Models of downward causation

Max Kistler

Models of downward causation

in : Jan Voosholz & Markus Gabriel (eds.), *Top-Down Causation and Emergence.* Synthese Library, vol. 439, Cham: Springer, 2021, p. 305-326.

Max Kistler
Université Paris 1 Panthéon-Sorbonne and IHPST
mkistler@univ-paris1.fr

Abstract

Two conceptual frameworks – in terms of phase space and in terms of structural equations – are sketched, in which downward causal influence of higher-level features on lower-level features is possible. The "Exclusion" principle, which is a crucial premise of the argument against the possibility of downward causation, is false in models constructed within both frameworks. Both frameworks can be supplemented with conceptual tools that make it possible to explain why downward causal influence is not only conceivable and compatible with the "Closure" principle, but also why it is often relevant to causally explain facts in terms of downward causal influence. It is briefly shown that 1) the analysis of downward causation in the two frameworks complements Bennett's (2003) analysis of overdetermination, 2) the analysis does not entail the failure of the "Closure" principle and 3) it does not require the postulate of synchronic downward causation.

1. Introduction

The idea that the mind causally influences the physical world is often claimed to be incompatible with physicalism. Physicalism is the doctrine according to which 1) everything is either physical or exclusively composed of physical parts, and 2) all properties of all objects supervene on the physical properties of those objects. There are stronger versions of physicalism, such as reductionism and eliminativism[1]. According to the construal of physicalism in terms of supervenience, mental properties, events, and processes are distinct from physical properties, events, and processes. Thus the question arises whether mental events, or indeed any kind of higher-level events, in a sense of "higher-level" to be specified shortly, can influence[2] physical events.

I will address this question within the framework of physicalism construed in terms of supervenience. Kim (1998; 2005) has developed a strong argument according to which downward causation is incompatible with this construal of physicalism: It can never be literally correct that a mental event causes a physical event, because the causes of physical events are always exclusively physical[3].

[1] According to the former, all real properties are reducible to physical properties, and according to the latter, strictly speaking, there are only physical properties. According to these strong forms of physicalism, the question whether the mind influences the physical world does not really arise, either because there is no mind (eliminativism) or because the mind is itself physical (reductionism).

[2] When I speak of influence, I always mean causal influence. I am using "cause" and "influence" as synonymous, the only difference being stylistic.

[3] Here I use the term "event" in Kim's sense, as the instantiation of a property by some object at some time. I shall later express the same question as being (in the model of state spaces, see below) 1) whether

This result is troubling because it seems obvious that its conclusion is wrong and that our minds do influence physical events. Here is a case of such a "downward" influence. My thoughts about the relation of the mind to the body, together with my desire to make my thoughts publicly known, cause my fingers to move over the keybord and indirectly cause words to appear on the screen of my computer. Downward influence from psychological on physiological features of persons can be studied experimentally. Psychotherapy has been found to influence brain function in many psychiatric disorders (Barsaglini 2014). Obsessive-compulsive disorder (OCD) is correlated with hypermetabolism in, among other regions of the brain, the right caudate nucleus. It has been found (Baxter et al. 1992) that behavioral therapy of patients suffering from OCD results in decreased rates of glucose metabolism in the head of the right caudate nucleus of their brains.

There are at least two interpretations of what it means to characterize such influences as being "downward". The *reductionist* interpretation uses the hypothesis of a hierarchy among the sciences that is structured by partial and local reduction relations between theories[4]. Few philosophers today think that the history of science tends towards unification, in the sense that all sciences tend to get reduced, directly or indirectly, to physics (Dupre 1993, Cartwright 1999). However, it is not controversial that there are cases of successful reductive explanations of certain laws or theories. These successful cases of local reductions can justify the hypothesis that the sciences concerned by those reductions are ordered in a partial hierarchy. Thermodynamics has been (in part) reduced to statistical physics and certain simple forms of learning by classical conditioning have been reduced to neuroscience. By virtue of these reductions, statistical physics lies lower in the hierarchy of the sciences than thermodynamics, and neuroscience lower than psychology. The causal influence of psychotherapy on the glucose metabolism in the caudate nucleus of a patient's brain is a case of downward causation in this hierarchical sense because psychotherapy modifies the properties of persons at the level of psychology, whereas the features of glucose metabolism in the caudate nucleus belong to neurophysiology, which lies at a lower level than psychology.

A second interpretation of the concept of downward causation derives from the distinction between properties of a whole object and properties of the object's parts. In a *mereological* sense of levels, the properties of a whole object lie at a higher level than the properties of its parts[5]. Levels can be locally defined and structured in local hierarchies by mereology. Levels in the reductionist sense, as defined by partial and local reductions, do not coincide with levels in the mereological sense[6]. The mouse's

---

mental facts can be causally responsible for physical facts or (in the model of structural equations, see below) 2) whether mental properties can influence physical properties.

[4] This *reductionist* concept of levels corresponds to Craver's "levels of science" (Craver 2007, p. 172), whereas the *mereological* concept of levels (see below) corresponds to Craver's "levels of composition" (Craver 2007, p. 184). For the relevant concept of reduction, see Nagel (1961) and Schaffner (1967).

[5] Only some parts have "constitutive explanatory relevance" (Craver 2007, p. 140) in the context of mechanistic explanation. In an "interlevel experiment" of the "top-down" variety (Craver 2007, p. 145), an experimenter manipulates a property of a system and observes the downward effect of this manipulation at the level of such a constitutively relevant part. However, the meaning of the term "downward" need not be restricted to constitutively relevant parts. Downward causation in the mereological sense exists within physics: Heating (i.e. a modification of a macroscopic property) of piece of Nickel modifies the properties of microscopic parts of that piece of Nickel (Kistler 2017).

[6] The two concepts of level do not have the same extension. 1) Many sciences study objects that lie at different mereological levels, in particular because they study, as does neuroscience, whole mechanisms as well as their parts; 2) many objects are studied by sciences that lie at different reductionist levels: Proteins are studied by physics (e.g., in X-ray cristallography), chemistry, and physiology.

perceiving the cat approaching it causes the mouse to flee: here, the cause (the mouse's perceiving) and effect (the mouse's fleeing) lie at the same level (in both the hierarchical and the mereological sense) because both events are changes in properties of the same object, i.e. the mouse. However, the same perception also causes a given determinate muscle fiber in the mouse's left rear leg to contract. The perception's causing the contraction of the fiber in the mouse's leg is downward causation in the mereological sense, because the muscle fiber is a part of the mouse, and thus, its contraction is a lower-level property compared to the property of perceiving the cat. It is also downward causation in the reductionist sense, because perception belongs to psychology, the contraction of muscle fibers belongs to physiology, and physiology lies at a lower reductionist level than psychology. For lack of space, I will concentrate in what follows on downward causation in the reductionist sense.

The levels required to analyze the concepts of same-level, downward, and upward causation are always locally defined, with respect to the objects or systems involved in the causal interactions under enquiry. It is not plausible that locally and partially defined reductionist levels (or, for that matter, mereological levels) can somehow be merged into a unique global hierarchy of levels[7]. However, the analysis of *local* downward causation does not require the existence of such a unique global hierarchy of levels.

We have analyzed one part of what it means to say that some influence is a case of downward causation, by explaining what it means to say that the effect lies at a lower level than the cause. It remains to be analyzed what is meant by "cause" or "causal influence". In what follows, I sketch two models of causation that provide frameworks for making sense of downward causation[8]. The first model is based on the notion of phase space. The second model elaborates the framework of structural equations. Within each of these models, we will evaluate the argument against downward causation in the following form[9].

1) (Closure) *The causal closure of the physical domain*. If a system p has at t_1 a physical property R_1, then there is, at each time t_0 preceding t_1, a physical property N such that the fact that p has N at t_0 is causally responsible for the fact that p has R_1 at t_1.

2) (Exclusion) *Principle of causal exclusion*[10]. If the fact that p has N at t_0 is causally responsible for the fact that p has R_1 at t_1, there cannot be any property M

---

[7] For reasons against the existence of a unique global hierarchy of levels, see Eronen (2013; 2019) and Voosholz (2020).

[8] In (Kistler 2017), I have explored whether it is possible to make sense of downward causation by modifying the framework of analyzing causal influence in terms of interventions (Woodward 2003) and using it as a complement to the account of causation in terms of transference (Kistler 2006a; 2013). In (Kistler 2017), downward causation is interpreted in terms of the mereological notion of levels, according to which a property P is at a higher level than property Q if and only if Q characterizes a proper part of the object characterized by P.

[9] There are many versions of this argument and of its premises "Closure" and "Exclusion". For an overview, see (Robb and Heil 2018). Kim's own version of the argument aims at establishing that mental properties are not downward causes by construing them in terms of higher-*order* predicates (Kim 1998, p. 83). This leaves open the question whether there is downward causation by higher-*level* properties as defined in this chapter. For critical analyses of Kim's construal of mental properties in terms of higher-order predicates and his use of the distinction between levels and orders in the context of his argument against downward causation, see (Kistler 2006b) and (Gozzano 2009).

[10] This principle has the consequence that one complete *causal explanation* of the fact that p has R_1 at t_1 by the fact that p has N at t_0 excludes other independent causal explanations of the fact that p has R_1 at t_1 by other facts about p at t_0. However, our "Exclusion" principle is weaker than Kim's "principle of explanatory exclusion": "There can be no more than a single complete and independent explanation of any one event"

distinct from N, and in particular no property M at some level higher than N, such that p has M at t_0 and such that the fact that p has M at t_0 is also causally responsible for the fact that p has R_1 at t_1.

3) *No downward causation*. Therefore, no higher-level level property M is such that the fact that p has M at t_0 is causally responsible for the fact that p has R_1 at t_1.

2. Downward causation in the framework of phase space

The framework of dynamical systems theory provides one way to think about causal influence (Hitchcock 2012)[11]. The decision of a person to raise her arm causes her arm to rise. This is downward causation both in the reductionist sense, because the cause is psychological and the effect physiological, and in the mereological sense, because the arm is a part of the person taking the decision: Let us assume that a person is identical with her body[12]. We can conceive the body of person p as a physical system whose state at time t_0 can be represented as a point p_0 in the body's state space. The number of dimensions of this state space equals the number of degrees of freedom the system possesses. Each degree of freedom corresponds to a way in which the system can change. A classical mechanical system of n particles has 6n degrees of freedom, 3 degrees of freedom for the position of each particle and 3 degrees of freedom for the velocity of each particle, in each of the 3 Cartesian dimensions. The space state of a system as complex as a human body in interaction with its environment, has a very large number of dimensions.

Here is a way of representing causality in this framework. Take as cause the state p_0 of the system at t_0, represented by a point in the system's state space. The system evolves through time according to the dynamical equations governing the system, following what is called its trajectory, which can be represented by its position in state space as a function of time. A point p_i on that trajectory at time t_i is a cause of all points p_j on the trajectory at later times $t_j > t_i$, and an effect of all points p_k on the trajectory at earlier times $t_k < t_i$. [13]

The question of what was the cause, at t_0, of p's raising her arm at time t_1, can be interpreted in at least two ways. On one interpretation, the effect is an event in the sense of something that is identified as the whole content of the region of space-time where it is happening (Quine 1960; 1995; Kistler 2006a). When we ask for the cause of the state p_1 of the person at t_1, the answer can be found by following the trajectory of her body, considered as a physical system, backwards in time, up to t_0. According to this interpretation, the state p_0 of the person at time t_0 is the cause of her state p_1 at the later time t_1.

However, when we ask questions about causes, we are typically interested in answers that give us more information than that. We want to know, not only what, at t_0, was the cause of p_1 at t_1, but also what it was about that cause that is responsible for the

---

(Kim 1988a, p. 233). He later calls this stronger principle the "principle of determinative/generative exclusion" (Kim 2015, p. 17). This strong principle is not plausible because one fact can have both a causal and a non-causal explanation, which can be independent of each other (Kistler 2006b).

[11] Hitchcock calls this model of causation "Laplacean causation" (Hitchcock 2012, p. 46).

[12] This assumption is of course controversial. Cf. Lowe (2000a, chap. 2).

[13] We shall see that an analysis of causal relations in the conceptual framework of dynamical systems is compatible with more traditional conceptions of causation as a relation between two events or between facts about substances. Causal statements expressed in the language of points, regions and trajectories of phase space can be translated in the language of events and facts, and vice versa.

fact that p_1 has a certain property[14]. We want to know, not only what was the trajectory of p up to p_1, but also what made her raise her arm at t_1, or, in other words, what fact about p_0 was *causally responsible* for the fact that she raised her arm at t_1?[15] Why would we want to have such additional information? One reason is that it gives us the means to generalize to other similar cases. A second reason, linked to the first, is that such information has counterfactual implications that we can use for planning actions. If what is causally responsible is the fact that p took at t_0 the decision to raise her arm at t_1, we know that, in general, had she not taken that decision, her arm would not have risen. We know also that, if she took a similar decision in other circumstances, her arm would probably rise. When we ask what it was about p_0 that made it the case that p_1 is a state of a person raising her arm, we can interpret this question as the application to this particular case of the more general question: For any person p and any time t_0, what is it about p at t_0 that makes it the case that p raises her arm at $t_1>t_0$?

At this point, the issue arises whether we can make sense of the possibility that psychological properties causally influence the body. Such downward influence is controversial because there always seem to be several complete causes (and therefore causal explanations) of the bodily movements that are part of actions. Even if we accept that the fact that p raised her arm at t_1 is caused by her decision at t_0 to raise her arm, the same fact also seems to be the effect of a physiological fact: that the relevant part of her brain showed a specific pattern of activity at t_0.

To address the question of how these facts about p at t_0 are related to each other in the framework of dynamical systems, we need the conceptual tool of *regions* of state space. A predicate describing the system determines a whole region of the state space of the system, consisting of all states in which the system satisfies the predicate. If p_1 is the point representing p at t_1, the predicate "raises her arm" corresponds to a region R_1 that includes p_1. All points in R_1 represent possible states of the system in which the person raises her arm. Points outside R_1 represent possible states in which she does not raise her arm.

Saying what made it the case that p raised her arm at t_1 requires finding a feature (or property) of p's state p_0 at t_0 such that the fact that p had property P_0 at t_0 is causally responsible for the fact that p had R_1 at t_1. The property P_0 must satisfy the following requirement: P_0 must be such that for all points of the state space within P_0, these points lie on trajectories that lie, at t_1, within R_1[16].

Here is a way of representing the search for the causally relevant property P_0. If we trace back in time up to t_0 the trajectory of all points within region R_1, we obtain a region R_0, which can be called the *projective state* of R_1 at t_0, or the reverse image of R_1 at t_0. R_0 contains all and only those points in state space at t_0 that represent possible states that evolve towards a state within R_1. The more the system is sensitive to initial conditions, or in other words "chaotic", the more R_0 will be spread out in state space.

---

[14] On the distinction between these two questions, the corresponding two sorts of causal information, and the metaphysical and logical relations between causal relations between events and relations of causal responsibility between facts, see Kistler (1999; 2006a; 2014).

[15] According to some conceptions of actions, the bodily movement of the arm that rises is only part of the action. According to Dretske (1988), the action consists in the whole process starting with the decision and ending with the bodily movement.

[16] Knowing that p has P_0 at t_0 also provides someone who doesn't yet know what p will do at t_1 with the means to predict that she will do R_1 at t_1. It also justifies the counterfactual judgment about a situation at t_2 in which p does *not* have P_0, that if p had had P_0 at t_2, she would have had R_1 (i.e. raised her hand) a little later.

Let us suppose that there is a predicate P, in science or common sense, whose extension at t_0 in the space state of the system is entirely included within R_0. The fact that p has P at t_0 is sufficient for p's trajectory lying within R_0 at t_0. R_0 is the projective state of R_1, so that the fact that p is P is also sufficient for its having R_1 at t_1. In our example, if R_0 is the projective state of the fact that p raised her arm (R_1) at t_1, and P is the predicate "decides to raise her arm", the fact that p decides at t_0 to raise her arm (which is represented by the fact that the trajectory of p lies within P at t_0) is causally responsible for the fact that she raised her arm at t_1 (which is represented by the fact that her state lies within R_1 at t_1). Insofar as P is a psychological predicate and R_1 is a region corresponding to a physiological predicate, we have a case of downward causation. Possessing P at t_0 is sufficient for possessing R_1 at t_1: All points within P lie on trajectories that cross R_1 at t_1. However, it is not necessary. Many points within R_0 lie outside P. Such points represent states of p at which p does not possess P but nevertheless lies on a trajectory that leads through R_1 at t_1.

There may also be predicates such as P*, sketched in figure 1, which is *almost* necessary and *almost* sufficient for R_1, in the sense that it comes close to picking out the same region as the projective state R_0. Almost all states that satisfy P* are on trajectories that lead through R_1. This means that P* is almost sufficient for R_1, or is sufficient with exceptions for R_1. And almost all states that lie outside of P*, i.e. that do not have P* are such that their trajectories do not lead through R_1. This means that P* is almost necessary for R_1, or is necessary for R_1 with exceptions.

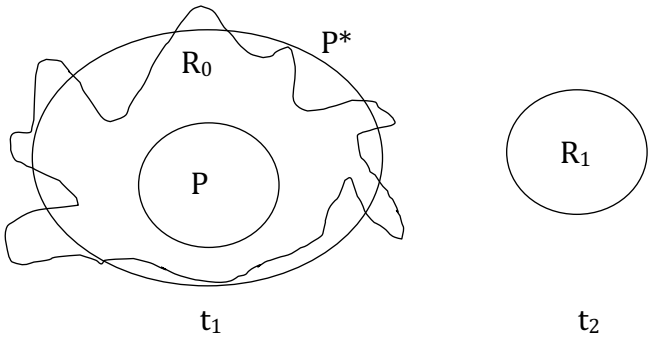


Fig. 1. R_0 is the projective state of R_1, where R_1 represents the region of the state space of a person p corresponding to the extension of the predicate "x raises her arm". Having P at t_0 is a sufficient (but not necessary) condition for having R_1 at t_1: all possible states of the system represented by points within P have trajectories that run through R_1. P* is *almost* necessary and *almost* sufficient for R_1. The figure is similar to fig. 3 in Hitchcock (2012, p. 49).

We are now in a position to evaluate the argument against the possibility of downward causation.

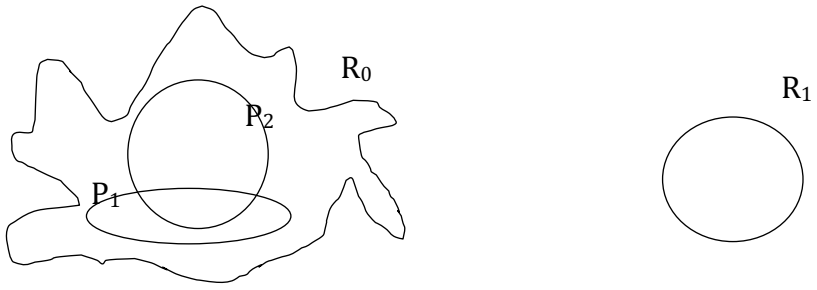
1) (Closure) *The causal closure of the physical domain*. If a system p has at t_1 a physical property R_1, then there is, at each time t_0 preceding t_1, a physical property N such that the fact that p has N at t_0 is causally responsible for the fact that p has R_1 at t_1.

2) (Exclusion): *Principle of causal exclusion*. If the fact that p has N at t_0 is causally responsible for the fact that p has R_1 at t_1, there cannot be any property M distinct from N, and in particular no property M at some level higher than N, such that p has M at t_0 and such that the fact that p has M at t_0 is also causally responsible for the fact that p has R_1 at t_1.

3) *No downward causation*. Therefore, no higher-level level property M is such that the fact that p has M at t_0 is causally responsible for the fact that p has R_1 at t_1.

Concerning premise (1), Hitchcock (2012) argues that there seems to be no principled reason for thinking that there will always be a physical property that plays the role of P (or P*, for that matter). "What reason do we have for thinking that P will correspond to some physical property? That is, why think that the similarity shared by all of the states in P will involve the values of some simply specifiable physical parameter – e.g. having a kinetic energy or angular momentum within some specific range?" (Hitchcock 2012, p. 50). However, I think that we can interpret Closure in this framework so as to see why it is generally taken to be true by physicalists. Hitchcock's point seems plausible only insofar as we mean by "physical property" a property that corresponds to a simple predicate in physical vocabulary[17]. However, there is a more charitable interpretation of Closure. The conjunction of the description of the positions and velocities of all atoms constituting system p may be longer than what can possibly be expressed within the limited time and space available to us. However, this is no reason to deny that the corresponding physical property exists, in the sense that it corresponds to a well-defined region within R_0. Let us interpret the condition that N is a physical property of p at t_0 as meaning that there is a region including p_0 that corresponds to the physical properties of all physical constituents of system p at t_0 (even though these properties cannot be actually described given human limitations). In this "metaphysical" interpretation, premise (1) seems plausible and can be accepted.

According to Exclusion (premise (2)), there cannot be two different properties M and N such that p has both N and M at t_0, and such that each of the facts that p has N and that p has M is by itself causally responsible for the fact that p has R_1 at t_1. The Exclusion principle is often stated with the proviso that there are exceptional cases of "genuine overdetermination".[18] Genuine cases of overdetermination are situations in which "R_1 is somehow being caused twice over" (Hitchcock 2012, p. 50). This is not the case here. Say P_1 is a physical property and P_2 a psychological property, both with extensions contained within R_0, as sketched in figure 2. There seems to be nothing problematic in allowing that facts involving both P_1 and P_2 are causally responsible for the fact that p is R_1 at t_1. The causal responsibility of the fact that p had P_2 at t_0 for the physical fact that p has R_1 at t_1 is downward. Its causal responsibility is not threatened by the existence of physical facts about p at t_0 (that p has P_1 at t_0) that are also causal responsible for the fact that p has R_1 at t_1.



[17] The interpretation of Closure according to which there is at each time t_0 preceding t_1, a cause N that is a sufficient condition for R_1 and that can be described with a short expression in the vocabulary of physics, might be called, with Flanagan (1992, p. 98) "linguistic physicalism". Such an interpretation is much stronger and less plausible than the metaphysical interpretation suggested in the text.

[18] "If an event *e* has a sufficient cause *c* at *t*, no event at *t* distinct from *c* can be a cause of *e* (unless this is a genuine case of causal overdetermination)" (Kim 2005, p. 17). What Kim calls "event *e*" is in our terminology the fact that p has property R_1 at t_1. For Kim's notion of event, see Kim (1973).

t_0 t_1

Fig. 2. R_0 is the projective state of R_1, where R_1 represents the region of the state space of a person p corresponding to the extension of the predicate "x raises her hand". P_1 and P_2 represent two properties of p at t_0 such that both are sufficient conditions for having R_1 at t_1: all possible states of the system represented by points within both P_1 and P_2 have trajectories that run through R_1. The figure is similar to fig. 4 in Hitchcock (2012, p. 50).

This is enough to show that, in the dynamical systems framework, the existence of downward causation is conceivable and plausible, in the sense that psychological features of a person may causally influence her physiological features. However, the framework also shows that there can be situations where both a physical (or physiological) and a psychological causal *explanation* is available for some physiological fact, and where the psychological (downward) causal explanation is the more *relevant* one[19]. Here is why. R_1 corresponds to a coarse grained predicate, "raising one's arm". There are many different ways of performing this act, differing along many dimensions, such as the particular angle of the elbow that is reached at the end of the movement, the speed of the movement, and the exact states of all muscle fibers that are constituents of the movement. It is plausible that the causes of these physiologically different processes will be spread out in state space and that any predicate whose extension comes close to R_0 would be a very long disjunction. The extension of some specific detailed description P_k of the cause of one particular type of raising one's arm will cover a very small subregion within R_0. The corresponding physical property P_k is sufficient but not necessary for R_1. However, the extension of the psychological predicate "decides to raise her arm" might resemble P* in diagram 1 above. It is plausible that most arm raisings are caused by decisions to do so – only few small subregions of R_0 lie outside P* - and most decisions to do so lead to arm raisings – only a few small subregions of P* lie outside R_0. If that reflects the situation, P* comes closer to specifying a necessary and sufficient condition for R_1 than any predicate P_k in physiological or physical vocabulary. P* comes closer than any such P_k to expressing a difference-maker for arm-raising: A condition X is a difference-maker for condition Y if X is necessary and sufficient for Y, so that the trajectories of all states within X run through Y, and no trajectories of states outside X run through Y. A causal explanation of R_1 in terms of the difference-making psychological property P* is more relevant and therefore preferable to a causal explanation in terms of a sufficient but not necessary physiological or physical condition P_k.

3. Downward causation in the framework of structural equations

Recent years have seen much work dedicated to developing the method of representing the search for causes by models using structural equations (Pearl 2000, Spirtes et al. 2000). This method constitutes the conceptual basis of algorithms that are successful in the discovery of causal structure, especially in complex systems studied by sciences such as economics or epidemiology. I can here only present the fundamental conceptual structure of the formalism, following Halpern (2000) and limiting myself to

[19] Hitchcock (2012) doesn't mention this consideration in his discussion of Kim's argument of causal-explanatory exclusion in the framework of dynamical systems.

deterministic models with a finite number of dimensions. The construction of a structural equations (SE) model requires three steps.

In the first step, the system under study is represented by a finite set of variables, which correspond to the predicates characterizing the features of the system. There are two sorts of variables. Endogenous variables are such that their values are determined by other variables within the model, whereas the values of exogenous variables are determined in a way that is independent of other variables of the system. The structural equations describe the functional dependence of the endogenous variables on other (endogenous and exogenous) variables in the model. The first step of construction of the model consists in determining a signature S. S is a triple (U,V,F), where U is the set of exogenous variables, V the set of endogenous variables, and F a set of functions associating with each variable Y a non-empty set F(Y) of values of Y. F(Y) is the range of the values of Y.

The situation of arm-raising can be represented with the following simple signature: U contains the variable D representing p's decision to raise her arm, V contains the variable R representing the raising of p's arm, the function F(D) maps D on (0,1) and F(R) maps R on (0,1). The fact that D takes the value 1 represents the fact that person p takes the decision to raise her arm, whereas D=0 represents the fact that p doesn't take that decision. Similarly for R: The fact that R takes the value 1 represents the fact that p raises her arm, whereas R=0 represents the fact that she doesn't.

The second step introduces the formal means representing the dependence relations among the variables introduced in the first step. A causal model consists in a pair (S,E), whose members are 1) the signature S and 2) a set of structural equations. There is exactly one structural equation for each endogenous variable $X \in V$, expressing the value of X as a function of all other variables in $U \cup V$.

The simplest model we can build for our example consists in the equation R=D. It expresses the assumption that the question of whether (R=1) or not (R=0) p raises her arm is perfectly determined by a single factor, i.e. the value of D. If D=0, then R=0, and if D=1, then R=1. This model is of course very much oversimplified because it does not represent the many other factors that may influence whether a person raises her arm at a given moment. The decision may be overrun by interference of external or internal factors, so that the arm does not rise although the decision has been taken (D=1 but R=0), and the arm may raise for reasons independent of the decision, so that R=1 although D=0.

A third step consists in an assignment, which represents the application of the causal model to an actual situation. An assignment consists in attributing a value to each of the external variables. In our model, the assignment simply consists in attributing to the exogenous variable D one of its two values, 1 for situations in which the decision is taken and 0 for situations in which it is not taken.

The minimal model I have used as an example to introduce the SE formalism already shows, albeit in a rather trivial sense, that downward causation is conceivable in this framework. The equation R=D represents a downward influence because D is a psychological predicate and R a physiological one. In order to evaluate the argument against the possibility of downward causation that relies on the premises of Closure and Exclusion, we need to construct a slightly more complex model. Let N be the only exogenous variable in U, where N=1 represents one particular state of activity of the neurons in p's brain at t_0, and N=0 all other states of activity of the brain. As above, let D=1 represent the fact that p takes the decision to raise her arm, whereas D=0 represents the fact that she doesn't take that decision. Let V contain, as endogenous

variables, D and R (R=0 and R=1 are interpreted as above). Any adequate model should respect the local supervenience[20] of psychological variables over neurological variables, so that D must be a function of N. In other words, it must be excluded that one value of N is mapped to different values of D. The simplest structural equation would be D=N. However, in any realistic situation, D as defined will not be a function of N as defined. It is plausible that different values of D will be associated with N=0. If N=0, the brain is not in exactly the neurological state represented by N=1 but it may be in a state that differs from such a state only slightly, maybe by the activity of a single neuron in an area with large redundancy. In that situation, we will have N=0 and D=1 because that slight neural difference makes no psychological difference so that a person in that situation would still take decision D. However, there will be other states that are represented by the same value N=0 which differ a lot from the situation with N=1, so that no decision is taken, and D=0. Thus there is no functional dependence of D on N, as sketched in fig. 3.

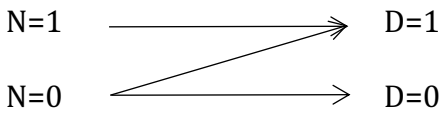


Fig. 3. One value of N is mapped on two different values of D: D is not a function of N.

A slightly more complex choice of signature can be used to model the situation in a more realistic way. Let N have 3 values instead of 2. Let N take value N=1, as before, when the brain of p is in a perfectly specific state of neural activation, which happens to be p's state at t_0. Let N take value N=2 when the brain is not in exactly the state corresponding to N=1 but in some other state that is also in the supervenience base of D, so that D=1 for both N=1 and N=2. Finally, let N take the value N=3 when the brain is in a state that is not in the supervenience base of D, so that D=0 whenever N=3. In this situation we can define the SE for D=F(N) with

F(N=1)=F(N=2)=1 and F(N=3)=0.

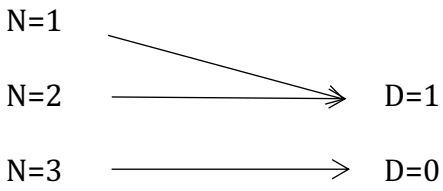


Fig. 4. No value of N is mapped on two different values of D: D is a function of N.

[20] According to externalism, the content of some mental states of a person depends on her social (Burge 1979) or physical (Putnam 1975) environment. Such mental states do not *locally* supervene on the person's physical state. We can leave such states to one side here. It can be doubted whether mental states that do not supervene locally can play a role in causing behavior. The challenge we are addressing in the present paper is whether one can make sense of the idea that mental states that do locally supervene can influence behavior, although the same behavior is also influenced by underlying physical states of the person.

Furthermore, let us suppose as before that D is necessary and sufficient for R, so that R=F(D), with F(D=0)=0 and F(D=1)=1.

This model contains two structural equations: D=F(N) represents the *non-causal* dependence relation between a psychological property and the underlying neurological property in its supervenience base[21]. R=F(D) represents the *causal* dependence between the decision D and the bodily movement R.

There is downward causation in this model, just as in the simpler first model. Indeed, the only SE representing a causal dependence represents a downward influence, of a psychological variable (D) on a variable representing bodily movement (R). However, this second model is rich enough for the challenge of Exclusion to arise.

The model respects "Closure", the first premise of the argument against downward causation sketched above. R depends both directly on D and indirectly on N. R is function of D and D is a function of N. Functional dependence is transitive; therefore R is a function of N.

What about "Exclusion", the second premise? This principle does not hold in the model I have sketched. Remember that the question we address is not about which event, interpreted as what fills a space-time zone, causes which event. This information is taken for granted. The information about the functional dependence of variables characterizing these events corresponds to *aspects, features* or *properties* of these events by virtue of which they influence each other. The functional dependencies represented by structural equations correspond to generalizations: The fact that R depends on D means that all events that resemble each other with respect to the variable D give rise to events that resemble each other with respect to the variable R. There is no reason why two such dependencies cannot coexist. No problem is created by the fact - if it is a fact - that R depends both on D and on N, in the sense that R is both a function of D and of N.

This double causal dependence is no case of "genuine overdetermination", which would be a situation in which R=1 is caused "twice over". Let me compare the situation with the situation of the firing squad where we have the clear intuition that the death of the victim is "overdetermined" in the sense of having several independent causes. The crucial difference with our model of mental causation is that the death caused by the firing squad is caused "several times", through several mutually independent paths, by several mutually independent particular events, which are located at different places, i.e. where the different soldiers of the firing squad stand. By contrast, R=1 is not "genuinely overdetermined" by the variables N=1 and D=1, or "caused twice over", because N and D are variables representing properties that the same person possesses at the same time. In terms of events, interpreted as what fills a space-time zone, there are just two particular events, which are related by a single causal relation: one event corresponds to the person at the time t_0 when the variables N and D have values N=1 and D=1, the other event corresponds to the person a t_1, when R has value R=1. Our model shows that an event can be causally influenced by two different aspects of some earlier event, or by two properties of that earlier event. This is so in particular when these aspects are not

[21] It is in general taken for granted that the dependence (and supervenience) of properties of wholes on properties of parts is a non-causal form of dependence (Kim 1974), insofar as it is a form of dependence without any temporal or spatial distance between the bearers of the two related properties. Some authors have recently argued that such dependence relations should be considered as causal nevertheless (Mumford and Anjum 2011, Leuridan 2012, Wilson 2018). I will leave this issue to one side and stick with the traditional thesis that properties standing in a supervenience relation based on the dependence of the properties of a whole on the properties of its parts do not stand in a causal relation. Schaffer (2016) uses structural equation models for both causal and non-causal grounding relations. However, Schaffer does not explore mixed models with both causal and non-causal dependence relations.

independent of each other: In our model, D is a function of N, which expresses the fact that the decision is nomologically dependent on the state of the person's brain.

However, even if the fact that R=1 causally depends on more than one factor does not entail that R is "genuinely" causally overdetermined, one might still argue that one of these factors – represented by the value of N - is fundamental, whereas the other – represented by the value of D - is only derivative. In other words, one might hold that a causal *explanation* of R by N is superior to an explanation of R by D, because the former explanation is in terms of a more fundamental variable. In other words, on might hold that even if R can be causally influenced "in parallel" by both N and D, the causal explanation of R in terms of the most fundamental variable N "excludes" all explanations in terms of less fundamental variables, such as D. One might hold in other words that even if a mental variable D can influence R in parallel to N, D is never explanatorily relevant. This reasoning depends on a principle of "causal-explanatory exclusion", according to which a causal explanation E_1 "excludes" other causal explanations E_2 of the same fact (even if both are correct) in the weak sense that E_1 is better than E_2 because E_1 explains the explanandum in terms of *more fundamental* variables than E_2.

Is it plausible that explanations in terms of more fundamental variables are always preferable? Explanations are assessed by two criteria: correctness and utility. The utility of an explanation depends on the interests and background knowledge of the explanation seeker (Bromberger 1966, van Fraassen 1980, p. 132-4), but it can also be evaluated in general terms of relevance. I would like to suggest that it is often more appropriate to causally explain a fact in terms of higher-level variables than to explain it in terms of more fundamental variables. The model of structural equations provides a straightforward criterion for comparing the relevance of various influences for the causal explanation of a given factor.

Each influence on R is expressed by a structural equation expressing a function R=F(X). This function can be injective or not[22]. A function Y=F(X) is called injective if there do not exist two different values $x_i \neq x_j$ of X that F maps on the same value of Y, so that $f(x_i)=f(x_j)$. The causal influence of X on Y is specific if and only if Y=F(X) is injective. Here is a criterion of relevance for causal explanations in terms of structural equations. If a variable Y depends on two factors X_i and X_j, and if $Y=F(X_i)$ is injective whereas $Y=F(X_j)$ is not injective, it is more relevant to causally explain Y in terms of X_i than in terms of X_j.

In the model sketched above, R is a function of both N and D. However, R is an injective function only of D, but not of N. The function R=F(N) is not injective, because F(N=1)=F(N=2)=R=1. However, function R=F(D) is injective because $F(D=0)\neq F(D=1)$. In our model, the variable D has not only downward causal influence on R, but it is more relevant to mention this downward causal influence in a causal explanation of R than to mention the parallel same-level cause N, because the downward causal influence is specific whereas the same-level influence is not[23].

---

[22] The analysis of specific causation in terms of the concept of an injective function is a variant of Woodward's (2010, p. 305) analysis, who builds on Yablo's (1992) notion of proportional causation and Lewis' (2000) notion of influence. My own use of the term "specificity" differs from Woodward's in that Woodward calls a function "specific" if it is both injective and surjective, whereas I use a weaker notion that requires only injectivity but not surjectivity. A function Y= f(X) is surjective if and only if, for every value y_i of Y there is some value x_j of X such that $y_i=f(x_j)$. Griffiths et al. (2015) provide a quantitative measure of the specificity of X for Y on terms of the mutual information between variables X and Y. See also Calcott (2017).

[23] List and Menzies (2009) analyze downward causation in terms of the notion of realization-insensitivity. However, their account leads to what they call "downward exclusion", according to which the causal

4. Other accounts of downward causation

There is downward causation in both models I have sketched. I would like to briefly put this result in the perspective of other proposals for making sense of downward causation. Both of our models provide ways of escaping the conclusion of the argument against downward causation by challenging "Exclusion". This result can be seen as establishing the cogency of compatibilism: A higher-level variable (or a higher-level feature of a system) D can exercise downward causal influence on a lower-level variable R even if there also are lower-level variables N that exercise low-level causal influence on the same variable R.

4.1. A counterfactual criterion for compatibility

Karen Bennett (2003) has developed a defense of compatibilism that relies on the fact that a mental sufficient cause *m* and a physical sufficient cause *p* of the same effect *e* can coexist insofar as the former depends on the latter, or in other words, is determined by the latter. In such a situation, *m* and *p* are not "overdetermining" *e*. If *m* and *p* did overdetermine their effect *e*, the following two counterfactuals would both be (non-vacuously) true:

O_1: If *m* had happened without *p*, *e* would still have happened.

O_2: If *p* had happened without *m*, *e* would still have happened.

Given that *m* is determined by *p*, O_2 is vacuous (Bennett 2003, p. 483-7), and so the compatibilist can deny that both O_1 and O_2 are non-vacuously true.

Our models can be seen as providing a complement to Bennett's demonstration. Bennett provides a criterion for the conceivability of downward causation: downward causation is conceivable if O_2 is vacuous. Our models clarify how O_2 can be vacuous in situations in which the issue of downward causation arises[24]. The semantic evaluation of O_1 and O_2 in our models yields the result that O_2 is vacuous. Thus, to the extent to which the models are adequate, there is downward causation in the situations represented by the models.

The reason for which O_2 comes out vacuous is the same in both models: The counterfactual situation described by the antecedent of O_2, in which *p* is present but *m* is absent, does not respect the relation between *p* and *m* in the actual world: Given the actual laws of nature, *p* determines *m*. In the SE model, the antecedent of O_2 corresponds

---

influence of a higher-level variable D on a lower-level variable R excludes the existence of a parallel low-level causal influence of N on the same variable R. It would by a mistake to judge, as List and Menzies (2009), but not Woodward (2010, p. 288) do, that all causation is specific (Kistler 2017; McDonnell 2017). My suggestion that the higher-level cause D is more relevant for the causal explanation of R than the lower-level cause N if the function R=F(D) is injective whereas R=F(N) is not injective, seems to be compatible with, and complementary to, Woodward's (2020) analysis. If both a higher-level variable U and a lower-level variable L, where the values of U are multiply realized by the values of L, cause E, with U being a downward cause of E, Woodward explains that it can be more relevant to mention U rather than L as a cause of E if U has a "uniform effect on E" (Woodward 2020, manuscript p. 23), whereas L is "causally independent of E conditional on U" (Woodward 2020, manuscript p. 22).

[24] Kim also suggests that the task of establishing the existence of certain causal relations cannot be accomplished simply by making it plausible that certain counterfactuals have certain truth values. "Merely to point to the apparent truth, and acceptability, of certain mind-body counterfactuals as a vindication of mind-body causation is to misconstrue the philosophical task at hand." (Kim 1998, p. 71) What is needed in addition is providing "an answer as to why these counterfactuals hold, that is to say, to find the relevant truthmakers" (Gozzano 2017, p. 301).

to a situation where N=1 but D=0. This situation is impossible in the model because it contradicts the functional dependence of D on N: N=1 is mapped on D=1. The functional dependence expressed in the SE D=F(N) represents the fact that D is grounded on N, which entails in turn that D supervenes on N, so that there cannot be a change in the value of D, while N is held fixed, which is the content of the antecedent of O_2. Thus, O_2's antecedent is false in all worlds that share our actual laws of nature.

In terms of dynamical systems, all systems that share the phase space of some actual cognitive system in our world, i.e. all systems that share the actual laws of nature, are such that $P \subset M$. The antecedent describes a system whose position in phase space lies within P but not within M. There is no system of that sort that corresponds to the actual laws of nature. Therefore the antecedent of O_2 is nomologically is impossible and O_2 is vacuous.

4.2. Rejection of Closure

Orilia and Paolini Paoletti (2017) claim that the acceptance of downward causation leads to "the rejection of causal closure" (Orilia and Paoletti 2017, p. 34). They justify this claim by using Yablo's (1992) framework of proportional causes. According to Orilia and Paolini Paoletti, the search for an adequate causal explanation of a bodily movement is constrained by the conception of that movement as being of a certain type, which has a certain degree of determination. Let us suppose that John's decision at t_1 to raise his arm causes, at t_2, an event at which his arm raises. That arm-raising event exemplifies a very specific sort of arm-raising, which they call R_{321}. The same event however also exemplifies a whole series of less and less specific, or more and more abstract types of arm-raising. R_{32} is the property of raising one's arm up to a height lying in a certain interval, with a speed lying within a certain interval, etc. but leaving open many more specific details, concerning e.g. the positions of the hand and fingers. R_3 might be the even more abstract property of raising one's arm, in any way whatsoever. The effect of John's decision corresponds to a particular degree of determination in that hierarchy. Let us suppose it is R_{32}. Now, they argue, the cause that is proportional to the exemplification of that property is the decision, i.e. a mental cause. By contrast, the exemplification of the underlying physical property of the person's brain and body is too specific to be proportional to the exemplification of R_{32}. This results indeed from the application of Yablo's criteria of proportionality: to be a proportional cause of *e*, a cause *c* must be 1) required for *e* and 2) enough for *e*. The physical cause underlying the decision is not required for R_{32}. Different physical events would have caused movements very similar to the actual movement exemplifying R_{321}: the movements they would have caused would still have belong to the determinable type R_{32} of arm-raising. Only John's decision to perform an action of type R_{32} is proportional to R_{32} (in the sense of being both required and enough for the exemplification of an event of precisely that type).

However, according to Orilia and Paoletti, the result that there is a downward causal influence from the decision to the arm-raising has been reached in a way that entails "the rejection of Causal Closure in the form suggested by Kim" (Orilia and Paolini, p. 34), according to which[25]: "If a physical event has a cause that occurs at *t*, it has a physical cause that occurs at *t*" (Kim 2005, p. 43).

[25] Contrary to the version we have used, this formulation of the closure principle leaves it open (following at this point Lowe 2000b) whether every physical event at *t* has a physical cause at every instant *t** earlier than *t*.

This result may seem surprising, insofar as it seems to contradict the compatibility of downward causation with the existence of a parallel underlying process of physical causation, which characterizes our two models of downward causation[26]. However, the contradiction is only apparent. It can be overcome by making explicit the different terminological choices underlying the two analyses. Orilia and Paoletti use the word "cause" only to make reference to proportional causes, in Yablo's sense of being both required and enough for a given effect. In that terminology there is indeed no physical cause happening at the same time as John's decision because all physical types of event are too specific to be required for the type of event R_{32}. In our own terminology, there may well be such a physical cause at the time of the mental cause. Closure can be accepted insofar as a cause is an event of a type that is sufficient for a given type of effect. In our terminology, and in our two models of downward causation, the mental cause of R_{32} at time *t* coexists with a physical cause at time *t*, but only the mental cause is specific (or proportional), which is why it is in general more appropriate to mention the mental cause in a causal explanation of why an event of type R_{32} has happened[27]. Our terminology seems preferable to Orilia and Paoletti's insofar as it makes it possible to say that there are non-specific causes.

4.3. Must downward causal relations necessarily be mediated by a synchronous top-down determination relation?

Carl Gillett (2016; 2017) argues that downward causal relations are necessarily mediated by synchronic top-down determination relations. Gillett's argument runs as follows.

1) In a first step, Gillett argues that there are numerous scientific examples of strong emergence ("S-emergence"). The instance of a property F in object *s* is a case of S-emergence if it a) contributes to determining powers of some parts of *s* and b) contributes to "powers causally resulting in effects at their own level" (Gillett 2017, p. 258). In other words, F is emergent if and only if it a) is a higher-level property of a composed object *s*, b) gives objects that possess it causal powers at its own level, i.e. makes them capable of influencing properties of other objects at the same level, and c) modifies the causal powers of the parts of *s*. Focusing on condition c), the S-emergence of property F requires that the fact that *s* has F modifies the causal powers of some of *s*'s

[26] Hendry also judges that "the existence of strong emergence in chemistry is incompatible with the causal closure of the physical" (Hendry 2017, p. 160). Anjum and Mumford (2017) say that downward causation requires "that causal closure should be rejected" (Anjum and Mumford 2017, p. 106). These authors do not explicitly argue against the hypothesis that all cases of downward causation are accompanied by parallel low-level causation. Their reasoning might be this. Given that a higher-level cause has to be postulated to causally explain the effect, there can be no base-level cause that can explain that effect. Thus, there is no causal closure at the base level. This reasoning relies on an oversimplification. In many cases in which the postulate and use of a higher-level cause is justified, that higher-level cause is only necessary to "specifically causally explain" the effect, not to explain the effect, tout court. Thus, it is often justified to introduce a higher-level variable and to use it to causally explain an effect although it is also possible to explain that same effect at a lower level. One reason for which the higher-level explanation may be better is that it is specific whereas the lower-level explanation lacks specificity.

[27] Woodward draws a similar distinction between David Lewis' (2000) terminology and his own, where the notion of specificity is used to "distinguish in a useful way *among* causal relationships, rather than treating it as a 'criterion' of causation" (Woodward 2010, p. 304; italics Woodward's). Lewis (2000) takes "influence", which is similar to causal specificity, to be characteristic of causation as such. Woodward's terminology is preferable to Lewis' (2000) insofar as it is compatible, whereas Lewis' is not, with the existence of non-specific causes.

parts, where this modification of the parts by the whole is synchronic, in the sense that *s* possesses F at the same time *t* at which *s*'s parts possess those powers that are modified by *s*'s possession of F.

2) In a second step, Gillett argues that this top-down determination relation cannot be causation[28]. A relation between the instance of a higher-level property P of an object *s* at *t* and an instance of a lower-level property Q of a part *p* of *s*, at the same time *t*, cannot be causal. The reason is that (*s*,P,*t*) and (*p*,Q,*t*) are temporally and spatially co-located, whereas causation requires the localizations of cause and effect to have no spatio-temporal overlap. Gillett dubs such synchronic top-down determination relations of properties of parts of composed objects by properties of those whole objects "machretic" (Gillett 2017, p. 257) determination relations.

3) From 1) and 2), Gillett draws the conclusion that downward causal relations, where an S-emergent property instance G at *t* influences the instance of some lower-level instance P at some later time *t**, can only be indirect: the influence of (*s*,G,*t*) (the exemplification of G by *s* at *t*) on (*p**,P*,*t**) (the exemplification of lower-level property P* by lower-level individual *p** at *t**, where *t** is later than *t*) must be mediated by a synchronous "machretic" top-down determination relation from (*s*,G,*t*) on (*p*,P,*t*) (the exemplification of lower-level property P by lower-level individual *p*, which is a part of *s*, at *t*).

For lack of space, I cannot here do full justice Gillett's analysis of machretic determination and downward causation. Let me just note that both models we have sketched above make sense of downward causation without positing any synchronous downward determination relation of the sort of Gillett's machresis.

Gillett's argument shows is that 1) the definition of S-emergence seems to entail machresis, 2) there are scientifically plausible cases of S-emergence, 3) machresis is non causal, and 4) given the acceptance of machresis, one can conceive of downward causation as mediated by machresis. However, this argument does not show that there cannot be downward causation that is not mediated by machresis.

The postulate of machresis, i.e. of a relation of synchronic top-down determination raises the following worry. If higher-level property G of complex object *s* at *t* can be given a "compositional explanation" (Gillett 2017, p. 246), in terms of the bottom-up determination of G by the parts p_1, … p_n of *s* and the lower-level properties P_i of those parts at the same time *t*, and if the higher-level property G also determines, at the same time *t*, in the reverse top-down direction, some lower-level property P_i of part p_i, it seems to follow that there can be (non-trivial) self-determination: The lower-level property P_i of part p_i determines G of *s*, which itself determines P_i of part p_i, all synchronously at *t*[29]. It would seem that models of downward causation that avoid the consequence that there can be non-trivial self-determination are preferable to those that do have that consequence.

Conclusion

[28] This move makes Gillett's account escape the objection based on Kim's (1999/2010, p. 35/6) argument according to which emergence entails that there are situations of "mutual causal interdependence" (Kim 1999/2010, p. 36), in which an object *x* is caused to acquire P at *t* although, at that same moment *t*, *x* already possesses P and exercises the causal determinative powers inherent in P.

[29] This argument, according to which the existence of mutual metaphysical determination entails, via the transitivity of metaphysical determination, the implausible consequence that contingent facts determine themselves, has a similar structure to the argument (Kistler 2013) according to which the interpretation of mutual nomic dependence as causal has the implausible consequence that contingent facts cause themselves.

I have sketched two conceptual frameworks that leave room for downward causation. Downward influence of higher-level features of complex systems on lower-level features of these systems can be represented in the framework both of dynamical systems and of structural equations. The "Exclusion" principle, which is a crucial premise of the argument against the possibility of downward causation, is false in both types of models. Furthermore, both frameworks can be completed with conceptual tools that make it possible to justify why downward causal influence is not only conceivable and compatible with the "Closure" principle, but also why it is often relevant to causally explain facts in terms of downward causation[30].

References

- Barsaglini Alessio et al. (2014), The effects of psychotherapy on brain function: A systematic and critical review. *Progress in Neurobiology* 114, p. 1-14.
- L.R. Baxter Jr. et al. (1992), Caudate glucose metabolic rate changes with both drug and behavior therapy for obsessive-compulsive disorder. *Arch. Gen. Psychiatry* 49, p. 681-689.
- Bennett, Karen (2003), Why the Exclusion Problem Seems Intractable, and How, Just Maybe, to Tract It. *Nous* 37, p. 471-497.
- Sylvain Bromberger, Why-Questions (1966), repr in *On What We Know We Don't Know*, Chicago, University of Chicago Press and Stanford, CSLI Press 1992, p. 75-100.
- Burge, Tyler (1979), Individualism and the Mental, in: P.A. French, T.E. Uehling and H.K. Wettstein, *Midwest Studies in Philosophy*, vol IV, Minneapolis: University of Minnesota Press.
- Calcott, Brett (2017), Causal specificity and the instructive-permissive distinction, *Biology and Philosophy* 32, p. 481-505.
- Cartwright, Nancy (1999), *The Dappled World. A Study of the Boundaries of Science*, Cambridge, Cambridge University Press.
- Craver, Carl (2007), *Explaining the Brain. Mechanism and the Mosaic Unity of Neuroscience*, Oxford: Oxford University Press.
- Dretske, Fred (1988), *Explaining Behavior*, Cambridge, MA: MIT Press.
- Dupré, John (1993), *The Disorder of Things. Metaphysical Foundations of the Disunity of Science*. Cambridge, MA: Harvard University Press.
- Eronen, Markus I. (2013), No Levels, No Problems: Downward Causation in Neuroscience, *Philosophy of Science* 80(5), p. 1042-1052.
- Eronen, M. (2019), The levels problem in psychopathology, *Psychological Medicine,* p. 1-7. doi:10.1017/S0033291719002514
- Flanagan, Owen (1992), *Consciousness Reconsidered*, Cambridge, MA: MIT Press.
- Gillett, Carl (2016), *Reduction and Emergence in Science and Philosophy,* Cambridge, Cambridge University Press.
- Gillett, Carl (2017), Scientific Emergentism and Its Move beyond (Direct) Downward Causation, in Michele Paolini Paoletti and Francesco Orilia (eds.), *Philosophical and Scientific Perspectives on Downward Causation*, New York: Routledge, p. 242-262.

[30] I thank Simone Gozzano and an anonymous referee for this volume for their helpful comments.

- Gozzano, Simone (2009), Levels, Orders and the Causal Status of Mental Properties, *European Journal of Philosophy* 17 (3), p. 347-362
- Gozzano, Simone (2017), The Compatibility of Downward Causation and Emergence, in Michele Paolini Paoletti and Francesco Orilia (eds.), *Philosophical and Scientific Perspectives on Downward Causation*, New York: Routledge, p. 296-312.
- Griffiths, Paul E., A. Pocheville, B. Calcott, K. Stotz, H. Kim, R. Knight (2015), Measuring Causal Specificity. *Philosophy of Science* 82, p. 529-555.
- Halpern, Joseph (2000), Axiomatizing Causal Reasoning. *Journal of Artificial Intelligence Research* 12, p. 317-337.
- Hendry, Robin (2017), Prospects for Strong Emergence in Chemistry, in Michele Paolini Paoletti and Francesco Orilia Michele Paolini Paoletti and Francesco Orilia (eds.), *Philosophical and Scientific Perspectives on Downward Causation*, New York: Routledge, p. 146-163.
- Hitchcock, Christopher (2012), Theories of causation and the exclusion argument. *Journal of Consciousness Studies* 19, p. 40-56.
- Kim, Jaegwon (1973), Causation, Nomic Subsumption, and the Concept of Event, *Journal of Philosophy* 70, p. 217–36. Reprinted in J. Kim, *Supervenience and Mind*, Cambridge, Cambridge University Press, 1993, p. 3–21.
- Kim, Jaegwon (1988a), Explanatory Realism, Causal Realism, and Explanatory Exclusion, Midwest Studies in Philosophy 12, p. 225-240.
- Kim, Jaegwon (1974), Non-Causal Connections, repr. in J. Kim, *Supervenience and Mind*, Cambridge: Cambridge University Press, 1993, p. 22-32.
- Kim, Jaegwon (1998), *Mind in a Physical World*, Cambridge, MA: MIT Press.
- Kim, Jaegwon (1999/2010), Making Sense of Emergence, repr. In J. Kim, *Essays in the Metaphysics of Mind*, Oxford: Oxford University Press, 2010, p. 8-40.
- Kim, Jaegwon (2005), *Physicalism, Or Something Near Enough*. Princeton: Princeton University Press.
- Kistler, Max (1999), Causes as events and facts. *Dialectica* 53, pp. 25-46.
- Kistler, Max (2006a), *Causation and Laws of Nature*, Londres, Routledge.
- Kistler, Max (2006b), The Mental, the Macroscopic, and their Effects. *Epistemologia* 29 p. 79-102.
- Kistler, Max (2013), The Interventionist Account of Causation and Non-causal Association Laws. *Erkenntnis* 78, p. 65–84.
- Kistler, Max (2014), Analysing Causation in Light of Intuitions, Causal Statements, and Science, in B. Copley, F. Martin (eds.), *Causation in Grammatical Structures*, Oxford University Press, 2014, p. 76-99.
- Kistler, Max (2017), Higher-Level, Downward and Specific Causation, in Michele Paolini Paoletti and Francesco Orilia (eds.), *Philosophical and Scientific Perspectives on Downward Causation*, New York: Routledge, p. 54-75.
- Leuridan, Bert (2012), Three Problems for the Mutual Manipulability Account of Constitutive Relevance in Mechanisms. *British Journal for the Philosophy of Science* 63: 399-427.
- Lewis, David (2000) Causation as Influence. *The Journal of Philosophy* 97, p. 182-197.
- List Christian and Peter Menzies (2009), Non-Reductive Physicalism and the Limits of the Exclusion Principle. *Journal of Philosophy* 106, p. 475-502.
- Lowe, E.J. (2000a), *An Introduction to the Philosophy of Mind*, Cambridge:

Cambridge University Press.
- Lowe, E.J. (2000b), Causal Closure Principles and Emergentism. *Philosophy* 75, p. 571-585.
- McDonnell, Neil (2017), Causal exclusion and the limits of proportionality. *Philosophical Studies* 174, p. 1459-1474.
- Mumford, Stephen, and Rani Lill Anjum (2011), *Getting Causes from Powers*, Oxford: Oxford University Press.
- Nagel, Ernest (1961), *The Structure of Science*, London, Routledge and Kegan Paul.
- Orilia, Francesco and Michele Paolini Paoletti, Three Grades of Downward Causation, in Michele Paolini Paoletti and Francesco Orilia (eds.), *Philosophical and Scientific Perspectives on Downward Causation*, New York: Routledge, p. 25-41.
- Pearl, Judea (2000), *Causality: Models, Reasoning, and Inference*, Cambridge University Press.
- Putnam, Hilary (1975), The Meaning of 'Meaning', in: H. Putnam, *Mind, Language, and Reality: Philosophical Papers, Vol. 2*, Cambridge: Cambridge University Press.
- Quine W.V.O. (1960), *Word and Object*, Cambridge,MA: MIT Press.
- Quine W.V.O. (1985), Events and Reification, in: E.LePore and B. McLaughlin (eds.) *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, Oxford: Basil Blackwell.
- Robb, David and John Heil (2018), Mental Causation. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/mental-causation/
- Schaffer, Jonathan (2016), Grounding in the Image of Causation. *Philosophical Studies* 173, p. 49-100.
- Schaffner, Kenneth (1967), Approaches to Reduction, *Philosophy of Science* 34, p. 137-147.
- Spirtes, Peter, Glymour, Clark and Scheines, Richard (2000), *Causation, Prediction and Search*, Second edition, Cambridge (Mass.), MIT Press.
- Van Fraassen, Bas C. (1980), *The Scientific Image*, Oxford: Oxford University Press.
- Voosholz, Jan (2020), Top-Down Causation Without Levels, in: *Top-Down Causation and Emergence* (this volume), eds. Markus Gabriel and Jan Voosholz, XXX–XXX. Springer.
- Yablo, Stephen (1992), Mental Causation. *Philosophical Review* 101, p. 245-280.
- Wilson, Alastair (2018), Metaphysical Causation. *Nous* 52, p. 723-751.
- Woodward, James (2003), *Making Things Happen*, New York: Oxford University Press.
- Woodward, James (2010), Causation in biology: stability, specificity, and the choice of levels of explanation. *Biology and Philosophy* 25, p. 287-318.
- Woodward, James (2020), Downward Causation Defended, in: *Top-Down Causation and Emergence* (this volume), eds. Markus Gabriel and Jan Voosholz, XXX–XXX. Springer.