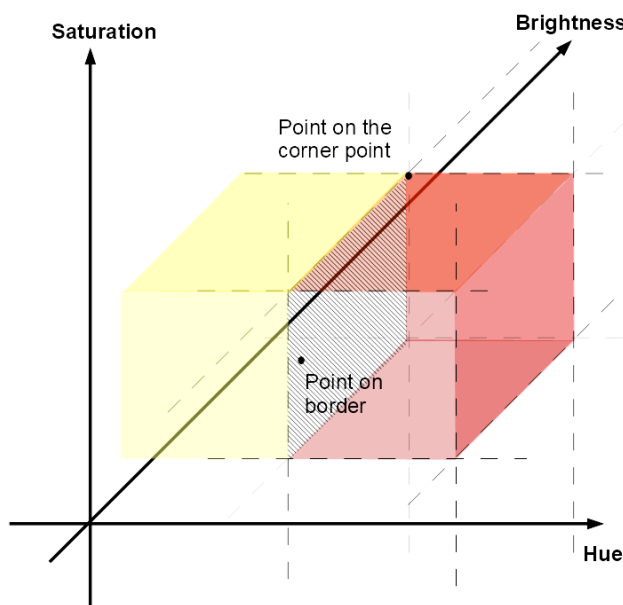
Figure 5.3-2

Remember that Menzies proposes a theory of counterfactual dependence in which there are worlds equally similar to each other[112]. As we saw, in section 4.2.2 there is a set of worlds equally close to actual world without the property referred to in the antecedent of the relevant counterfactual. If in the actual world some ordinary shade of red, such as

---

[111] Note that the exact shape of the relevant regions makes no difference with respect to the conclusion of my argument. The only thing that really counts is that there should be a surface between the two regions. The shape of the surface is only decisive with respect to the number of closest neighbours on the other side.

[112] In a theory of counterfactuals where every two worlds can be ordered with respect to their closeness to the actual world the situation discussed here would create problems. But we are not dealing with such a theory.

scarlet is present (implying that the property is somewhere inside, not on any surface of the red cuboid on Figure 5.3-2) then it only has red neighbours in the similarity space as all of the neighbours have red as a determinable. In that case counterfactual iii. below is true, whereas iv. is false, therefore the putative lower-level cause is not a cause.

   iii.   Target is x (x=certain shade of red) □→ pigeon pecks

   iv.   Target is not x (x=certain shade of red) □→ pigeon does not peck

Nevertheless, there are borderline cases we also need to consider checking the status of determinate-level counterfactuals. If there is at least one true lower-level counterfactual the downwards exclusion criterion fails. As we saw, in the case of some determinate colours, identified by the points on Figure 5.3-2, the actual world is such that there are immediate neighbours to the actual colour both with and without the same determinable. In those cases,
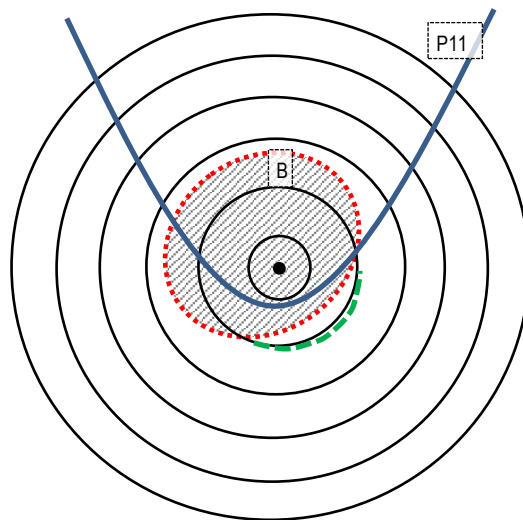
Figure 5.3-3

the status of counterfactual iv. is less straightforward, as in some closest worlds there is no pecking, while in others there is. However, this is not a problem for Menzies. On his semantics the determinate in question should be required for the effect to occur. So, if there is at least one closest world with the effect in place iv. comes out as false.

Figure 5.3-3 depicts such a situation in terms of possible world semantics using the notation introduced for the neuroscience research example in section 5.2. The P11 region is a region like red on Figure 5.3-2 and it shares a border with a region like yellow, in this case closest possible ¬P11 worlds. The B region signifies the behavioural outcome (the pecking of the pigeon or the reaching behaviour of the monkey in the neuroscience scenario). On the

diagram there are some ¬B worlds among the closest possible ¬P11 worlds situated on the right-hand side of the second sphere around the point signifying the actual world followed by a green dashed curve. It is neither covered by the cross-hatched B region nor is inside the P11 region. When the possibility space is structured in this way the second criterion for proportionate causation fails. Let us take a look at it again:

B. ¬P11 (¬red) □→ ¬B (¬pecking) is true in world w if and only if ¬B is true in all the ¬P11-worlds within the smallest ¬P11-permitting sphere of worlds around w.

For the reformulated exclusion argument to work when multiply realized properties are causes the relevant lower-level counterfactuals with reference to a realizer in the antecedent and reference to a common outcome property in the consequent should be false in general. That they should be false independently of which realizer of the multiply realized higher-level cause property is picked out. This condition is satisfied on the assumption that realization is strictly analogous with property determination. The analogy ascertains that realizers cannot be proportionate difference-makers of the higher-level outcome for which the higher-level property is a proportionate difference-maker.

To sum up, the second criterion of the proportionality constraint formulated for lower-level causal relations holds true for all possible cases of where the relevant higher-level property is a determinable of the realizer property in the antecedent of the relevant lower-level counterfactuals. Now that this question is settled, in the next section, I will marshal arguments against taking realization to be a version of the determination relation. After that I will investigate the consequences of rejecting the analogy.

5.4 Multiple realization is not the determinable-determinate relation

According to Yablo (1992), as Increased specificity (see section 5.3.1) is true for realized-realizer property pairs as well, they stand in the determinable-determinate relation and consequently that relation provides the best understanding of multiple realization and mental causation. Let us see why we should believe Yablo. There are at least three essential features that resulted from the analysis of the determinable-determinate relation in terms of Increased specificity that are relevant here (see: Wilson 2009 and Funkhouser 2014).

(D1) A determinable has multiple determinates: red has many shades; there are different, more specific ways of being red. Here lies the basis for the analogy with multiple-realization. A higher-level property has many realizers. A determinable has many determinates.

(D2) An object instantiating a determinate property must also instantiate a determinable property and vice versa. Higher-level properties must be realized by some lower-level property and lower-level realizer properties must realize a higher-level property.

(D3) Determination is asymmetric and irreflexive. The same is true of multiple realization. Only lower-level realizers can multiply realize a higher-level property, it doesn't work the other way around. And nothing can multiply realize itself.

(D4) Determinable and determinate properties are distinct at least as property types. Scarlet is a kind of red, scarlet is not identical to red. Also, realized and realizer properties are distinct, and if realization is a kind of property determination the idea that higher-level realized properties are distinct from their lower-level realizer properties is motivated, so the non-reductive physicalist seems to have the basis for running her standard argument for higher-level causal autonomy.

Because the relations share such features, people like Yablo and Wilson believe that there is a strict analogy between the realized-realizer relation and the determinable-determinate relation. If the analogy holds, the reformulated exclusion argument seems to work as intended.

However, if the analogy between determinable-determinate properties and realized-realizer properties breaks down, or if one aims to solve the exclusion problem without relying on this analogy, then measuring closeness, or distance in general in the similarity space that is required for evaluating counterfactuals is way less straightforward. In what follows, I would like to summarize the argument to the effect that there is an important disanalogy between the two relations, so one is forced to use independent measures of similarity.

5.4.1 The disanalogy between multiple realization and property determination

Funkhouser (2006, 2014) have argued effectively that it is not possible to interpret the connections between mental properties and their realizers based on the determinable-determinate property model. His argument against the idea proved to be successful and was part of the reason why Menzies was searching for an independent solution to the exclusion problem based solely on counterfactual patterns (see: Menzies 2007, 2008).

Let's see how the argument goes. Take a mental property, a belief realized by a neural network. The question is whether one can take a neurological state to be a more determinate propositional attitude. Or in line with the Qua principle: is it true that a neurological state specifies a determinable propositional attitude? It seems to be straightforward that mental properties are multiply realized, but are they determinables of their realizers?

Throughout the reconstruction below, I will follow Funkhouser (2006) in talking about beliefs, but my model of propositional attitudes for modelling beliefs will be slightly different.

Propositional attitudes are relied on both in philosophy and some strands of psychology like social psychology. In philosophy the notion has central importance in many fields (from philosophy of mind to philosophy of language), and it is a widely accepted model. The situation is similar in social psychology, although there are differences. In psychology the intensity of attitudes is taken to be an important dimension, philosophers are usually not interested in that. For the discussion below I am borrowing the intensity dimension from social psychology.[113]

A propositional attitude expresses the stance of a person towards some kind of a propositional content that has to be expressible in language. Propositional attitudes can be both conscious and subconscious. The simplest scheme of a propositional attitude is (x Fs that y) (see: Schiffer 1992, 1995), where x is a person, y is a proposition representing states of affairs that can be either true or false. F is x's attitude towards the propositional content. In psychology it can be any kind of attitude, positive or negative (belief, disbelief, doubt, fear, sadness). Philosophers usually talk about beliefs and desires. An attitude in psychology, e.g. a fear, has a level of intensity as well.
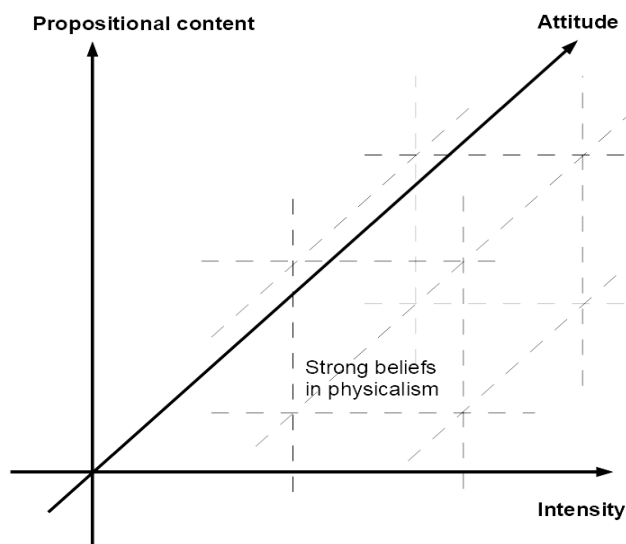
Figure 5.4-1

In the case of beliefs, they would talk about a level of confidence. Therefore, a propositional attitude as a mental state has at least three determination dimensions: propositional content,

---

[113] For more details on the psychological treatment see: Baumeister & Finkel 2010, chapter 6.

attitude and intensity. Knowing this one can easily situate for example "strong beliefs in physicalism" in the determination space provided by the three dimensions (see: Figure 5.4-1).

Although two of these determination dimensions of propositional attitudes are qualitative rather than quantitative, the resulting picture is akin to the case of colours. Theoretically, distance between particular propositional attitudes can be measured in a 3D coordinate system along the three determination dimensions. A broader kind of belief can be located in a cuboid with more determinate beliefs of that determinable falling into that region. For example, a "strong belief in non-reductive physicalism" belongs to the cuboid on Figure 5.4-1 "Strong belief in physicalism", it is contained by a cuboid inside the former cuboid (it is a proper subset of it) and it would count as a determinate of that broader mental property. Points inside both cuboid regions count as super-determinate propositional attitudes, e.g. an immovably strong belief in a special version of non-reductive physicalism.

If all this is so, do we have space for further determinates that could be considered as even more specified? It doesn't seem to be the case. But this is what we would need to be able to include the realizer of a determinate belief into the picture. But even if we start with a determinable, if a determinate property has to be situated inside the determination space of its determinable propositional attitude property then brain states described at the neural level provide no match for a more determinate property in the determination space of beliefs.
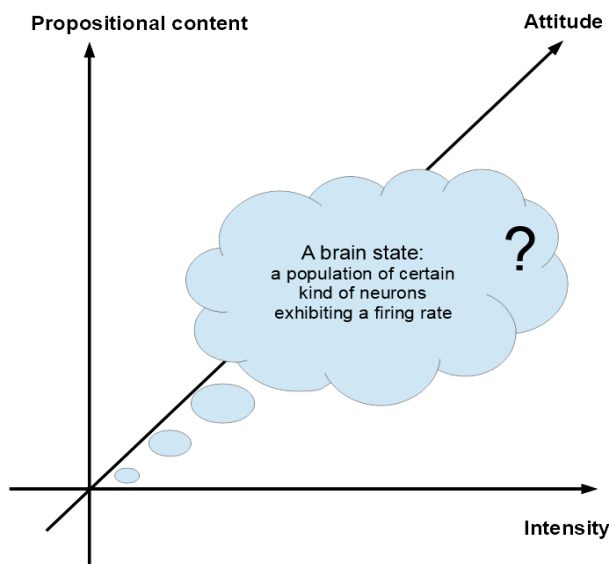
Figure 5.4-2

A brain state has at least two, but plausibly more, determination dimensions: the first concerns the type of neurons participating in a brain process and the second has to do with the firing rate of a population of such neurons. Obviously, the neural level description can be made way more sophisticated, but even if we count in more features it is implausible that those will match the determination dimensions of propositional attitudes. Higher-level mental properties and their neural realizers have distinct determinations dimensions in the sense that they don't share any determination dimensions (see: Figure 5.4-2). Therefore, it is not possible to situate neural realizer states in the determination space of mental properties. As Ehring puts it (1996:474): "physical realizers of the mental will not differ mentally at all, as they should if they are determinates of the requisite mental states". This amounts to the following argument:

- If multiple realization is determination, then different physical realizers of a mental property M should differ with respect to M.

- Different physical realizers of a mental property M do not differ with respect to M.

- Therefore: multiple realization cannot be modelled as determination.

For the determinable-determinate model to work different realizers of mental properties should differ mentally in the same way as all superdeterminate colours differ in respect of colour. Unfortunately, they don't. This becomes even clearer observing that the same super-determinate mental state could have different physical realizers. From this it follows, that realizers are not determinates of mental properties (Funkhouser 2006:565):

- Super-determinate mental property M can have different physical realizers.

- The different physical realizers are not different M-wise.

- Therefore: multiple realization cannot be modelled as determination.

If this conclusion holds then I can go a step further to argue that the tacit analogy with colours in Menzies' discussion of mental causation and realization is misleading. We know that Menzies aimed to provide a solution independent of the determinable-determinate relation as an interpretation of realization, so plausibly this wasn't what he intended to suggest. But as the example of colours is based on the determinable-determinate relation it does suggest that the closest worlds without the actual realizer are worlds with a different realizer of the same higher-level property. However, if realizers can be really diverse as many believers of multiple realizability would require, then there is no guarantee for that. The closest worlds to the one with the actual realizer might have content that is different from any realizers of the relevant realized property. Plausibly, this content would be something like a brain state that is not a realizer of the relevant mental property.

The reason why this could happen is that if we move away from clear cases of determinable-determinate properties to examples of multiple realization, then similarity relations in lower-level realizer terms are not channelled by higher-level sameness anymore. Or in other words, the determination space defined by the higher-level property kind cannot be used to measure the distances of the realizers. Realizers have to be described, considered along their own, independent determination relations divorced from the guidance provided by the higher-level realized properties.

If the possible realizer properties are diverse enough, as they can be if they do not belong to the same lower-level property spectrum or determination dimensions, their proper descriptions might plausibly exclude the kind of immediate closeness we saw in the case of different shades of red. Descriptions of lower-level events in terms of networks of neurons firing or in terms of computer chips allowing certain flows of electricity allows for unexpected

dimensions of resemblance relations. So, my concern is that it would be miraculous if we found a different realizer in the vicinity of the actual world for all possible cases of realization.

If this point holds water, then it is also possible that the truth value of the relevant lower-level causal statements based on lower-level similarity relations shifts with the context. It might be that in some cases, the closest non-actual realizer worlds have different realizers of M in place, in other cases they have nothing realizing M. So, in some cases the actual realizer is a cause of B, but in other, plausibly rare cases it is not a cause of B depending on the structure of the actual world.

However, detailed descriptions of neural networks and electric circuits are not easy to compare. Described in the right way, they might still be close to each other in the eyes of a Laplacian scientist. Therefore, a convincing argument requires more work. One of my aims in section 5.5 is to investigate cases where it is plausible that lower-level realizers are causes of their higher-level effects on Menzies's account. But before that let me try to answer an influential argument by Wilson (2009) that aims to save the idea that multiple realization should be understood as property determination. I don't think that a successful refutation of this argument is necessary to argue against Menzies, as he himself (Menzies 2008) rejected the idea that realization can be modelled on the determinable-determinate relation, therefore one could simply show the consequences of that commitment. Nonetheless, in recent literature this issue became a focus of discussion, so I think it is important to put the main options on the table and consider them more seriously. I don't think that I have a knock-down counterargument against interpreting realization as determination, but I do think that what I have to say shows that it is an unattractive option.

5.4.2 Metameric colours, determinates and realization

According to Wilson (2009), Funkhouser is too quick in concluding that mental properties don't have, or cannot have, neurological determination dimensions. She makes a case against Funkhouser's argument based on the example of metameric colours. The argument roughly goes like this. In the end, as Funkhouser also admits, science decides what counts as a property kind. The example of metamerism in colour science suggests that it is legitimate to characterize colours by extra information beyond hue, saturation and brightness in explicitly physical terms. Wilson interprets the kind-term colour as being science-relative. In traditional colour science it has three determination dimensions (see: Figure 5.3-1), but it has further determination dimensions in the context of physical colour science. She argues, by analogy, that it should be legitimate to talk about mental properties determined not only in terms of mental features, but also in terms of differences in realizer brain states.

First, to get to a basic understanding of metameric colours, let us go through the relevant facts of colour science. Certain pairs of materials seem to have identical colour under fixed lighting conditions, but they appear to have distinct colours under alternative lighting conditions. This is the so-called metameric failure. It occurs in situations where two material colour samples match when viewed under one light source, usually one that is close to bright white light, but not under another kind of light source. Fluorescent lights provide the strongest example of potential game changers (see: White & Jenkins 2001)[114]. Such pairs of colour samples are called metameric pairs.

The reason for the perceived shift in colour with changing background lighting conditions can be traced back to a difference in the so-called spectral power distribution (SPD)

---

[114] Another accessible and useful source of information on this topic is a series of Wikipedia entries starting with the obvious one: https://en.wikipedia.org/wiki/Metamerism_(color)

of the light reaching the human eye. SPD is the proportion of total light at visible wavelengths. It describes the total physical information of a bundle of visible light waves. The theory explaining how the human eye reacts to incoming physical light is called trichromacy theory. The human eye, because of its physiology, reduces SPD information to three sensory quantities, leaving out part of the physical information present in incoming light waves, which means that the colour space we experience represents only a subset of the information physically contained in the light reaching our retina. As a result, different SPDs might give rise to the same subjective colour experience. In textbooks on optics and colour science, they not only talk about metameric colour pairs, but also about metameric lights, lights that have different composition in terms of their SPDs, while giving rise to the same subjective colour experience (see: White & Jenkins 2001).

We already discussed one of the ways traditional colour science identifies colours in a 3-dimensional colour space (see: Figure 5.3-1). Now we can see that from the point of view of physical optics, the colours identified in the 3-dimensional colour space can be characterized further in terms of the SPD that causes the subjective colour experience. It is time to go through a more detailed explanation for the existence of so-called metameric colour pairs and metameric failures. Note, that by talking about metameric colour pairs we need to accept the supposition that colours are there to track certain traits of objects in the world. Characterising two colours as metameric pairs we say the following: there are two different objects in the world that cause the same colour perception under certain lighting conditions, but the very same objects might cause systematically divergent perceptions under other sets of conditions. What the perceived colour of an object aims to track on this view is the surface reflectance traits of the object, but in cases of metameric failure it fails to track

those traits properly. However, if we attend to the SPD information of incoming light waves, we have a means to track surface reflectance properties without such failure[115].

Let's see in more detail what happens in the case of metameric failure. First, we need to know that the light reaching our eyes when looking at an object originates from a light source and it is reflected by the surface of the object. Normally, we attribute colour properties to objects reflecting light from a light source like the sun or a light bulb. The light source emits light that has a characteristic SPD. The object selectively absorbs frequencies of the total light reaching its surface[116], while the remaining light gets reflected towards perceivers. The reflected light has a different SPD composition than the light that originated from the source as some of that energy has been absorbed by the object. It is possible that two objects under identical lighting conditions reflect metameric lights, lights that cause the same colour experience but have different SPDs. This is a case when one is faced with metameric colour pairs. By changing the light source but leaving the two objects untouched it might happen that we get different kinds of reflected lights that are not metamers of each other. This constitutes a case of metameric failure.

Wilson thinks that metamerism provides a case where two colours can differ without a difference in their subjective appearance in accordance with the difference principle. For Wilson (2009:163) the example implies two things. First, the determination dimensions of colours are science relative. For metameric (or physical) colour science colours (metamers) have an extra, purely physical determination dimension in terms of SPD beyond hue,

---

[115] This is an oversimplified picture of colour science, SPD is not enough to track color perceptions and metameric effects, the task requires more, e.g. knowledge of how the brian adjusts colour experience with changes in environmental conditions. But for the purposes of this discussion there is no harm in simplifying things.

[116] It gets transformed into thermal energy by electrons vibrating at the same frequency.

saturation and brightness. In traditional colour science colours only have three determination dimensions. As she summarizes her moral drawn from the examples: "Different sciences may treat the same determinable as having different determination dimensions" (Wilson 2009:163). Second, by analogy psychological determinables may also have explicitly physical determination dimensions. Wilson allows that beliefs have neural determination dimensions. Relative to psychology they only have content, attitude and intensity, but relative to neuroscience they have at least one further neural dimension. To make the case more vivid, Wilson relies on an example from psychopharmacology:

> "*different forms of depression depend* on whether the depression results from disorders in serotonin-based neuronal circuits or in dopamine based neuronal circuits; depending *on which transmitter is involved*, different drugs are likely to produce a beneficial effect" (Wilson 2009:164 my italics).

Assuming that the example stands, from the point of view of neuroscience, depression described relative to psychology cannot be a fully determinate mental state, a superdeterminate depression has features invisible for psychologists without MRI scanners or other such equipment. Two mentally identical depressions can be different and can be sorted by their neural properties.

I would disagree with Wilson on the interpretation of the depression example[117]. According to recent publications on the topic (Korb 2015), dopamine-system dysfunction is responsible for low motivation, whereas serotonin-based circuitry is responsible for how one processes emotions and develops moods. Serotonin- and dopamine-based mental states are

---

[117] The source referenced for this item in Wilson's paper is personal conversation with a college (Wilson 2009:164, footnote 22), therefore I thought that it would be important to check more recent publications on the neuroscience of depression.

both constitutive of depression, not different kinds of depression. It is also questionable that the two different circuits could be taken to be different realizers of depression, in the sense required for multiple realization, as it is not true that the pure, qualitative mental aspect of the depression is independent of changes to realization. Suppose that depression is multiply realized by the two circuits. If it were based on dopamine circuits the depression as a mental state would be different compared to the case when it is based in serotonin circuits. The first case of depression would be characterized by low motivation, whereas the second would be characterised by problems in processing emotions.

One might also try to think of the different depressions as a case of kind splitting (see: Craver 2004), but it doesn't seem to be a proper example of that either. The canonical example of kind splitting comes from research into memory. In the case of this research psychologists discovered different forms of memory (procedural, declarative, episodic, etc.) by identifying different brain structures performing memory-related functions. It turned out that intervening on those different brain structures one can disrupt different fields of memory, not memory as such. Antidepressant research suggests that a particular depression can be treated via interventions into any of the two underlying systems, while problems of declarative memory cannot be treated by interventions into the realizer base of procedural memory.

In the end, the science suggests two things. First, these distinct neural circuits realize different aspects of depression as such. Low motivation and bad mood are aspects of depression, not different "forms" of depression. Second, to the extent that the neuronal dysfunctions are different in different cases, the corresponding mental states as mental states are different as well. It would be more apt to say that these depressions are different both as mental states and as neural (dopamine- and serotonin-level based) states and the second

explains the former. However, for the sake of the argument I will accept the plausible assumption that there are good examples reflecting what Wilson envisioned, cases where there is no mental difference with substantial difference at the level of realization.

Let us see the reasons Wilson put forward for accepting metameric colours as a good case of determination in terms of physical properties (see: Wilson 2009:162-163). First, metameric colours are called colours in physical optics which indicates that they are taken to be specific kinds of colour. Second, most colour research is aimed at understanding colours in terms of retinal SPDs, not simply in terms of hue, saturation and brightness, and this is the approach that helps photographers and designers of different kinds of screens to fit their products to human vision. The role SPDs play in these accounts seems to be compatible with taking colours to be partly constituted by SPD. Third, colours are understood as visual perceptual properties and SPDs are an integral part of the process of visual colour perception. Forth, thinking about colours further determined by SPDs has a unifying explanatory value. It helps to decide what we experience when we are experiencing colours: something in the head, something outside the head or something in between, a relational property. At a fine-level of grain where colours are further determined by SPDs, colours can be seen as tracking the surface properties of objects. It is also possible to think about colours as relational properties resulting from the interaction of the environment and our visual perceptual system. This is a motivating argument not recognized by Funkhouser.

In Wilson's conceptualization a metamer is a colour appearance property in the same way as colour in traditional colour science, and it is partly individuated by a broadly physical feature, retinal SPD. A metamer can be seen as a specific kind of colour that has at least four determination dimensions, hue, saturation, brightness and SPD. Considering the above this might sound acceptable. However, I should note that for Wilson metamer has the same

determinable as colour in traditional colour science. She is explicit in committing herself to this view drawing morals from the case of metamerism:

> "That different sciences may treat *the same determinable* as having different determination dimensions reflects that different sciences and their associated laws may treat the same phenomena at different levels of metaphysical grain." Wilson (2009:163, my italics)

I would like to put forward three arguments against Wilson' approach. Firstly, whatever one thinks about metamer as property kind, the above highlighted assumption is in clear contradiction with Funkhouser's determination framework. Here, I follow Funkhouser's line of argument (Funkhouser 2014:5). As we saw, for him the essence of a property kind is defined by the determination dimensions of the property. A determinable is a region in the determination space of the property in question. But if a metamer and a colour have different determination dimensions then by definition it is impossible to treat a determinable in the determination space of metamer as being the same as a determinable in the determination space of colour. We are talking about two property kinds having different essences. If the different sciences talk about different properties or different determination dimensions, then they must talk about different determinables as well.

So, Wilson's metameric morals as they were formulated suffer from a conceptual problem. She claims that the moral of metamerism can be smoothly accommodated into Funkhouser's framework, but this doesn't seem to be true. Now, one might say that this is simply a definitional issue, and we might be able to turn the table back on Funkhouser as metamer is still a perfectly fine property in itself. This might be true. But even if it is, it remains

contradictory to claim that different sciences posit the same kind, but with science relative determination dimensions and essences.

Secondly, to strengthen the case against Wilson's view, I would like to add an argument against accepting metamer as a property. The argument follows from Funkhouser's framework but was not recognized by him as a possible response to Wilson. It can be shown, that to say that metamer is a property kind is to say that there are properties certain determination dimensions of which realize the remaining determination dimensions of the same property. This implies that the SPD features that are said to determine metamer properties stand in a supervenience relation to the pure appearance features (hue, saturation, brightness) of the very same determination space. The latter supervene on the former. This leads to an internal tension in the concept of a determination space.

What is important to pin down as a consequence of the supervenience relationship highlighted is that in the case of a metamer property a particular parameter value corresponding to the SPD dimension fixes the parameter values in all other determination dimensions of the determination space. There is some wiggle room for changing the SPD parameter without changing the other parameters, because metameric lights have the very same combination of hue, saturation and brightness accompanied by different SPDs. A similar problem arises when we approach from the other side of the supervenience relation. The parameter values assigned to the three traditional determination dimensions of the determination space fix which phase of the determination dimension assigned to SPDs contains the actual parameter value. This implies a strong dependence relation between the

determination dimensions that is in stark contradiction with Funkhouser's independence criterion for determination dimensions (see criterion (3) in section 5.3.1)[118].

Wilson accepts Funkhouser's model and argues that that framework allows for the example of metamers as an example of science-relativity of property kinds. I disagree. The problem is simply this: the very idea of a determination space implies that the dimensions as parameters are independent, this means that we can move around freely in any dimension independently of others. This is definitely false in the case of metamerism.

The reason why SPD might seem to be a simple determination dimension is that metameric colour pairs (material samples we experience to have the same colour under certain lighting conditions) can be differentiated in terms of their SPD fingerprints. There is one way to accommodate SPDs into the determination space model. SPD is not a simple parameter, it encodes quite complex information, the total physical information of light. If the information it contains could be factored into four independent parameters three of which corresponds to hue, saturation and brightness while the remaining one provides information concerning metamerism then it would be reasonable to say that those parameters are the determination dimensions of metamer properties.

Matching SPD based, and traditional colour models requires a so-called colour appearance model of which there are many in circulation[119]. The aim of these models is to map SPD information onto the appearance parameters of human colour vision. The mapping

---

[118] I am only flagging this problem here, but if there are further relations between the parameters in the different determination dimensions, relations that usually go together with inter-level supervenience, the problem proliferates. Explanation, constitution, etc. are relations that are supposed to go together with supervenience in cases of realization, so the problem I raised in connection to supervenience might go even deeper. I restricted the discussion to supervenience as it provides sufficient grounds for making a clear argument.

[119] An accessible source on the topic is Wikipedia: https://en.wikipedia.org/wiki/Color_appearance_model

is based on knowledge of the physiology of vision, that is, on the sensitivity of the three different type of cone cells in the human eye to certain wavelengths of the light spectrum. Based on this mapping it is possible to predict metameric effects from SPD information. However, the mapping created for this purpose does not provide us with an independent parameter for metameric fingerprints. Scientists know how to calculate the colour appearance parameters from SPD information and the sensitivity curves of the cone cells in our eyes, but the identification of metameric pairs is based on the full SPD fingerprint in those cases where the calculations resulted in the same appearance parameters (see: White & Jenkins 2001).

This is the reason why, if we take SPD to be a determination dimension for colours, we get a supervenience relation between the different determination dimensions of the determination space created. The dimensions are not independent. Changes in hue, saturation or brightness introduce changes in SPD and also certain changes in SPD result in changes in the other three dimensions.

The determinable-determinate relation as a theory of how the realization relation works was supposed to shed more light on the realization relation itself. But now we are faced with perplexing puzzles instead of enlightenment. How can one determination dimension of the same property kind realize another? A similar issue would arise even in the case of higher-level broadly physical properties, like hardness in solids or temperature. The existence of a determination space that includes both the putative realizer and realized properties creates a serious problem for the concept of a determinable property itself.

Let us go back to Wilson's psychopharmacology example. If it were the case that the same kind – depression – depended on different kinds of brain circuits in different cases, and as a result different medication were required to ease the symptoms, then the case would

show why deep explanation in terms of local reductions can be highly important for science. The identification of the realizer in terms of the underlying mechanism not only explains the higher-level realized property, capacity, but it also suggests realizer-specific methods to intervene into how the system behaves. This is natural, as lower-level processes are supposed to implement higher-level capacities; therefore, it is fair that knowing the realizer can suggest realizer-specific means for changing the higher-level property[120].

But is this a good enough justification for talking about depression as a mental kind specified further in terms of its realizers? That move creates more problems than it solves, as it results in properties one physical determination dimension of which realizes all the other mental determination dimensions. Such properties give rise to the same problem that we saw in the case of SPDs. Let us think in terms of superdeterminate properties in a determination space. Imagine that propositional attitudes have one extra determination dimension beyond the three dimensions we saw on Figure 5.4-1, call it the realization dimension. On the original conception it was possible to change the attitude, the content and the intensity parameters independently of each other. Introducing realization as a 4th dimension violates this picture. Different propositional attitudes (maximally determined in the dimension of content, intensity and attitude) supervene on certain phases, parameter intervals of the realization dimension. Therefore, moving from one phase in the realization dimension to the other we are getting changes in all other, mental determination dimensions as well. It is the same problem as if by changing the intensity of a propositional attitude we could change its content. Or by changing the saturation of a colour we could change its brightness.

---

[120] Sober (1999) suggests a similar picture on multiple realization and local reductive explanations.

If one takes the relation between maximally determinate physical realizers and maximally determinate psychological properties to be realization, a relation that is different from the determinable-determinate relation, it remains meaningful to talk about realizer specific interventions without bringing about these problems.

The third and maybe the least serious counterargument against Wilson's view is that it might have some unwelcome consequences for physicalists[121]. The view implies that superdeterminate "mental" states have both physical and mental determination dimensions. But by the same logic this also means that "physical" realizers have mental determination dimensions as they are part of the very same determination space. However, if a property has both mental and physical dimensions, then it is not a mental or physical property but a mixed property. Let us call such property mixtures PM properties and kind mixtures PM kinds. There are at least two problems with accepting such properties into our picture of the world.

First, for physicalists the aim of reduction is to explain a mental property in terms of a physical property. What they want, if they are not of the eliminativist conviction, is to find the image of the mental property in the physical realm and to be able explain those things that were previously explained in terms of the higher-level properties in terms of lower-level physical properties. PM properties don't seem to be useful in fulfilling this promise, they seem to contaminate the physical realm with mental properties instead. Second, it seems that a purely physical description of the physical goings on would also be impossible if PM properties were accepted, as physical properties would have mental determination dimensions as well. That would cause a lot of trouble for physicalism. Including psychological properties and their

---

[121] The worry concerning physicalism presented here is motivated by a presentation held by Jonas Christensen at the Metaphysics Reading Group in Durham 06/05/2016

physical realizers as components of a common determination space only obscures and confuses inter-level relations, so I advise against this approach.

As we already saw in section 5.1, if realization is not a case of the determinable-determinate relation, then we cannot rely on the determinable-determinate relation in the evaluation of lower-level counterfactuals the falsity of which is required for downwards exclusion by Menzies and List. In the next section, I will investigate the status of realizer-level causal statements again as the conclusion of this section has important consequences with respect to their evaluation.

5.5   Exclusion without determinables and determinates?

If realization relation is not to be understood in terms of the determinable-determinate relation, as I hoped to have shown criticizing Wilson's approach in the last section, then there is no motivation for the assumption that realizer states are immediate neighbours in the similarity space, so the closest worlds without the actual realizer have another realizer in place. The implicit analogy of firing neurons with the case of colours is unjustified.

As we saw, discussing causation and counterfactuals, the closeness of worlds should be evaluated on the basis of the description of the cause event. In the classic colour examples developed by Yablo, changing the shade of colour to a neighbouring one in the similarity space does not make a difference to the outcome. Can we provide a similarly effective justification for the proximity of lower-level realizers without relying on the determinable-determinate relation between realized and realizer properties?

To decide the question, we need to have a good grasp on the relevant lower-level descriptions, so we need to attend to multiple realization claims. We either accept the way Menzies and List (2009, 2010) and Woodward (2008) have tacitly conceptualized multiple realization, which seems to reflect the commitments of the composition-based approach to multiple realizability (see: section 1.4.3.2), or we accept the flat view of multiple realization (see: section 1.4.3.1) suggested by Polger & Shapiro (2016). As I hope to show, this leads to an inconvenient dilemma. The former theory has a hard time providing a justification akin to the case of colour determination. Choosing the flat account isn't really promising either. For proper multiple realization to occur this theory expects qualitatively different mechanisms at the lower-level and even if one finds multiple mechanistic realizations for some realized property, because of the dissimilarity of the realizers, downwards exclusion remains elusive.

Let see how things plays out in the case of mental properties realized by firing neurons. Reporting the mentioned research done on monkeys, Menzies and List say:

"The neural signals that encode the monkeys' intentions to reach for certain targets were recorded as *averages of the firing rates* (spikes per second) of individual neurons. But clearly the *same aggregate firing rate in a group of neurons is consistent with a lot of variation in the behaviour of individual neurons*. For example, very different temporal sequences of neural firings can give rise to the same firing rate. So, an intention to reach for a certain target can be realized in many different ways at the level of individual neurons." (Menzies & List 2010, my italics)

The italicized parts suggest that our neuroscientists are interested in an overall pattern at the neural level, average neural firing rates in a brain region with respect to a restricted population of neurons. I will call it AV. The researchers don't measure all lower-level features
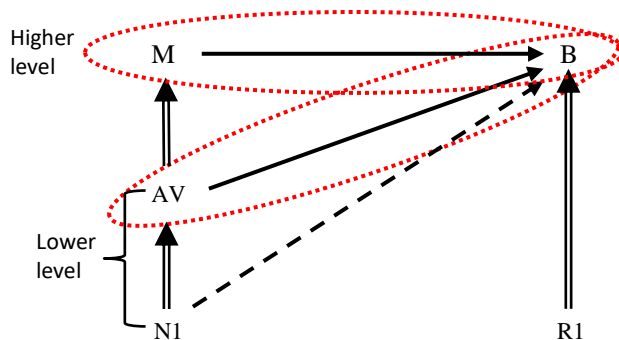

Figure 5.5-1

of particular neurons, they wouldn't be able to do so in a technical sense, their predictions are not based on particular cellular properties in the relevant neuron population (N1).

What makes them capable of tracking the intentions (M) of their monkeys on the neural level is the AV pattern. On the left-hand side of Figure 5.5-1[122] the reader can track the relations between the properties mentioned. B on the right-hand side signifies the behavioural effect

---

[122] On this figures single tail arrows depict causal connections, arrows with dashed tails signify causes that are excluded by other causal connections and two tailed arrows refer to the supervenience relation. Dotted ellipses highlight the active, proportionate causal relations.

of the intention. As a rough approximation we can say that the research identified the image

of a monkey's intention to do B in the lower-level realm (Musallam el al. 2004). I take it that

in philosophical parlance we can say that the AV pattern is considered to be identical with the

intention at least tentatively. Note, that both M and AV are proportionate difference-makers

of behaviour B. It is clear from the quoted passage that Menzies and List believe that M is

multiply realized by different neural states (N1, ...Nn).

I start with the least interesting criticism one can possibly formulate, denying the

multiple realizability claim. I agree with others, who observed that this is a cheap strategy

(Pernu 2014), but I think it to be a good way into saying something much more interesting.

Shapiro (2012) put the idea the following way: the firing pattern itself is the relevant

realization base for the intention, therefore, the monkey research does not provide a suitable

example for multiple realization as such[123]. Interestingly, Menzies charges the neuroscientists

in his case study with doing exactly the same and being unreflected reductive physicalists. He

says: "these experimenters are best seen as reductive physicalists who could plausibly claim

that the new exclusion argument is not effective against their view" (Menzies 2013:67).

An approach that endorses the tacit interpretation of the researchers reflects the

commitments of Shapiro's flat view of realization discussed in section 1.4.3.1. By the lights of

this approach, the higher-level functional kind is realized by the same structural features, the

same neural-level mechanism[124] in all cases when the intention is instantiated by a monkey

---

[123] Shapiro (2012) put forward a parallel, but different argument for this view from the one I present below, even though he could have argued simply on grounds of his interpretation of multiple realization the same way as I do. This is probably because he tried to develop an internal criticism of Menzies and List. However, I agree with Pernu (2014) that his argument is on the wrong track, so I won't consider it here.

[124] We are talking about the average firing rate of a group of specific kinds of neurons organized in a specific manner, in specific brain region. This roughly how a neurological description looks like.

brain. There is no neuroanatomical-level scientific taxonomy according to which different particular activations of the same group of specific neurons that conform to the same overall pattern would belong to a different kind[125].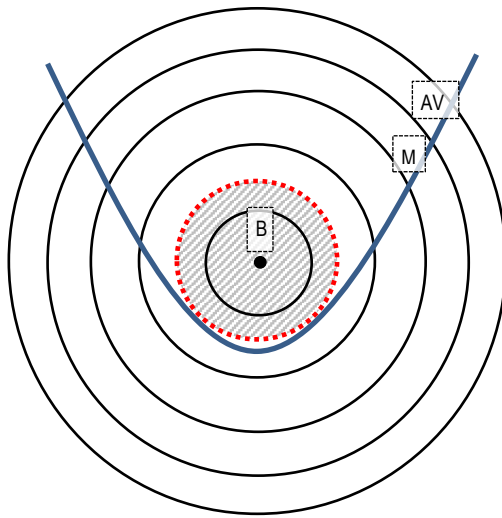 So, at best we are faced with a mild case of multiple realization, where variability between instances of the same realizer is irrelevant to the realization of the higher-level property. Analysed from this perspective, Menzies' argument is disarmed because the original multiple realizability claim is denied. B is proportionately dependent on M in the same way as it is on AV, so downwards exclusion fails (see: Figure 5.5-2). In such a setting one could establish a bi-conditional bridge-law between the two properties.

Figure 5.5-2

This is similar to the strategy by which Shapiro (2012) tried to disarm Menzies' example of downwards causal exclusion, claiming that it collapses to the case of identity. However, he did not realize that Menzies allowed for cases of causal compatibility in his framework, so he thought that this result cannot even be accommodated by Menzies in any possible way (Shapiro 2012:16). This is not so. As we saw, causal compatibility is allowed by

---

[125] AV in a specific group of neurons is the relevant realization base. AV is consistent with uninteresting variation in the behaviour of individual neurons. It is consistent with a case where one neuron fires at a lower rate if some other fires at a higher rate, it is also consistent with many different temporal sequences of neural firings. However, neurons exhibiting the same firing pattern together can vary in a bunch of other properties like osmotic pressure, axon length, number of dendritic branches, amounts of neurotransmitters produced, but these are largely irrelevant with respect to the basic organisation. Such differences can be disregarded for the purposes of mechanistic explanation and prediction.

the reformulated exclusion principle. In turn, that allows for inter-level property identity together with other yet unexplored cases of compatibility. As the case of mental-neural identity is compatible with Menzies' framework, it is unlikely that followers of Menzies would be devastated by Shapiro's argument, especially if they can find further working cases of downward exclusion. And why can't they, if multiple realization by alternative mechanisms is possible?

Even though, on Shapiro's terms, good enough empirical examples of multiple realization are hard to come by (see: Shapiro 2004, Polger & Shapiro 2016), his case against multiple realization is not a priori[126]. Therefore, we should ask the question what if M is robustly multiply realized? Shapiro (2012:15) did ask this question but, not acknowledging the option of inter-level compatibility, he did not go far in evaluating the consequences.

If there were other possible realizers of M like AV in our example, we would have to handle them somehow in the context of the reformulated exclusion argument. In cases of robust multiple realization, we need to consider distinct realizers without realization-relevant commonalities among those realizers. The relevant working parts of those systems are considerably different, which might mean different kinds of neural modules (different cells in a different organization, with different activation patterns), but even way more diverse kinds of material systems as well. In all such cases, the relevant structural property would be substantially different, as it would be based on different lower-level properties and different scientific methods of "averaging" (see: section 1.4.2).

---

[126] He argues that accepting his constraints on the notion (see footnote 24 in section 1.4.3.1) it is empirically implausible that multiple realization is as widespread as many philosophers think and the more complex the realized function in question is the less plausible multiple realization becomes (see: Shapiro 2004).

To show the consequences of this scenario, I will use a much-discussed toy example, the case of human brains and computer brains. Suppose that in human brains an intention M is realized by neural firing rates of a specific kind of neural structure in the right interval. Call this property AV1. So far, nobody knows much about computer brains capable of the same functions as human brains, but let us suppose that to realize M they require the activation of differently organized electric circuits, call this property AV2. Note, that AV1 and AV2 are supposed to be qualitatively different mechanisms. Therefore, AV1 and AV2 don't share a common determination space. As we saw, similarity measures are paramount to evaluating the relevant counterfactuals and the evaluation is based on what is taken to be the essence of an event, on what aspect of an event is highlighted by our description.

On grounds of the above suppositions a forceful argument can be formulated against the view that distinct realizers are always neighbours in the similarity space. M has two possible realizers AV1 and AV2. The counterfactual involving AV1 and B can either be true or false depending on the similarity relation between AV1 and AV2 worlds. It tells us whether the closest ¬AV1 world is an AV2 world or a (¬AV1 & ¬AV2) world. There is no a priori guarantee for either of these options, but from an empirical point of view, accepting the hypothetically different descriptions I provided for AV1 and AV2, a great distance between them is much more plausible then proximity.
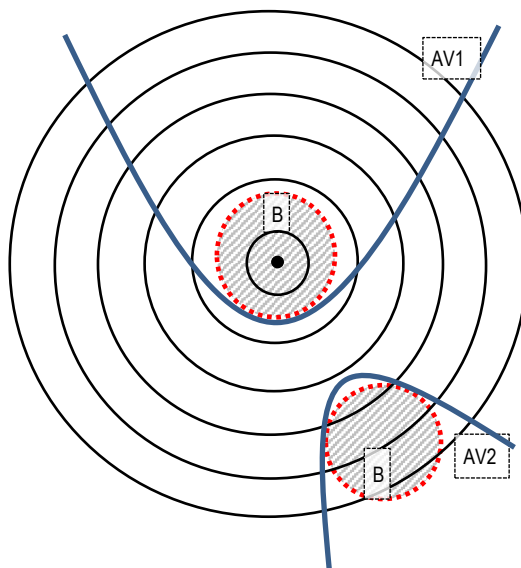
Figure 5.5-3

The distance between AV1 and AV2 worlds should be measured on grounds of the descriptions used. If the AV1 and AV2 descriptions belong to the same level of description and

they are fundamentally different as in the case of human brain and computer brains, then it is plausible that we get closest ¬AV1 states that are not AV2 states.

Suppose that descriptions AV1 and AV2 are fundamentally different, because we are talking about fundamentally different mechanisms. If AV1 refers to human brains, this makes it highly probable that a ¬AV1 state is a state of a human brain region that is not capable of realizing M. The same would go for ¬AV2 states. This scenario has an unexpected consequence for Menzies' account. If the closest ¬AV1 world does not contain AV2, then in the actual world AV1 comes out as a proportionate difference-maker for B, as the presence of AV1 guarantees the presence of B and its absence guarantees the absence of B, even though M is not uniquely realized by AV1.

This can be seen on Figure 5.5-3. The actual world has AV1 and B in place. If we move to the closest ¬AV1 worlds on the second circle around the point signifying the actual world B disappears. The closest world in which B occurs again is on the fourth circle and that world is contained by the AV2 region. Both criteria of the proportionality constraint are satisfied by AV1. The same would be true of AV2. This result depends on the way the realizers are described. If that is acceptable the result follows.

This result does not affect the truth of the relevant higher-level causal statement. M is still a proportionate cause of B. From the perspective of the higher-level causal description AV1 and AV2 are both on the closest M permitting sphere, because the higher-level description structures the similarity space around itself. 'M causes B' is still a realization-insensitive causal statement as all closest M worlds that differ only in terms of M's realizers are B worlds and all closest ¬M worlds are ¬B worlds.

As it is possible that both AV1 and M are proportionate causes of B, this is a clear case of causal compatibility (see section 4.3.3 for the conditions). M is a cause of B, so compatibility

conditions (i) and (ii) are satisfied. The same holds for (iii) as B is absent in all closest ¬AV1 worlds that are M-worlds, as there are no closest ¬AV1 worlds that are M worlds. Note that this is a special case of compatibility, one that wasn't recognized by Menzies. In section 4.3.3 I have already shown that such scenarios are theoretically possible: here I have provided a scenario that fits the bill.

What the above case proves, is that the truth value of a corresponding lower-level causal statement can be independent of the truth value of the higher-level statement, contrary to what Menzies and List believed. When a mental property is robustly multiply realized at some level of description, cases of downward exclusion might exist, but inter-level causal compatibility seems to be much more likely. The scenario I highlighted above is simply not on the chart of possible cases that Menzies and List (2009, 2010) provided (see section 4.1, Table 4.1-1). Everything I have shown in this section relied on the premises Menzies and List accepted in their work, only the multiple realization example and the interpretation were changed. I take it that my results demonstrate an important internal tension in Menzies' framework.

However, it might be that the above argument wouldn't be acceptable for Menzies and List, they would simply go back to their original monkey research example, telling us that lower-level variability is still a fact. Maybe that example is not good enough to justify a claim for robust multiple realization, but who cares? Any kind of lower-level variability is enough to run their argument. I disagree with the anticipated defence.

Even though philosophers disagree with respect to the transitivity of the realization relation (see: Aizawa & Gillett 2009 vs. Polger & Shapiro 2016), they don't disagree when it comes to the level-bound nature of multiple realization claims (see: section 1.4.3). This agreement reflects the conviction that the recognition of variation is bound to a level of

description. As I showed earlier, the same is true of the evaluation of counterfactual claims. If realizer and realized properties are not connected via the determinable-determinate relation, the evaluation of lower-level counterfactual claims should be done based on the description provided for the event in the antecedent.

Now, on the dimensioned view of realization, realization-relevant lower-level variability is easier to come by. After we got rid of those entities and properties of the realization base that play no role in the realization of the realized property, we get a population of lower-level entities, their properties and relations composing and realizing the higher-level property. According to this view, variation in the exact configuration of those entities, their relations to each-other or their properties constitutes proper multiple realization even if it is not a case of robust multiple realization, so it seems that this approach agrees with the way Menzies and List view the case of the monkey brain research. So, one might simply object that the right theory of realization would deliver exactly the result Menzies expected. However, I think that we don't need to decide which account is better in general. The difficulty observed in the case of the Polger-Shapiro account exists in the context of the Gillett-Aizawa account as well.

It is certain that not just any kind of change in the configuration or properties of the relevant lower-level component entities results in a state that realizes the higher-level property in question. To decide whether a configuration of realizer entities P11 is a difference-maker for the higher-level outcome, we should see how that particular description looks and check whether the closest ¬P11 states are realizers of the relevant realized property M.

Suppose that the neural-level description of lower-level affairs includes a network of certain type of neurons situated in an area of the brain[127]. All possible realizer states (P11, P12, …, P1n) are described at the same neurocytological level. As Aizawa and Gillett (2009) remind us drawing on recent research concerning memory formation, proteins relevant for such processes can be multiply realized by different proteins. "There are many combinations of amino acids that can be bound into chains that have" the very same realization relevant property (Aizawa & Gillett 2009:197-198). Suppose further that one component realizer in P11 is such a protein PR1, there are other such proteins PR2 and PR3 that could do the same work. One way of going to a closest possible ¬P11 world is to go to a world which is P11 in all respects except that it is not a PR1 world.

A typical protein is a combination of around 200 amino acids, so the available set of combinations has an immense number of members and we know from empirical research only a handful of those combinations can replace PR1 functionally. We also know, that the biologically available replacement proteins are not too close to each other, they are many steps away from each other in terms of their amino acid components. Therefore, empirically it is highly likely that the closest possible amino acid combination to PR1 is not PR2 or PR3. This abstract example gives some plausibility to the view that closest possible ¬P11 worlds are worlds where realizers of the relevant mental property M are absent.

If one accepts the semantics for proportionate causation proposed by Menzies and List a lower-level realizer state is a difference-maker only if it has no neighbours in the similarity space that realize the very same property M, where M is a relevant cause for the effect we are interested in. This is definitely a high bar. It is not impossible that all realizer

---

[127] In the monkey example we are talking about the parietal reach region of the primary motor cortex which is responsible for the control of execution of movements among other things.

states described at a certain level have at least one other proximal neighbour that is also a realizer state. However, downward causation fails even if there is one possible realizer of M that has no immediate neighbour that is an alternative realizer for M at the same time. Whenever this is true, we have a case of inter-level compatibility at hand.

Note that with respect to the decision concerning distances in the similarity space, it doesn't help to know that the different realizer states are realizing the same M, because, in agreement with Menzies, we gave up on the view according to which realized and realizer properties stand in the determinable-determinate relation. Lower-level states are only comparable in determination dimensions applicable to them on their own right. Therefore, we lack grounds for arguing that in the closest possible world where the actual realizer ceases to exist, the effect is absent as well.

What made Menzies and List think that small variations in the realizing states are proximate neighbours in the similarity space? As far as I can see, they were influenced by how their case study was framed conceptually. A certain level of average firing rate (AV) as a property and a particular firing pattern as a property can be described as determinable and determinate property pairs. It is plausible to think that there are many different particular patterns that conform to AV and particular firing patterns are more determinate AVs, even if it wouldn't be easy to spell out the determination space that describes the relationship between AV and more determinate AVs. Suppose that this interpretation is acceptable, and this is what explains why Menzies and List thought about their case study the way they did. In that case, I don't see why Menzies (2008) distanced himself from Yablo's determinable-determinate interpretation of the realization relation. If tacitly the model put forward by Menzies and List relies on the property determination interpretation of the realization relation, and they have no other justification for supposing that realizer states are immediate

neighbours in the similarity space (see: section 5.2), then they have failed to achieve the theoretical goal of inventing an argument that preserves the virtues of Yablo's account, based solely on a theory of causation.

If this section's argument holds water, then in cases of putative downward exclusion, we must consider that whether or not the closest possible world has a relevant realizer state in place is a contingent matter. It is contingent on the state of the lower-level system in the actual world and on what resides in the closest possible worlds where this state is absent, on what are the closest properties in the lower-level similarity space. This is because we need to evaluate the truth-value of counterfactuals iii. and iv. of section 5.1 and these involve only lower-level descriptions in their antecedents.

I have shown, that if the evaluation of the lower-level causal statements in done based on the lower-level descriptions provided independent of the higher-level description, there is a good chance that a example that look like a case of downwards exclusion turns out to be a case of inter-level causal compatibility. In other words, contrary to what Menzies and List claim, multiple realizability is not a sufficient condition for downwards exclusion.

In the section that follows, I develop an independent argument to prove the very same point from a different angle. I aim to show, that the same might be true even if the lower-level realizers are closest neighbours in the similarity space.

5.6   Shifting background conditions and local causal compatibility

In section 5.2 I discussed the idea of causal invariance, the sensitivity or insensitivity of causal relations to changes in background conditions. I also explained how this idea was put to use by Menzies, an application that gave rise to the idea of realization sensitivity. Here, I would like to go a step further and to show that sensitivity to realization and sensitivity to general background conditions are different but probably not independent features of causal relations and the difference has interesting consequences with respect to the causal autonomy of higher-level properties in the context of the reformulated exclusion argument. The arguments in this section are independent of the conclusions of section 5.5 as this argument does not depend on considerations concerning the distance between putative cause properties in the similarity space.

5.6.1   Two kinds of sensitive causation

The idea of sensitivity in Menzies' discussion originates from Woodward. He considers both kinds of sensitivity when explaining the idea, but he never investigates the two aspects separately, nor does he investigate their possible interactions. He explicitly tells us that we should consider them together (Woodward 2006:3)[128]. In his paper on the exclusion problem he highlights the issue of realization sensitivity (Woodward 2008:241) but does not consider the interaction of realization sensitivity with sensitivity to background conditions. Here, I am interested in exactly that.

---

[128] In the paper referred here he is not interested in realization, what he considers as one aspect of sensitivity is whether a causal relationship remains stable under slight changes in the instantiation of causes and effects. He is also explicit in saying that when he explores the issue of sensitivity, he will allow for both kinds of variations without separating them.

Remember that realization insensitivity is the idea that a higher-level causal relation is usually expected to be robust, unbroken if changes are made to the way the cause property is realized. For Menzies sensitivity to realization, as he defined it with List (see section 5.2), is a binary issue. A relation is either sensitive or insensitive to realization. This is one aspect of sensitivity. The second aspect, what I will call sensitivity to background conditions, occurs when we fix the manner in which a property is instantiated or realized and only attend to changes in the background circumstances. Borrowing the manner of presentation from Woodward (2006), we should attend to the evaluation of the following counterfactuals relevant for the sensitivity of the causal relation between C and E:

1a. C □→ E (under set of background conditions BG1)

1b. ¬C □→ ¬E (under set of background conditions BG2)

1c. If a C-like event occurred under some background condition BG11, an E-like event would occur.

1a. is the presence condition for the causal relation in question. It tells us that the counterfactual relation between C and E (where the realization or the manner of instantiation of properties C and E is fixed) holds under BG11. 1b. tells us the same in the case of the non-occurrence condition. I won't discuss 1b. as the other conditions are deemed to be far more important in the literature and similarly for my aims here. 1c. tells us that there is a particular setting, background condition BG11 that allows for variation in the way C and E are instantiated or realized.

The idea I would like to test and prove in this section is the following. Whether a causal relation is sensitive in the dimension of realization sensitivity might depend on the background conditions that hold. And also, that the sensitivity of a causal relation to

background conditions might depend on the way the relevant properties are instantiated or realized.

Why is this interesting? Simply because Menzies and List argued that multiply realized higher-level cause properties have different "difference-making powers" than their realizers:

"these properties can typically have multiple physical realizations, they are not identical to physical properties, and further they possess causal powers that differ from those of their physical realizers." (Menzies and List 2010:108)

What is even more, when they explain their account of higher-level causation, they claim something stronger:

"we employ this theory [proportionate causation] to specify the conditions under which an instance of a higher-level, special-science property can have causal powers not possessed by the instance of the physical property that realizes it." (Menzies and List 2010:109)

In short, their claim is that downwards exclusion scenarios prove that multiply realized higher-level properties have different and more difference-making powers than their realizers. This is a serious claim on the side of Menzies and List. If higher-level properties have causal powers not possessed by their realizers, then higher-level properties cannot inherit their causal powers from their realizers. As some form of causal inheritance is presupposed by most theories of physical realization the denial of causal inheritance amounts to ontologically serious autonomy for higher-level properties.

One of my aims here is to provide more precise conditions for downwards exclusion based on further considerations concerning the sensitivity of causal relations to background conditions. At the same time, I would like to show that the sensitivity of causal relations to

background conditions is a factor that provides grounds for rejecting the conclusion according to which multiple realization is enough for higher-level autonomy and it does so independently of the argument of section 5.5. In my view, multiple realizability is a necessary, but not a sufficient condition for autonomy.

For Menzies and List the only alternative interpretation of the status of valid higher-level causal claims seemed to be causal compatibility. Moreover, the only viable case of compatibility seemed to be the identity of the relevant higher and lower-level properties, although Menzies (2013) claimed to be open to other interpretations of compatibility. Below, I would like to show on different grounds than in section 5.5 that identity[129] is not required for causal compatibility and that it is consistent with the multiple realization of the higher-level property in question.

As I said, downwards exclusion by higher-level properties is a serious claim as it is a prima facie rejection of causal inheritance that is a crucial premise for most physicalists. Below, I will attempt to formulate an objection showing that causal inheritance might be reasonable even when a property is multiply realized. To do that I will set aside everything I said in previous sections of this chapter and will investigate this idea independently. What I will show in the section that follows is that the difference-making powers of higher-level properties might depend on their lower-level realizers even on Menzies' account.

---

[129] The same applies to singular realization, for most physicalists singular realization would imply identity. Some emergentists offer autonomy solutions that are compatible with singular realization.

5.6.2 Proportionate difference-making and changing background conditions

In this section I will investigate two simplified, but, in many respects, realistic examples of multiple realization. The example of mechanical hardness in different kinds of material comes from condensed matter physics and is motivated by a similar example used by some authors in the multiple realization literature. The example shows that higher-level causal relations can be sensitive to background conditions in a manner that interferes with their supposed insensitivity to how the higher-level cause property is realized. This leads to unexpected cases of inter-level causal compatibility and results in the claim that multiple realization is insufficient for downwards exclusion to occur.

The second example comes from the more practical context of caving. Traditionally there are different available ignition systems for carbide lamps widely used in caving and speleology, but one is used way more often than the other for a reason that sheds light on an interesting feature of the realization sensitivity of higher-level causal relations. It shows that realization sensitivity might apply to cases of upwards exclusion. It does not mean that the higher-level property is not causally potent, it only means that under the particular background conditions in place only one particular realizer property can manifest its powers to cause the outcome. Therefore, upwards exclusion might also be a local phenomenon.

5.6.2.1   The case of local inter-level causal compatibility

Let us turn to an example of multiple realizability the original version of which I came across reading Aizawa and Gillett (2009). Their central example for multiple realization has to do with mechanical hardness in teeth and teeth fillings. These different materials, different metal alloys and enamel all share the physical property of having a Knoop hardness somewhere between 350-400 kg/mm^2. So, t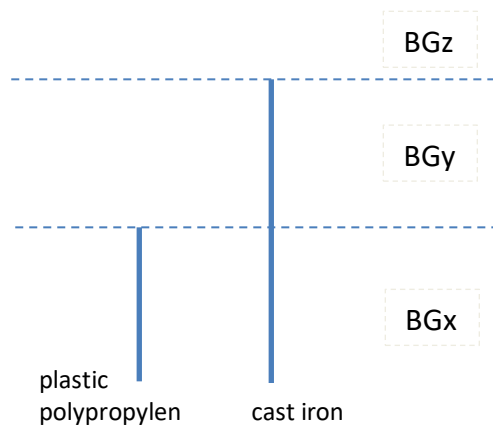hey are different in lower-level terms, but the same in higher-level terms in a quite straightforward sense. Their hardness (among some other mechanical properties) makes them perfect replacements to serve the very same function of crunching food in our mouths. In what follows, I will rely on an example similar in spirit that I contrived while studying the scientific literature on mechanical hardness. That will help me to highlight some aspects of multiple realization that are highly relevant for the evaluation of the idea of realization sensitivity. So, forget eating and teeth, for my toy example we will put the hardness property to use in a mortar in the form of a durable shaft that has to have the mentioned level of Knoop-hardness to be able to serve its purpose of smashing things into pieces.

Figure 5.6-1

In the example the Knoop hardness (350-400 kg/mm^2) will serve the role of the higher-level property required for the smashing of those softer things that are put into the mortar. This property is multiply realized by the different materials the shaft is made of in different cases. It could be realized by cast iron and polypropylene a form of plastic as well. The shared hardness property is multiply realized by these materials on any standards of multiple realization; however, I will concentrate on what the example suggests interpreting it on flat

account. As I clarified in my section 1.4.3, according to a number of philosophers (Shapiro 2004; Francescotti 2014), proper multiple realization requires different lower-level mechanisms for the same higher-level function, which is more than mere difference in constituent parts. Hardness is realized by different mechanisms in the different kinds of materials, as these materials have different micro-level bond structures with different kinds of active binding forces[130] (see: Gilman 2009). Now, the example materials are close to each other in terms of hardness, in fact they can acquire the very same value on the Knoop-hardness scale, but they are far away from each other in terms of their melting points. Cast iron melts at around 1150 ^{0}C, the plastic at around 250 ^{0}C. Let's make some serious simplifying assumptions. For the sake of the argument suppose that the higher-level causal relation between the instantiated hardness property (C) and the smashing (E) it brings about can only be realized by the two materials mentioned. The only background condition we will consider for the causal relation between C and E is background temperature.

The realization sensitivity of the higher-level causal relation can be modelled in counterfactual terms by moving from possible worlds with cast iron as a realizer for hardness to worlds with plastic as a realizer. The sensitivity to background conditions can be modelled by changing the temperature. Figure 5.6-1 summarizes the situation. The vertical lines

---

[130] Gilman (2009) details the physics of mechanical hardness in different kinds of solids. In the introduction, he highlights the relevant history of the physics of mechanical hardness as well. A half century ago physicists believed that a unified treatment of hardness properties will be provided on some microphysical basis. Later it turned out, that the mechanisms that underly hardness (defined as resistance to permanent deformation), toughness (the ability to absorb mechanical energy by permanent plastic deformation without fracturing) and many other mechanical properties are material kind specific. Polymers and metals, like plastic and iron have radically different kind of internal atomic/molecular structures. Their mechanical properties are based on different structural features and kinds of bonding relations. Therefore, calculating the higher-level hardness properties from lower-level information requires different models and equations.

represent the range of background conditions under which a realizer can realize C. The dashed horizontal lines represent the boundaries between interesting regions of possible background conditions. The BGz region has a temperature above 1150 ^{0}C, BGy has it between 1150 and 250 ^{0}C and BGx has it between 250 and -20 ^{0}C.

If the actual background condition for the causal relation between C and E falls into the BGx region the C-E relation is realization insensitive[131]. With changes in realization the higher-level causal relation remains constant, so downwards exclusion takes place. However, BGy is a special region, there the plastic shaft melts or burns away. Only the cast iron shaft would remain operational, which means that our hardness property is singularly realized in the BGy region[132], under that particular set of background conditions. In the BGz region, there are no available realizers for the higher-level property in question.

The situation in the BGy region sounds like a case of inter-level causal compatibility. As we saw in section 4.3.3, the only case of inter-level causal compatibility Menzies could find[133] was the identity of lower and higher-level properties. However, the case of the cast

---

[131] Here I set aside the issue that two realizers might be far away from each-other in the similarity space. This wouldn't be an issue in the interventionist theory as the general semantic framework differs from the Menzies & List interpretation. The ideas presented here are independent of the conclusions of the previous chapter.

[132] If realization is understood the way Gilman (2009) as a physicist and Shapiro (2004) suggest, so the same higher-level property is realized by different atomic-level mechanisms, then diamond comes out as a good example for something singularly realized as the kind of material that we take to be the hardest solid. Nothing else realizes the extrem hardness value of diamonds (7000 kg/mm^2) regardless of background conditions, only one kind of bond structure that develops only between carbon atoms. Singular realization does not imply identity, but the rejection of identity requires motivation, like insufficient explanation of the higher-level property by lower-level features.

[133] In cases of inter-level compatibility both the higher and the lower-level property comes out as a proportionate difference-maker cause. Menzies highlights this scenario as a logical possibility compatible with his model. However, on grounds of what one can see from his writings on the topic, he wasn't a believer of the identity theory. Most of what he had to say about the scientific examples of mental causation suggests that he rejected

iron shaft is not a case of identity. It is a local case of singular realization, however globally, in our world, the Knoop-hardness of the shaft is multiply realized. Therefore, if changes in background conditions are allowed by the semantics utilized for the relevant counterfactuals[134], then the compatibility condition provided by Menzies applies to both identity and to singular realization localized to a particular possible world.

As I have shown already in section 4.3.3 the definition of compatibility allowed for more than identity. Now, I have a case at hand where the realized C property is not identical with the realizer property (general iron microfeatures), however both the higher-level property and the lower-level property are proportionate difference-makers. Why? Because it is true that in relevantly similar worlds, in other words, under fixed BGy conditions, nothing else realizes the required Knoop-hardness value, even though in worlds faraway in terms of how the worldly context is structured, somewhere in the BGx region, there are other available realizers. Therefore, even though the higher-level property is multiply realized, the counterfactual test for proportionate causation provides positive results for both the realizer and the realized properties. This example shows that the multiple realizability of a causally potent property is necessary, but not sufficient for downwards exclusion.

---

the identity theory and criticized the implicit reductive physicalism of neuroscientists. Therefore, I doubt that inter-level compatibility was taken to be a live option by Menzies.

[134] In the original 2009 Menzies and List article this wasn't the case, for reasons that remained unexplained in their paper they restiricted their model of counterfactuals to worlds that differ only in terms of the putative causally relevant properties. However, most theories based on counterfactuals allow for other kinds of modifications as well and the restriction can be undone without doing any harm in the Menzies & List model as well. In Lewis' classical model there can be faraway worlds where an event that counts as a cause in the actual world exists, but without the effect it has in the actual world. This is made possible by changes in the background conditions, in the causal field to use Mackie's classic term.

The moral of the above discussion can be extended to cases where one starts out with the dimensioned view of multiple realization in mind and it can also be applied to monkey research example Menzies and List used to illustrate their view. I started with a simplified example accepting the flat view of realization, because it made it easier to highlight the essence of the argument. However, the same logic applies in the context of the reformulated exclusion argument independently of the view assumed concerning multiple realization. Let us consider the monkey research example. Realizers of the monkey's intention M are neurologically different even if the difference is quite small. In different realizers, different particular neurons fire and the neurons are in slightly different states. Imagine that your task is to disrupt the neural process that leads to the expected outcome of the intention by lower-level interventions. You will find that there are possible interventions that would disrupt the workings of certain realizers, but not of others. This is because of the mentioned differences between the realizers. Therefore, theoretically it is plausible that there are background conditions under which only one possible realizer could bring about the expected outcome, because under those circumstances all other possible realizers would be disrupted by some environmental factor or other. This is plausible even if it is true that there are circumstances under which most realizers would be effective in bringing about the expected effect and there are interventions that could stop all realizer of M from bringing about the same effect.

To sum up, inter-level causal compatibility can occur in the absence of inter-level property identity. This scenario is interesting as I have found an example of inter-level compatibility that Menzies and List allowed for but could not provide. This case should be interesting for Menzies and List as the case identified does not require the identity of property types at different levels, it can be satisfied even if one believes that multiple realizability is the only possible relation between properties at different levels.

5.6.2.2   The case of local upwards exclusion

Menzies generalized about upwards exclusion (see section 4.3.1) based on a single example he borrowed from Woodward. I think, this example introduced a bias into the description of the phenomenon, so it would be advisable to develop alternative examples.

In the original example among the croupier's possible hand movements only one causes the ball to fall into a particular slot, but the croupier has no fine enough motor control over the relevant hand movement. So, according to Woodward and Menzies, the voluntary hand movement, the croupier's throw of the ball is not a proportionate difference-maker for the outcome of interest, however an uncontrollable micro-level feature of the throw comes out as a proportionate difference-maker. What we see in this example is a causally irrelevant macro property that supervenes on a causally relevant lower-level property and also on some causally irrelevant lower-level properties. So far, so good.

Note however, in the above scenario nobody would consider the higher-level realized property as the cause of the outcome. I think it would be more interesting to investigate a case in which a higher-level property has the potential to cause the outcome of interest, therefore we would consider it to be a serious candidate cause, but the circumstances are such that only a particular realizer of that property can manifest its power to bring the effect about. Relying on the classical exclusion argument, Kim claimed that lower-level properties always exclude multiply realized higher-level properties from causal efficacy. He investigated cases where the higher-level mental property seems to be a proper cause from the point of view of a psychologist and drew the surprising conclusion that only the lower-level realizer of that property has causal efficacy. So, for Kim upwards exclusion ruled supreme, but for him lower-level properties ruled over all seemingly causally potent higher-level properties.

I think such kind of causal competition is possible in Menzies' framework as well. Consider the following example. In caving and speleology people used to use two different kind of carbide lamps to delve into the dark below. Most caves have humid air and a lot of water is dripping and flowing around, therefore it is practically unavoidable that the caver and her equipment becomes at least somewhat wet. Now, older carbide lamps had an ignition device that utilized the same mechanism as a traditional flint lighter. When the flint striker gets humid, temporarily it loses its ability to produce sparks and to ignite the acetylene gas that flows out of the flame tip at the end of the gas tube, which might be a great pain in a dark cave. On more developed carbide lamps, they installed a piezoelectric ignition device that is way less prone to failure under humid conditions. One can ignite the lamp even in rainfall[135]. It is plausible to say that ignition as a capacity is multiply realized by the flint striker and the piezo igniter. The different igniters have different strengths and weaknesses, and this is what manifests itself in the example of the carbide lamps.

Let us make the simplifying assumption that there are only these two possible ignition devices and imagine a situation where a caver in a wet cave (call this background condition BGa) has to reignite her carbide lamp. Is it true in the actual world that the higher-level property of being an ignition device (C) is the proportionate difference-maker for the reappearance of the flame? It is not. In the actual world, under background conditions BGa only the piezo igniter does the trick (see Figure 5.6-2). In the closest possible BGa world where the flint igniter occurs the flame (E) remains absent. This is a clear case of upwards exclusion as the piezo igniter comes out as a proportionate difference-maker for the occurrence of the

---

[135] I can back this up with my personal experience. In 1998, on a dark summer night, along with 2 other cavers, I was trying to find my way back from a cave to a small village in the Carpatians. The only carbide lamp that worked in the rainfall was the one with a piezoelectric igniter.
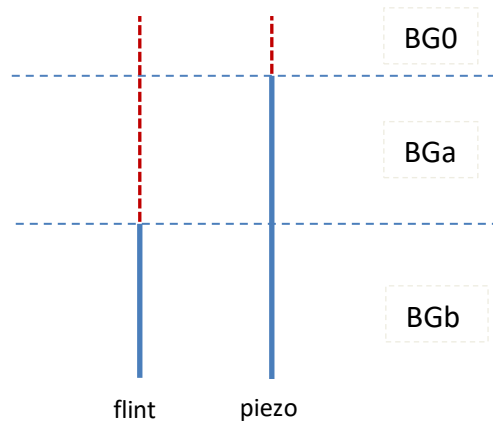
Figure 5.6-2

flame (E) as the flame is absent in all closest worlds with an igniter in place that don't have a piezoelectric igniter in place.

Let me highlight that under conditions BGa both igniters can occur, so the higher-level property can be instantiated by different underlying structures. The catch is that in the case of the flint igniter, the relevant causal power of the higher-level property is prevented from manifesting its powers by the environment. The flint igniter does retain its potential to produce sparks and will be able to manifest it under dry conditions again, even though it cannot manifest it under circumstances BGa.

This example shows, much like the one in the last section, is that the multiple realizability of a causally potent property is necessary, but not sufficient for downwards exclusion. For Menzies and List multiple realization seemed to be enough for downwards exclusion. If my examples hold water, they provide reason to reject that view. Now we can see, even when a causally potent property is multiply realizable it might be that it does not exclude its realizers from causal relevance under certain background conditions. Under those conditions the opposite is true, the higher-level property gets upwards excluded by one of its realizers. However, under different background conditions BGb, on your terrace on a dry day, both igniters would work perfectly. Under such background conditions downwards exclusion takes place. This shows that the judgement concerning the direction of exclusion is not only dependent on the multiple realizability of the higher-level property, but also on the background conditions in place in the actual world.

The second moral of my toy example concerns the difference-making powers of higher-level properties. Menzies and List claimed that multiply realized higher-level causes have "more" difference-making powers than their realizers simply because they are multiply realized. Well, it seems that sometimes they do, but only when the background conditions are such. Under certain background conditions multiple realization implies that higher-level properties have more difference-making powers than their realizers. However, there might be clear cases when the multiple realizability of the higher-level cause property is insufficient for realization insensitivity and inter-level compatibility occurs. Therefore, I take it that Menzies wasn't successful in proving the general claim that multiply realized higher-level properties possess difference-making powers not possessed by their realizers. Sometimes this is true, but not always.

5.6.3 Localized difference-making powers

I started chapter 2 with some preliminary discussion of causation. There I already pinned down that causal relations are bound to certain background conditions. What we saw in section 5.2 and in the last two sections strengthens this conviction. Now, I would like to raise a point based on the insights I gathered.

When Kim argued for local reductions to save mental causation from the danger of epiphenomenalism his argument went like this. Because of causal exclusion we should go for local reductions and should suspend the ontological seriousness of mental discourse that can only provide real unity for species or kind specific mental properties. Mental kinds like pain-in foxes or pain-in-humans are causally heterogeneous as on Kim's view physicalists should accept that localized (structure specific) higher-level properties are inheriting their causal powers from their necessarily diverse realizers:

> "this principle enables us to give a specific interpretation to the claim that [the different local realizers] are heterogeneous as kinds: the claim must mean that *they are heterogeneous as causal powers* that is, they are diverse as causal powers and enter into diverse causal laws." (Kim 1992:17, my italics)

I don't aim to advocate Kim's approach to higher-level properties or his idea of the inheritance of causal powers. I would only like to summon the spirit of his argument to try to translate the problem he posed to the kind of causal language utilized by Menzies who explicitly rejects all versions of the causal inheritance principle. When arguing against causal inheritance, Menzies spells out the difference between realized and realizer properties in terms of "difference-making powers". Causal autonomy applies to cases where the difference-making powers of the higher-level property instance are not a subset of those of

the lower-level property instance. In such cases lower-level candidate causes don't make a proportionate difference to the outcome, remember that proportionality in itself provides an exclusion principle, so lower-level property instances can't transmit their causal powers to higher-level properties as in the end with respect to the relevant outcome they don't have any. The main aim of Menzies' project was formulated this way:

> "we employ the difference-making account of causation to determine whether the thesis of the Physical Determination of the Causal Powers of Mental States is true."
> (Menzies and List 2010:115)

For Menzies, causally relevant higher-level properties are not causally homogeneous, contrary to Kim's conceptualization, but they have difference-making causal powers not possessed by any of their physical realizers.

My question is the following, is there a plausible way to argue in the framework created by Menzies that the difference-making powers of higher-level properties are somehow dependent on their realizers? I think there is and the conceptual scaffolding for this to make sense is only a short step away, most of the conceptual tools are already at hand.

If we fix the background conditions under which a higher-level causal relation between C and E holds and attend to the sensitivity of this relation to changes in the realization of the cause, we find the following: the realization sensitivity of the causal relation between C and E depends on the background conditions. Let me unpack this thought.

As we saw, there are two dimensions of sensitivity for causal relations. Sensitivity to realization and sensitivity to background conditions. As I established in section 5.6.2, the sensitivity to realization is contingent upon the background conditions in place. It might be, that the causal relation between the cause and effect variables is insensitive to realization

under certain sets of background conditions, but sensitive to realization under other sets of background conditions. Examples for the latter kind of case are provided by local inter-level causal compatibility and local upwards exclusion scenarios.

It is interesting that the scope of a causal relation, in other words the range of background conditions under which the relation holds is usually taken to be something that comes in degrees. There are fragile and robust causal relations and the difference between them is their degree of invariance with respect to background circumstances. Usually we like the more robust causal relations or laws of nature, but for certain purposes nature only provides relations with narrower scopes (for an extended discussion see: Woodward 2003, chapter 6). In my view, the same applies to the case of realization sensitivity.

Menzies and List treat realization sensitivity as if it were a binary issue. Either a relation is realization sensitive or it is not. And to some extent this treatment is justified. They were interested in the question when is it justified to favour higher-level causal relations over lower-level ones? The answer is, whenever the higher-level causal relation is insensitive to realization. What I would like to add here is that we shouldn't stop there. There are causal relations that are more sensitive to realization and there are others that are less sensitive. Sensitivity to realization, similarly to sensitivity to background conditions, comes in degrees. If one accepts this basic picture concerning causal relations, then there is only one step further to see that realization sensitivity can be a parameter that helps me to show that the difference-making powers of higher-level properties are dependent on their realizers in an interesting sense.

As we saw in the case of local inter-level causal compatibility, under everyday conditions a shaft made of plastic can be utilized to do the same job as a cast iron shaft. The difference is that the iron shaft is available to do that job under a much wider range of

background conditions than the plastic shaft and under certain circumstances it is the only realizer in town. This toy example of mine has shown that if the realization sensitivity of higher-level causal relations is tested under fixed background circumstances, as the semantic theory of Menzies and List suggests us to do, then we might get surprising results. Realization sensitivity under one set of conditions and insensitivity under others.

However, accepting that realization sensitivity comes in degrees, this thought experiment can be pushed further. It can be said that the degree of realization sensitivity of a causal relation depends on the background conditions in place. On Menzies' account this means that the difference-making powers of a higher-level property C might shift with changes in the background conditions. It is possible that causal relation C –> E is such that depending on the background conditions in place (BGa, BGb, BGc, BGd and BGe below) different sets of realizers are available for C. Below, the available realizers under certain background conditions are enlisted between the brackets after C:

in BGa: C (P11) -> E   sensitive

in BGb: C (P11...P1n) -> E   insensitive

in BGc: C (P21...P2n) -> E   insensitive

in BGd: C (P11...P2n) -> E   robustly insensitive

in BGe: C (P11...P3n) -> E   super insensitive

Depending on the size of the set of available realizers under the conditions specified the degree of realization sensitivity changes from sensitive to super insensitive. My point is that the shifting difference-making powers of the higher-level cause can only be explained in terms of the realizers available under a particular set of background conditions. Getting back to my earlier toy example, in an everyday context (BGx) it is an insensitive causal statement

that mortar shafts with a Knoop hardness around 350-400 kg/mm² do the smashing we need, while in a special context with high temperatures (BGy) it becomes sensitive to realization (in the case of local inter-level compatibility). What was added to the moral of this example in this section is that there are more options available than what this simplified picture offers.

I am talking about the causal heterogeneity of higher-level kinds or properties in a different sense than Kim, but the logic is not that far from Kim's. For Kim realized properties inherit their causal powers from their realizers (see section 2.2.4). If the realizers are different, then the higher-level causal powers are different. In Menzies' framework, even though we are not talking about the inheritance of the causal powers of individual realizers, the difference in available realizers leads to a background condition dependent difference in the difference-making powers of realized properties. This is reflected in the changing level of realization sensitivity that depends on the background conditions. The reason why different realizers are available or are active under different background conditions is that those realizers have somewhat heterogeneous causal powers and therefore they interact with their environment in a somewhat different manner. The sensitivity of higher-level causal relations changes with the background conditions exactly because of that causal heterogeneity.

So, if realization sensitivity comes in degrees and the degree of realization sensitivity for a causal relation depends on the background conditions, because those determine which realizers are available or active and which are not, then it is plausible to say that the difference-making powers of a higher-level property, in terms of its realization sensitivity, depends on what realizers of the putative higher-level cause are available under certain background circumstances. Therefore, it is meaningful to localize the difference-making powers of a realized property, to say that its difference-making powers are determined by the properties of its realizers.

5.7   Conclusion: The causal status of lower-level realizer properties

My general conclusion for chapter 5 is this:

- In case higher-level cause M of outcome B is multiply realized in any of the discussed senses, it is highly likely that at least some realizers of M are also proportionate difference-makers of outcome B.

This thesis contradicts Menzies and List who argued that only the higher-level property is a proportionate difference-maker in cases where multiply realizable properties are causes of some outcome. My thesis follows from starting points accepted by Menzies and List and some further assumptions concerning multiple realization that were never explicitly discussed by Menzies and List. Therefore, I claim to have provided an internal criticism of the proposed solution to Kim's old exclusion problem. The conclusion only concerns the Menzies and List approach to the exclusion argument. Others like Woodward (2008, 2015) formulated similar, but somewhat different, compatibilist solutions to Kim's exclusion argument based on a mature version of the interventionist theory of causation. What I have said in this chapter is no argument against those treatments of the exclusion problem.

I have provided two independent arguments for the general conclusion. One depends on the particular semantics for counterfactuals relied on by Menzies and List. According to this argument (section 5.5), without presupposing that the determinable-determinate relation holds between realized and realizer properties one cannot secure that among the closest possible worlds without the putative lower-level cause of an outcome, for which the realized property is a proportionate difference-maker, there are worlds with another realizer of the same realized property. This means that the multiple realization of a causally efficacious property does not licence a downwards exclusion scenario. Depending on one's view of

realization this conclusion can be made more or less plausible, but it is plausible enough on most mainstream views of realization.

The second argument says that (see section 5.6), independently of the relation between realized and realizer properties, whether a higher-level property inherits the difference-making powers of its realizers depends on the background conditions in place. It might be that under certain circumstances only one realizer of a higher-level property is available. Under such conditions the difference-making powers of the realized and realizer properties are identical, while they might not be identical under other circumstances.

Both arguments support the conclusion that (1) higher-level causal autonomy or downwards exclusion requires more than the multiple realization of higher-level properties (2) and that even if one accepts the possibility of serious higher-level causal autonomy it occurs less frequently than Menzies and List supposed.

References:

Aizawa, K. & Gillett, C. (2009). The (Multiple) Realization of Psychological and Other Properties in the Sciences. *Mind and Language* 24(2), 181–208.

Aizawa, K. (2018). Multiple Realization and Multiple 'Ways' of Realization: A Progress Report. *Studies in History and Philosophy of Science Part A* 68, 3–9.

Anscombe, G. E. M. (1971). *Causality and Determination: an Inaugural Lecture*. Cambridge University Press.

Árnadóttir, S. T. & Crane, Tim (2013). There Is No Exclusion Problem. In S. C. Gibb, E. J. Lowe, & R. D. Ingthorsson (Eds.), *Mental Causation and Ontology*. Oxford University Press: 248-266.

Baumeister, R. F. & Finkel, E. J. (Eds.). (2010). *Advanced Social Psychology: The State of the Science* (1 edition). Oxford; New York: Oxford University Press.

Baysan, U. (2015). Realization Relations in Metaphysics. *Minds and Machines* (3), 1–14.

Baysan, U. (forthcoming). Emergence, Function, and Realization. In S.C. Gibb, R.F. Hendry and Tom Lancaster (eds.), *Routledge Handbook of Emergence* (London: Routledge, forthcoming in 2019)

Beebee, H. (2004). Causing and Nothingness. In Collins, J. and Hall, N. and Paul, L. A. (eds.) *Causation and Counterfactuals*. MIT Press: 291–309.

Bennett, K. (2003). Why the Exclusion Problem Seems Intractable and How, Just Maybe, to Tract It. *Noûs* 37(3), 471–497.

Bickle, J. (1998). *Psychoneural Reduction: The New Wave*. MIT Press.

Bickle, J. (2019). Multiple Realizability. *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), Edward N. Zalta (ed.), Retrieved from:

https://plato.stanford.edu/archives/spr2019/entries/multiple-realizability/>.

Block, N. & J. Fodor (1972). What Psychological States Are Not. *Philosophical Review* 81, 159-181.

Bontly, T. D. (2005). Proportionality, Causation, and Exclusion. *Philosophia* 32 (1–4), 331–48.

Broad, C. D. (1925). *The Mind and its Place in Nature* (Vol. 2). Routledge and Kegan Paul.

Bunzl, M. (1979). Causal Overdetermination. *Journal of Philosophy* 134-150.

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Churchland, P. M. (1981). Eliminative Materialism and the Propositional Attitudes. *Journal of Philosophy* 78, 67–90.

Crane, T. (2001a). *Elements of Mind: an Introduction to the Philosophy of Mind.* Oxford: Oxford University Press.

Crane, T. (2001b). The Significance of Emergence. In C. Gillet & B. Loewer (Eds.), *Physicalism and its Discontents*. Cambridge University Press: 207-224.

Crane, T. (2008). Causation and Determinable Properties: on the Efficacy of Colour, Shape, And Size. In J. Hohwy & J. Kallestrup (Eds.), *Being Reduced: New Essays on Reduction, Explanation, and Causation*. Oxford University Press: 176–195.

Craver, C. (2004). Dissociable Realization and Kind Splitting. *Philosophy of Science* 71(5), 960–971.

Craver, C. (2007). *Explaining the Brain*. Calderon Press

Christensen, J. & Baysan, U. (2018). Why Incompatibilism about Mental Causation Is Incompatible with Non-Reductive Physicalism. *Inquiry,*

DOI: 10.1080/0020174X.2018.1560362

Davidson, D. (1970). Mental Events. In Lawrence Foster and J. W. Swanson (eds.), *Experience and Theory*, London: Duckworth

Demeter, T. (2009). Two Kinds of Mental Realism. *Journal for General Philosophy of Science* 40(1), 59-71.

Demeter, T. (2013). Mental Fictionalism: The Very Idea. *The Monist* 96(4), 483-504.

Dennett, D. (1991). Real Patterns. *Journal of Philosophy* 88, 27–51.

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

Dowe, P. (2001). A Counterfactual Theory of Prevention and 'Causation' by Omission. *Australasian Journal of Philosophy* 79, 216–226.

Dowe, P. (2004). Causes Are Physically Connected to Their Effects: Why Preventers and Omissions Are Not Causes. In Christopher Hitchcock (ed.), *Contemporary Debates in Philosophy of Science*. Basil Blackwell: 189-196.

Dowe, P. (2008). Causal Processes. *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.), retrieved from:

<https://plato.stanford.edu/archives/fall2008/entries/causation-process/>

Dowe, P. (2010). Proportionality and Omissions. *Analysis* 70(3), 446–51.

Ehring, D. (1996). Mental Causation, Determinables, and Property Instances. *Noûs*, *30*(4), 461–480.

Enç, B. (1983). In Defense of the Identity Theory. *Journal of Philosophy* 80, 279–298.

Fai, L. C. & Wysin, G. M. (2012). *Statistical Thermodynamics: Understanding the Properties of Macroscopic Systems*. CRC Press.

Fazekas, P. (2009). Reconsidering the Role of Bridge Laws in Inter-Theoretical Reductions. *Erkenntnis* 71(3), 303–322.

Fazekas, P. (2014). Pursuing Natural Piety: Understanding Ontological Emergence and Distinguishing it from Physicalism. *Dialectica* 68(1), 97–119.

Feigl, Herbert (1958). The 'Mental' and the 'Physical'. *Minnesota Studies in the Philosophy of Science* 2, 370-497.

Festinger, L. (1962). *A Theory of Cognitive Dissonance*. Stanford University Press.

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

Fodor, J. A. (1990). Making Mind Matter More. In *A Theory of Content and Other Essays, by Jerry Fodor*, 137-159.

van Fraassen, B. (1989). *Laws and Symmetry*, Oxford: Clarendon Press.

Funkhouser, E. (2006). The Determinable-Determinate Relation. *Noûs* 40(3), 548–569.

Funkhouser, E. (2014). *The Logical Structure of Kinds*. Oxford University Press. Retrieved from http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780198713302.001.0001/acprof-9780198713302

Gillett, C. (2003). The Metaphysics of Realization, Multiple Realizability and the Special Sciences. *Journal of Philosophy* 100(11), 591-603.

Gillett, C. (2010). Moving Beyond the Subset Model of Realization: The Problem of Qualitative Distinctness in the Metaphysics of Science. *Synthese* 177(2), 165–92.

Gillett, C. (2016). *Reduction and Emergence in Science and Philosophy*. Cambridge University Press.

Gillett, C. & Rives, B. (2005). The Nonexistence of Determinables: Or, a World of Absolute Determinates as Default Hypothesis. *Noûs* 39(3), 483–504.

Gilman, J. (2009). Chemistry and Physics of Mechanical Hardness. John Wiley & Sons.

Gibb, S. (2006). Why Davidson Is Not a Property Epiphenomenalist. *International Journal of Philosophical Studies*, 14, 407–422.

Gibb, S. (2010). Closure Principles and the Laws of Conservation of Energy and Momentum. *Dialectica 64*(3), 363–384.

Gibb, S. (2013). Mental Causation and Double Prevention. In S. C. Gibb, E. J. Lowe, & R. D. Ingthorsson (Eds.), *Mental Causation and Ontology*. Oxford University Press: 193-214.

Gibb, S. (2017). The Mental Causation Debate and Qua Problems. In Francesco Orilia & Paoletti, M. P. (eds.), *Philosophical and Scientific Perspectives on Downward Causation*. New York: Routledge: 268-277.

Glymour, C. et al. (2010). Actual Causation: A Stone Soup Essay. *Synthese* 175, 169-192.

Hall, N. (2004a). The Price of Transitivity. In Collins, J. and Hall, N. and Paul, L. A. (eds.) *Causation and Counterfactuals*. MIT Press: 281–304.

Hall, N. (2004b). Two Concepts of Causation. Collins, J. and Hall, N. and Paul, L. A. (eds.) *Causation and Counterfactuals*. MIT Press: 225–276.

Hausman, D. M. (1992). Thresholds, Transitivity, Overdetermination, and Events. *Analysis* 52(3), 159–163.

Hausman, D. M. (2005). Causal Relata: Tokens, Types, or Variables? Erkenntnis 63(1), 33–54.

Heil, J. (2013). Mental Causation. In S. C. Gibb, E. J. Lowe, & R. D. Ingthorsson (Eds.), *Mental Causation and Ontology*. Oxford University Press: 18-35.

Hendry, R. F. (2010a). Emergence vs. Reduction in Chemistry. In C. Macdonald & G. Macdonald (Eds.), *Emergence in Mind*. Oxford University Press: 205-221.

Hendry, R. F. (2010b). Ontological Reduction and Molecular Structure. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 41(2), 183–191.

Hendry, R. F. (2017). Prospects for Strong Emergence in Chemistry. In *Philosophical and Scientific Perspectives on Downward Causation*. Francesco Orilia, F. & Paoletti, M.P. New York: Routledge: 146-163.

Hempel, Carl G. (1966). *Philosophy of Natural Science*. Englewood Cliffs, Prentice-Hall.

Hitchcock, C. (2001). The Intransitivity of Causation Revealed in Equations and Graphs. *The Journal of Philosophy* 98(6), 273–299.

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

Horgan, T. (1993). From Supervenience to Superdupervenience: Meeting the Demands of a Material World. *Mind* 102(408), 555-586.

Jackson, F. & Philip Pettit (1990). Program Explanation: A General Perspective. *Analysis* 50(2), 107–117.

Jenkins, F., & White, H. (2001). *Fundamentals of Optics* (4th edition). New York: McGraw-Hill Education.

Kaiserman, A. (2017). Causes and Counterparts. *Australasian Journal of Philosophy* 95(1), 17–28.

Kim, J. (1973). Causation, Nomic Subsumption, and the Concept of Event. *Journal of Philosophy* 70(8), 217–236.

Kim, J. (1992). Multiple Realization and the Metaphysics of Reduction. *Philosophy and Phenomenological Research* 52(1), 1-26.

Kim, J. (1993). *Supervenience and Mind: Selected Philosophical Essays*. Cambridge University Press.

Kim, J. (1998). *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. MIT Press.

Kim, J. (1999). Making Sense of Emergence. *Philosophical Studies* 95(1–2): 3–36.

Kim, J. (2005). *Physicalism, or Something Near Enough*. Princeton University Press.

Kim, J. (2007). Causation and Mental Causation. In B. P. McLaughlin & J. D. Cohen (Eds.), *Contemporary Debates in Philosophy of Mind*. Blackwell: 227-242., reprinted in: Kim, J. (2010). *Essays in the Metaphysics of Mind*. Oxford University Press: 243-262.

Kim, J. (2010). *Essays in the Metaphysics of Mind*. Oxford University Press.

Kirchhoff, M. D. (2015). Species of Realization and the Free Energy Principle. *Australasian Journal of Philosophy* 93(4), 706-723.

Korb, A. (2015). *The Upward Spiral: Using Neuroscience to Reverse the Course of Depression, One Small Change at a Time*. New Harbinger Publications.

Krickel, B. (2018). *The Mechanical World: The Metaphysical Commitments of the New Mechanistic Approach*. Springer

Lewis, D. (1973). Causation. *Journal of Philosophy* 70, 556–567.

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

Lewis, D. (2000). Causation as influence. *Journal of Philosophy* 97(4), 182–197.

Lewis, D. (2004). Void and Object. In Collins, J. and Hall, N. and Paul, L. A. (eds.) *Causation and Counterfactuals*. MIT Press: 277–290.

Loewer, B. M. (2007). Mental Causation, or Something Near Enough. In Brian P. McLaughlin and J. D. Cohen (eds.), *Contemporary Debates in Philosophy of Mind*, Blackwell: 243–64.

Lowe, E. J. (1990). Conditionals, Context, and Transitivity. *Analysis* 50(2), 80–87.

Lowe, E. J. (2000). Causal Closure Principles and Emergentism. *Philosophy* 75(4), 571–585.

Lowe, E. J. (2013). Substance Causation, Powers, and Human Agency. In S. C. Gibb, E. J. Lowe, & R. D. Ingthorsson (Eds.), *Mental Causation and Ontology*. Oxford University Press: 153-172.

Mackie, J. L. (1974). *The Cement of the Universe*. Oxford, Clarendon Press.

Maslen, C. (2004). Causes, Contrasts, and the Nontransitivity of Causation. In Collins, J. and Hall, N. and Paul, L. A. (eds.) *Causation and Counterfactuals*. MIT Press: 341–358.

McDermott, M. (1995). Redundant Causation. *British Journal for the Philosophy of Science* 46(4), 523–544.

McDonnell, N. (2016). Events and Their Counterparts. *Philosophical Studies* 173(5), 1291–1308.

McDonnell, N. (2017a). Non-occurrence of Events. *Philosophy and Phenomenological Research,* DOI: 10.1111/phpr.12476, 1-17.

McDonnell, N. (2017b). Causal Exclusion and the Limits of Proportionality. *Philosophical Studies* 174(6), 1459–1474.

McGrath, S. (2005). Causation by Omission: A Dilemma. *Philosophical Studies* 123(1–2), 125–148.

Menzies, P. (2003). The Causal Efficacy of Mental States. In S. Walter & H.-D. Heckmann (Eds.), *Physicalism and Mental Causation*. Imprint Academic: 195-223.

Menzies, P. (2007). Mental Causation on the Program Model. In G. Brennan, R. Goodin, F. Jackson, & M. Smith (Eds.), *Common Minds: Themes From the Philosophy of Philip Pettit*. Oxford University Press: 28-54.

Menzies, P. (2008). The Exclusion Problem, the Determination Relation, and Contrastive Causation. In J. Hohwy & J. Kallestrup (Eds.), *Being Reduced: New Essays on Reduction, Explanation, and Causation*. Oxford University Press: 196-217.

Menzies, P. & List, C. (2009). Nonreductive Physicalism and the Limits of the Exclusion Principle. *Journal of Philosophy* 106(9), 475–502.

Menzies, P. & List, C. (2010). The Causal Autonomy of the Special Sciences. In C. Mcdonald & G. Mcdonald (Eds.), *Emergence in Mind*. Oxford University Press: 108-128.

Menzies, P. (2013). Mental Causation in the Physical World. In S. C. Gibb & R. Ingthorsson (Eds.), *Mental Causation and Ontology*. Oxford University Press: 58-87.

Menzies, P. (2017). Counterfactual Theories of Causation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2017 edition). Metaphysics Research Lab, Stanford University. Retrieved from:

<https://plato.stanford.edu/archives/win2017/entries/causation-counterfactual/>

Moore, M. S. (2009). *Causation and Responsibility: an Essay in Law, Morals, and Metaphysics*. Oxford University Press.

Mumford, S. & Anjum, R. L. (2011). *Getting Causes from Powers*. Oxford University Press, USA.

Musallam S., B.D. Corneil, B. Greger, H. Scherberger, & R.A. Andersen (2004). Cognitive Control Signals for Neural Prosthetics. *Science*, cccv (July): 258

Nagel, E. (1961). *The Structure of Science. Problems in the Logic of Explanation*, New York: Harcourt, Brace & World, Inc.

Nagel, E. (1970). Issues in the Logic of Reductive Explanations. In H.E. Kiefer & K.M. Munitz (eds.), *Mind, Science, and History.* Albany, NY: SUNY Press: 117–137.

Northcott, R. (2008). Causation and Contrast Classes. *Philosophical Studies* 139(1): 111–23.

O'Connor, T. & Wong, Hong Yu (2015). Emergent Properties. *The Stanford Encyclopedia of Philosophy* (Summer 2015 Edition), Edward N. Zalta (ed.), retrieved from:

<https://plato.stanford.edu/archives/sum2015/entries/properties-emergent/>

Oppenheim, P. & Putnam, H. (1958). Unity of Science as a Working Hypothesis. *Minnesota Studies in the Philosophy of Science* 2:3-36.

Paul, L. A. (2004). Aspect Causation. In Collins, J. and Hall, N. and Paul, L. A. (eds.) *Causation and Counterfactuals*. MIT Press: 205–25.

Papineau, D. (2001). The Rise of Physicalism. In C. Gillett & B. M. Loewer (Eds.), *Physicalism and its Discontents*. Cambridge University Press: 3-36.

Papineau, D. (2008). Must a Physicalist be a Microphysicalist. In J. Hohwy & J. Kallestrup (Eds.), *Being Reduced: New Essays on Reduction*, *Explanation and Causation.* Oxford: Oxford University Press: 126-148.

Pernu, T. (2014). Causal Exclusion and Multiple Realizations. *Topoi* 33 (2), 525–30.

Polger, T. (2004). *Natural minds*. Cambridge, MIT Press.

Polger, T. (2008). Two Confusions Concerning Multiple Realization. *Philosophy of Science* 75 (5), 537–47.

Polger, T. & L. Shapiro (2016). *The Multiple Realization Book*. Oxford University Press UK.

Putnam, H. (1960). Minds and Machines. In S. Hook (ed.), *Dimensions of Mind: A Symposium*, New York: Collier: 138–164; Reprinted in Putnam (1975): 362–386.

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

Putnam, H. (1975). *Philosophical Papers: Volume 2, Mind, Language and Reality*. Cambridge: Cambridge University Press

Raatikainen, P. (2010). Causation, Exclusion, and the Special Sciences. *Erkenntnis* 73(3), 349–363.

Reiss, J. (2013a). Contextualising Causation Part I. *Philosophy Compass* 8(11), 1066–75.

Reiss, J. (2013b). Contextualising Causation Part II. *Philosophy Compass* 8(11), 1076–90.

Richardson, R. (1982). How not to reduce a functional psychology? *Philosophy of science* 49, 125–137.

Russo, A. (2016). Kim's Dilemma: Why Mental Causation Is Not Productive. *Synthese* 193 (7): 2185–2203. https://doi.org/10.1007/s11229-015-0837-7.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press.

Schaffer, J. (2003). Overdetermining Causes. *Philosophical Studies* 114(1–2), 23–45.

Schaffer, J. (2004). Causes Need Not Be Physically Connected to Their Effects: The Case for Negative Causation. In Christopher Hitchcock (ed.), *Contemporary Debates in Philosophy of Science*. Basil Blackwell: 197–216.

Schaffer, J. (2005). Contrastive Causation. *Philosophical Review*, 114(3), 297–328.

Schaffer, J. (2013). Causal Contextualisms. In Martijn Blaauw (ed.), *Contrastivism in Philosophy: New Perspectives*. Routledge: 35–64.

Schiffer, S. (1992). Belief Ascription. *Journal of Philosophy* 89(10), 499–521.

Schiffer, S. (1995). Descriptions, Indexicals, and Belief Reports: Some Dilemmas (but not the Ones You Expect). *Mind* 104(413), 107–131.

Schrödinger, E. (1944). What Is Life? The Physical Aspect of the Living Cell. Cambridge University Press.

Shapiro, L. (2000). Multiple Realizations. *Journal of Philosophy* 97(12), 635–654.

Shapiro, L. (2004). *The Mind Incarnate*. Cambridge, Mass: MIT Press.

Shapiro, L. (2012). Mental Manipulations and the Problem of Causal Exclusion. *Australasian Journal of Philosophy* 90(3), 507–524.

Shapiro, L., & Sober E. (2012). Against Proportionality. *Analysis* 72(1), 89–93.

Sider, T. (2003). Review: What's so Bad about Overdetermination? *Philosophy and Phenomenological Research* 67(3), 719–726.

Shoemaker, S. (1981). Some Varieties of Functionalism. *Philosophical Topics* 12, 93–119.

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

Shoemaker, S. (2013). Physical Realization without Preemption. In S. C. Gibb, E. J. Lowe, & R. D. Ingthorsson (Eds.), *Mental Causation and Ontology*. Oxford University Press: 35-56.

Sklar, Lawrence (1993). *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics.* Cambridge University Press

Sklar, Lawrence (2015). Philosophy of Statistical Mechanics. *The Stanford Encyclopedia of Philosophy* (Fall 2015 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2015/entries/statphys-statmech/>.

Sober, E. (1999). The Multiple Realizability Argument against Reductionism. *Philosophy of Science* 66(4), 542–564.

Smart, J. (1959). Sensations and Brain Processes. *Philosophical Review* 68, 141–156.

Wilson, J. M. (2005). Supervenience-Based Formulations of Physicalism. *Noûs* 39(3), 426–59.

Wilson, J. M. (2009). Determination, Realization and Mental Causation. *Philosophical Studies* 145(1), 149–169.

Wilson, J. (2010). Non-reductive Physicalism and Degrees of Freedom. *The British Journal for the Philosophy of Science* 61, 279-311.

Woodward, J. (2000). Explanation and Invariance in the Special Sciences. *British Journal for the Philosophy of Science* 51(2), 197–254.

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

Woodward, James (2006) Sensitive and Insensitive Causation. Philosophical Review 115(3), 273–316.

Woodward, J. (2008). Mental Causation and Neural Mechanisms. In J. Hohwy & J. Kallestrup (Eds.), *Being Reduced: New Essays on Reduction, Explanation, and Causation*. Oxford University Press: 218-262.

Woodward, J. (2010). Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation. *Biology and Philosophy* 25(3), 287–318.

Woodward, J. (2015). Interventionism and Causal Exclusion. *Philosophy and Phenomenological Research* 91(2), 303–47.

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

Yablo S. (2004). Advertisement for a Sketch of an Outline of a Prototheory of Causation. In Collins, J. and Hall, N. and Paul, L. A. (eds.) *Causation and Counterfactuals*. MIT Press: 119–138.