The Causal Relevance of Content to Computation

Michael Rescorla

Department of Philosophy

University of California, Santa Barbara

Abstract: Many philosophers worry that the computational theory of mind (CTM) engenders epiphenomenalism. Building on Block's (1990) discussion, I formulate a particularly troubling version of this worry. I then present a novel solution to CTM's epiphenomenalist conundrum. I develop my solution within an *interventionist* theory of causal relevance, as espoused by Woodward (2003) and others. My solution departs substantially from orthodox versions of CTM. In particular, I reject the widespread picture of digital computation as formal syntactic manipulation.[1]

§1. Computation, content, and epiphenomenalism

According to *the classical computational theory of mind* (CTM), mental processes are digital computations. CTM has proved fruitful within philosophy (Fodor, 2008) and cognitive science (Gallistel and King, 2009). Many researchers regard it as our most promising theory of mental activity. Nevertheless, it faces chronic worries about *the causal relevance of content*.

[1] I am indebted to James Woodward for correspondence that exerted a decisive impact on this paper. I am also grateful to Ned Block, José Luis Bermúdez, Tyler Burge, Peter Carruthers, Kevin Falvey, Ian Nance, Agustín Rayo, Brad Weslake, and Eric Yang for helpful discussion of these issues. My research was supported by a fellowship from the National Endowment for the Humanities. Any views, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect those of the National Endowment for the Humanities.

Intuitively speaking, the contents of one's thoughts are causally relevant to subsequent thought and action. For example, suppose that Oscar's desire to drink orange juice causes an intention to open the refrigerator, which causes him to open the refrigerator. The content of Oscar's desire (*that I drink OJ*) seems causally relevant to his subsequent intention and behavior. But how can content play any significant causal role if mental activity is digital computation? On the standard picture, a digital computer is a "syntactic engine": it manipulates syntactic entities according to mechanical rules that allude solely to syntactic properties. Syntactic entities manipulated during computation may *have* meanings, but those meanings do not inform how computation proceeds. On this picture, content seems causally irrelevant to computation. To invoke Dretske's (1988) famous analogy, glass may shatter when a soprano screeches "Shatter," but her screech's meaning is causally irrelevant to the shattering. If mental activity is sensitive only to syntax, not semantics, then how can mental content be any more causally efficacious than the meaning of the soprano's screech? Epiphenomenalism looms.

Some readers will reply that epiphenomenalist worries arise without special appeal to CTM. The literature offers various arguments that neural or physical properties preclude mental properties from causal relevance. One might therefore insist that epiphenomenalism is a general problem extending far beyond CTM.

I disagree. Obviously, I cannot review the vast mental causation literature here.[2] I simply register my conviction, informed by the discussions of Burge (2007, pp. 344-382), Shapiro and Sober (2007), Woodward (2008), Yablo (1997), and many others, that there is no compelling *general* reason to fear epiphenomenalism. In contrast, I think that CTM encounters a serious and a distinctive epiphenomenalist challenge, a challenge that does not reflect more global epiphenomenalist worries. My suggestion should seem plausible. Almost no one really thinks

[2] For a survey, see (Bennett, 2007).

that mental content is causally irrelevant to mental activity. Yet it seems likely that content *is* causally irrelevant to the activity of an ordinary desktop computer. The computer manipulates syntactic items in accord with purely mechanical rules, paying no heed to what, if anything, those items mean. Semantics "makes no difference" to the computer's syntactic machinations. So epiphenomenalism appears true *of the desktop computer*. If a computer is our paradigm for mental activity, then a distinctive epiphenomenalist worry immediately arises.

In this connection, consider that Fodor's recent formulation of CTM includes the following crucial passage: "[c]omputations are operations defined over the *syntax* of mental representations; it is the syntax, rather than the content, of a mental state that determines its causal powers" (2008, p. 70). Fodor here comes perilously close to epiphenomenalism. If a mental state's content does not help "determine its causal powers," then isn't content causally irrelevant? By comparison, imagine a parallel thesis that replaces "syntax" with "neural properties": "the neural properties of a mental state, rather than its content, determine its causal powers." Few philosophers would accept this thesis as an uncontested premise. On the contrary, it seems to concede far too much to epiphenomenalism. The whole issue is whether content, not just neurophysiology, plays a significant causal role! Orthodox formulations of CTM concede a great deal, if not the whole game, to epiphenomenalism.

Proponents of CTM make various attempts to address their epiphenomenalist conundrum. I think that existing attempts fail. I want to present a new solution. My solution departs substantially from orthodox formulations of CTM. Specifically, I reject the nearly ubiquitous view that digital computation consists in "formal" syntactic manipulation. I will argue that my departure from orthodoxy is well-motivated by antecedently plausible doctrines about causation.

I begin by explaining more fully why CTM engenders a special epiphenomenalist worry (§2) and by examining previous attempts at alleviating the worry (§3). I then sketch the intuitive idea behind my solution (§4). The rest of the paper (§§5-11) develops my intuitive sketch more rigorously.

§2. An argument for epiphenomenalism

Why does CTM face such a serious and distinctive epiphenomenalist challenge? Block vividly formulates the basic worry (1990, p. 139):

> [A]ny Turing machine can be constructed from simple primitive processors such as *and* gates, *or* gates, and the like… Gates are sensitive to the syntactic forms of representations, not their meanings. But if the meaning of a representation cannot influence the behavior of a gate, how could it influence the behavior of a computer --- a system of gates?

Following Kazez (1995), I isolate three key premises underlying Block's formulation:

> **COMP:** Computations are sequences of operations on symbols performed by primitive processors.
>
> **PRIM:** Semantic properties of a symbol received as input by a primitive processor are causally irrelevant to the processor's output.
>
> **WHOL:** Suppose that, for each primitive processor in a computational system, semantic properties of the input received by that processor are causally irrelevant to the processor's output. Then semantic properties are causally irrelevant to the system's overall activity, including its behavioral outputs.

These three premises entail that semantic properties are causally irrelevant to a computational system's activity, including its behavioral outputs.

Not everyone accepts COMP. For instance, connectionists treat computation not as "symbol manipulation" but as evolution of weights and activations in a connectionist network. I will focus solely on digital computation, where COMP prevails. This "classical" paradigm is favored by many, though not all, cognitive scientists (Gallistel and King, 2009). As we will see, digital computation raises enough complexities to fill an entire paper. I therefore assume COMP.

A plausible computational model of the mind will not just posit "pure" processors whose inputs and outputs are both syntactic. It will also posit "perceptual transducers," which convert sensory inputs into syntactic outputs, and "behavioral transducers," which convert syntactic inputs into motor outputs. Given the distinction between syntactic, perceptual, and behavioral processors, let us reformulate PRIM and WHOL:

> **PRIM-SYN**: Semantic properties of a symbol received as input by a syntactic primitive processor are causally irrelevant to syntactic properties of the processor's output.
>
> **PRIM-MOT:** Semantic properties of a symbol received as input by a behavioral transducer are causally irrelevant to the motor output that the transducer yields.
>
> **WHOL-SYN/MOT:** Suppose that, for each primitive processor in a computational system, semantic properties of the input received by that processor are causally irrelevant to the processor's syntactic/motor output. Then semantic properties are causally irrelevant to subsequent syntactic/motor developments in the computational system.

By "motor output," I mean motor gestures as described in whatever terms our computational model employs: muscle activations, muscle contractions, limb velocities, etc.

CTM, COMP, PRIM-SYN, PRIM-MOT, and WHOL-SYN/MOT jointly yield a persuasive argument that mental content is causally irrelevant to behavior. By CTM and COMP, we can model the mind as a network composed of primitive processors, including perceptual and behavioral transducers. PRIM-SYN and PRIM-MOT state that semantic input to a processor is causally irrelevant to the processor's syntactic/motor output. WHOL-SYN/MOT immediately yields that semantic input to a processor is causally irrelevant to any *other* processor's syntactic/motor output. In other words, semantics is causally irrelevant to the system's overall activity, including motor outputs.

Can we generalize this argument from the syntactic to the neural or physical level? To do so, we must replace PRIM-SYN with an analogous premise that replaces syntactic properties with neural/physical properties. The new premise will look something like this:

> **PRIM-NEURAL:** Semantic properties of a neural/physical state are causally irrelevant to the neural/physical properties of "immediately" ensuing neural/physical states.

If we concede PRIM-NEURAL, then we concede that semantic properties are causally irrelevant to a vast range of neural/physical developments. We are well on our way towards epiphenomenalism. PRIM-NEURAL may be true, but we should hardly assume it from the outset. Epiphenomenalists owe us an excellent argument for PRIM-NEURAL.

The fundamental asymmetry here is that, even though few philosophers would readily concede PRIM-NEURAL, proponents of CTM almost universally embrace the analogous premise PRIM-SYN. Beginning with Fodor (1981), numerous philosophers have insisted that computation is *formal*, meaning roughly that semantics does not inform computational operations. As Fodor puts, computational operations "consist entirely of transformations of symbols; in the course of performing those operations, the machine is sensitive solely to syntactic properties of the symbols" (1987, p. 19). Similarly, Egan states that computational operations "are sensitive only to *formal* (i.e. *non-semantic*) properties of the representations over which they are defined, not to their content" (2010, p. 254). Even more explicitly, Block writes that "the *and* gate is sensitive only to the syntax, not meanings of its inputs, and likewise for the primitive processors postulated by cognitive science accounts of the mind" (1990, p. 142). Commentators almost universally hold that digital computation decomposes into elementary syntactic operations that are not sensitive to semantics. Depending on how we gloss the key term "sensitive," this position either entails or strongly suggests PRIM-SYN. That is why epiphenomenalism poses an *especially* serious problem for CTM.

§3. Strategies for avoiding epiphenomenalism

To say that epiphenomenalism poses a serious problem for CTM is not to label the problem insoluble. Can we somehow defuse Block's epiphenomenalist argument?

Proponents of CTM often suggest that we can combat epiphenomenalism by distinguishing various *levels of description*. In this spirit, let us differentiate PRIM-SYN and

> **PRIM-SEM**: Semantic properties of a symbol received as input by a syntactic primitive processor are causally irrelevant to semantic properties of the processor's output.

One can concede PRIM-SYN while rejecting PRIM-SEM. One can hold that semantic properties are causally irrelevant to computational operations *as described syntactically* but causally relevant to computational operations *as described semantically*. Peacocke (1994) advocates a view along these lines, although he does not focus specifically on logic gates. He concedes that "the total [causal] story of a sequence of events at the syntactic level will indeed say nothing about content," but he insists that there is also a "causal story at the content-involving level" that assigns an essential role to intentional content (p. 326).

In a related vein, Fodor (1990) argues that mental activity falls under *intentional laws*, i.e. laws that mention intentional content. He espouses a nomological theory of causation, according to which a property is "causally responsible" if the property plays an appropriate role in suitable (possibly *ceteris paribus*) laws (1990, pp. 137-159). He concludes that intentional properties are causally responsible, even though "the properties of mental states in virtue of which they are engaged by mental processes are intrinsic/syntactic" (1991, p. 298). On Fodor's view, formal syntactic operations implement intentional laws, thereby securing the causal relevance of mental content. Mental content figures in the total causal story about mental activity but not the total causal story *at the syntactic level*.

Unfortunately, such maneuvers do not help combat Block's epiphenomenalist argument. As reconstructed in §2, the argument's premises are CTM, COMP, PRIM-SYN, PRIM-MOT, and WHOL-SYN/MOT. The argument does not invoke PRIM-SEM, either implicitly or explicitly. Thus, one does not attack the argument by rejecting PRIM-SEM. More generally, one cannot rebut the argument by distinguishing the "syntactic causal story" from the "intentional causal story." Block's argument purports to show that *there is no* intentional causal story about a given bodily motion. Once one accepts the argument's premises, one must accept that mental content is causally irrelevant to bodily motion.[3]

Fodor and Peacocke might retort that mental content, while causally irrelevant to bodily motion as described in certain terms (e.g. *performing certain muscular contractions*), is causally relevant to bodily motion as described in other terms (e.g. *opening the refrigerator door*, or

[3] Another problem facing Fodor's treatment is that nomological theories of causation are very problematic (Block, 1990), (Woodward, 2008). For further discussion of Peacocke, see (Rescorla, forthcoming a).

perhaps *intentionally opening the refrigerator door*). However, this "dual explanandum" strategy still concedes a troubling epiphenomenalist consequence: the content of Oscar's desire to drink OJ is causally irrelevant to his muscular contractions. Intuitively speaking, Oscar's desire for OJ rather than water causally influences which muscular contractions he performs. If possible, we would like to preserve that intuition. At least, I would like to preserve it.

A good solution should tackle §2's argument directly: by rejecting PRIM-SYN, PRIM-MOT, or WHOL-SYN/MOT. Most likely, PRIM-SYN and PRIM-MOT stand or fall together. Thus, we should either reject PRIM-SYN or reject WHOL-SYN/MOT.

Block (1990) favors the latter option. He claims that a computer's "internal processors can be sensitive to *both* syntax and semantics" (1990, p. 151), even though *primitive* processors "are sensitive the syntactic forms of representations, not their meanings" (p. 139). Content is causally relevant to the network as a whole but not to any individual gate.

Although the solution I will propose borrows key ideas from Block's discussion, his specific proposal strikes me as unpromising. The basic difficulty, emphasized by Kazez (1995), is that WHOL-SYN/MOT seems overwhelmingly plausible. Suppose we wire together primitive processors G_1, … G_n to form a computational system. Some of the processors may be perceptual or behavioral transducers. Given PRIM-SYN and PRIM-MOT, a processor's semantic input is causally irrelevant to its syntactic/motor output. But then how can the processor's semantic input be causally relevant to *any* syntactic or motor aspects of computation? How can semantic input to a primitive processor influence subsequent syntactic/motor developments without influencing the processor's *own* syntactic/motor output?[4]

The present paper focuses not upon criticizing earlier treatments, but rather upon developing my own solution to CTM's epiphenomenalist worries. My favored solution is to

[4] (Rescorla, forthcoming a) offers further criticism of Block.

reject PRIM-SYN. I hold that a computational system's elementary syntactic operations can be sensitive to semantics as well as syntax. A primitive processor's semantic input can be causally relevant to that processor's output *as described syntactically*. As we have seen, virtually all commentators hold that primitive processors are sensitive to syntax but not semantics. By rejecting PRIM-SYN, I reject the orthodox formulation of CTM. I reject the popular picture of digital computation as formal syntactic manipulation.

§4. Rethinking PRIM-SYN

Following Block, I think the key issue here is *original* versus *derived* intentionality (Haugeland, 1985), or *intrinsic* versus *observer relative* meanings (Searle, 1980). These phrases have various connotations I want to avoid, so I will instead speak of *indigenous* versus *inherited* meaning. *Inherited meanings* arise when a system's semantic properties are assigned to it by external observers, either through explicit stipulation or through tacit convention. Nothing about the system helps generate its own semantics. *Indigenous meanings* arise when a system helps generate its own semantics (perhaps with ample help from its evolutionary, design, or causal history, along with other factors). The system helps confer content upon itself, through its internal operations or its interactions with the external world. Its semantics does not simply result from external assignment. Thus, the central dichotomy is between systems whose intentionality is entirely due to an external system versus systems that help generate their own intentionality.

For instance, words in a book have inherited meanings, because they have intentionality only through their connection to human linguistic conventions. Likewise, a simple network of gates has only whatever intentionality an external observer assigns to it. Quite plausibly, the same analysis applies even to fairly sophisticated computational systems, such as pocket calculators and probably even desktop computers. In contrast, the human mind has indigenous meanings. Of course, external observers can assign inherited meanings to a system that has its own indigenous meanings. The key point is that some systems *only* have inherited meanings, while others also have indigenous meanings.

Classical cognitive science assumes that a suitably sophisticated computational system has indigenous meanings. If we connect various primitive processors (including perceptual and behavioral transducers) into a suitably sophisticated computational network, then the network's syntactic manipulations help confer indigenous meanings on its internal states. A sufficiently sophisticated robot has content partly by virtue of the robot's own syntactic activity. Many philosophers, including Block (1990) and Parisien and Thagard (2008), accept this claim. Searle (1980) rejects it. To reject it is to abandon the whole enterprise of modeling intentional creatures as computational systems. Thus, we may safely assume it here.

Inherited and indigenous meanings bear radically different relations to a computational system's underlying syntactic manipulations. Inherited meanings result from external imposition, so they can vary arbitrarily relative to the syntactic manipulations upon which they are imposed. Certain semantic interpretations may seem more *natural* to us than others. For instance, base-10 interpretation of a standard calculator seems more natural than a base-13 interpretation, since the former but not the latter allows us to interpret the calculator as performing useful arithmetical calculations. But the syntactic manipulations themselves do not favor one interpretation over any other (Haugeland, 1985, pp. 121-122). The situation is quite different for indigenous meanings. To illustrate, suppose that representation *r*, as used by robot *X*, has the indigenous meaning *red*. If we hold fixed *X*'s syntactic activity (including the interface with transducers) while varying *X*'s other properties, then how much can we vary the meaning of *r*? Could *r* have meant *OJ*, or

Paris, or *Richard Nixon*? Could a physical system endow *r* with one of those alternative indigenous meanings while manipulating syntax as *X* does? That seems doubtful. Thus, *X*'s syntactic activity (including the interface with transducers) constrains *r*'s meaning. We cannot arbitrarily vary indigenous meanings while holding syntax fixed.

This crucial contrast suggests a further contrast regarding causal relevance. Since we can arbitrarily vary inherited meanings relative to syntactic machinations, inherited meanings do not *make a difference* to those machinations. They are imposed upon an underlying causal structure. By comparison, indigenous meanings do not vary arbitrarily relative to syntactic machinations. They arise partly from the system itself, not merely from external assignment. Quite plausibly, then, they *make a difference* to the system's activity. It is plausible that they inform the system's causal structure, because they emerge partly from within that very structure. *Indigenous* meanings may be causally relevant to computation, even though *inherited* meanings are not.

I claim that indigenous meanings are causally relevant to computational operations, including primitive syntactic manipulations. A primitive processor is not sensitive to semantics when isolated from other processors, but it becomes sensitive to semantics once embedded in a computational system with indigenous meanings. Elementary syntactic operations become sensitive to semantics once suitably immersed in a computational system that helps confer meaning on its own internal states. The rest of my discussion develops this intuitive idea.

§5. Interventionist theories of causal relevance

The first step towards a rigorous development is to choose a systematic theory of causal relevance. Like many philosophers, I favor a counterfactual approach. In particular, I favor an *interventionist* counterfactual approach (Spirtes, Glymour, and Scheines, 1993), (Pearl, 2000), (Woodward, 2003, 2008), (Woodward and Hitchcock, 2003). Roughly, interventionists hold that *C* is causally relevant to *E* just in case *E* would change if we were to manipulate *C* appropriately. In this section, I present basic aspects of interventionism. §6 explores how the interventionist framework applies to mental causation. §§7-11 uses interventionism to develop my intuitive solution to CTM's epiphenomenalist worries.

My intuitive solution does not depend upon the details of interventionism. For instance, I think one could develop a similar solution within Lewis's (2000) theory of causal influence. I choose interventionism for two reasons. First, it handles various phenomena that elude rival accounts, including Lewis's theory (Woodward, 2003, pp. 133-149). Second, it meshes well with current scientific practice (Woodward and Hitchcock, 2003). By invoking interventionism, I risk alienating philosophers who favor some alternative treatment of causal relevance. But a comparable risk is inevitable whenever discussing mental causation. There is no way to address these issues systematically without endorsing *some* specific doctrines about causal relevance.

I take Woodward's (2008) version of interventionism as my starting point. Woodward focuses primarily on *type* rather than *token* causation. He primarily studies claims of the form

Whether one smokes is causally relevant to whether one contracts cancer

rather than claims of the form

Jones's smoking was causally relevant to his contracting cancer

Token-causation raises many complexities, such as preemption, not directly relevant to my concerns. I therefore focus upon type-causation, even though a complete treatment of mental causation would also address token-causation.

Interventionists emphasize the general notion of *variable*. The *values* of a variable are possible states of the system under consideration. The simplest variables take two values. For any property *P*, there is a variable whose two values correspond to some object either having or not having *P*. For any event-type *e*, there is a variable whose two possible values are occurrence versus non-occurrence of an *e*-type event. In general, a variable (e.g. velocity) may assume more than two values. Interventionists regiment type-causal claims by treating them as relations among variables. Thus, the central locution of an interventionist theory is

Variable *X* is causally relevant to variable *Y*.

For instance, suppose that the baseball will break the window just in case its velocity exceeds 4 m.p.h. Let variable VELOCITY have possible baseball velocities as values. Let variable WINDOW have two values, reflecting whether the window breaks. Then

VELOCITY is causally relevant to WINDOW,

corresponding roughly to the English paraphrase: *which velocity the baseball has is causally relevant to whether the window breaks*.

Notably, this regimentation treats causation as *contrastive*. Type-causal relations obtain between variables, which enshrine contrast classes of possible values. There are strong reasons for treating causal relevance as contrastive (Schaffer, 2005), (Woodward, 2003, pp. 145-146). For instance, let VELOCITY* have baseball velocities above 4 m.p.h. as values. Then we can say that VELOCITY is causally relevant to WINDOW, but VELOCITY* is not. This regimentation captures the following intuitive distinction: *which velocity the baseball has is causally relevant to whether the window breaks*, but *which velocity the baseball has in the range above 4 m.p.h is causally irrelevant to whether the window breaks*.

The basic idea behind interventionism is that *X* is causally relevant *Y* just in case manipulating *X* is a way of manipulating *Y*. To make this idea precise, we introduce the notion of an *intervention*.

Intuitively, an intervention on X with respect to Y is a change in X that affects Y, if at all, only through the change in X and not through an independent causal route. Thus, an intervention on X with respect to Y must not alter any variable that exerts a causal influence on Y independent of whatever causal influence X exerts on Y. For instance, suppose we alter the baseball's velocity from 2 m.p.h. to 10 m.p.h. by throwing it harder. We thereby intervene on VELOCITY with respect to WINDOW, because any change in WINDOW results from our change in VELOCITY. In contrast, suppose we alter VELOCITY by exploding a bomb whose shock wave both slows the baseball *and* shatters the window. We thereby manipulate VELOCITY, but we do not intervene upon it (with respect to WINDOW). Our manipulation exerts a causal influence on WINDOW that does not run through VELOCITY. An intervention on X with respect to Y must not alter confounding variables, including common causes of X and Y.

The intuitive idea behind interventionism is that X is causally relevant to Y just in case Y would change if suitable interventions on X were to occur. VELOCITY is causally relevant to WINDOW, because suitable interventions on baseball velocity change whether the window shatters. WINDOW is causally irrelevant to VELOCITY, because intervening on whether the window shatters (e.g. by smashing it with a hammer) yields no change in the baseball's velocity.

Interventionists formulate this intuitive idea in various ways. The formulation most useful to us runs as follows (Woodward, 2008): X is causally relevant to Y iff there are distinct values x, x^* of X and y, y^* of Y such that, for some background circumstance B:

An intervention that sets $X=x$ occurs in B □→ $Y=y$

An intervention that sets $X=x^*$occurs in B □→ $Y=y^*$,

where "□→" is the counterfactual conditional. Note the appeal to background circumstances. Short-circuits are causally relevant to fires, because an intervention that causes a short-circuit would lead to a fire. However, this interventionist counterfactual prevails only when oxygen is present, along with other suitable background conditions. Thus, Woodward's account demands only that there exist a background B relative to which desired interventionist counterfactuals prevail. For smoothness of exposition, I will sometimes omit mention of background conditions.

A central task for interventionism is to define "*I* is an intervention on *X* with respect to *Y*" more precisely. Woodward (2003, pp. 94-114) offers a definition that strips the notion of any anthropomorphic overtones. I will not discuss his definition here, except to note one important feature: it freely cites causal relevance between variables other than *X* and *Y*. Thus, Woodward elucidates interventions in terms of causal relevance and causal relevance in terms of interventions. He does not seek a reductive theory of causal relevance. His goal is simply to illuminate how various notions, such as causal relevance and intervention, relate to one another.

An intervention *I* on *X* with respect to *Y* need not leave fixed *all* other variables that causally influence *Y*. For instance, *I* can alter variables that fall on the causal path $I \rightarrow X \rightarrow Y$. A subtler point, which the current interventionist literature does not sufficiently emphasize, is that *I* can alter a variable *Z* that is not "independently manipulable" from *X*. We can describe a physical system through numerous distinct variables, reflecting different ways of type-identifying the system's states. In many cases, two variables *X* and *Z* will "overlap," so that they do not reflect truly independent parameters. In such a case, an intervention on *X* need not leave *Z* fixed, because *X* and *Z* are not independently manipulable. Here are three examples.

Baseball velocity: Let variable THRESH (for *threshold*) have two values, reflecting whether the baseball's velocity exceeds 4 m.p.h. THRESH is causally relevant to WINDOW, because intervening on THRESH yields a change in WINDOW. VELOCITY is also causally relevant to

WINDOW, because altering VELOCITY from 2 to 10 m.p.h yields a change in WINDOW. If *I* is an intervention on THRESH, then *I* alters VELOCITY as well, even though VELOCITY does not occupy an intermediate position in the causal route *I*→THRESH→WINDOW. Why does *I* nevertheless count as an intervention? Because THRESH and VELOCITY are not independently manipulable. A change in THRESH and a change in VELOCITY are not independent causal influences on WINDOW, so an intervention on THRESH can legitimately change VELOCITY.

A "high-level" variable and its supervenience base: A more interesting case concerns the relation between "lower-level" variables and the "higher-level" variables cited within macrophysics and the special sciences. Any manipulation of a "higher-level" variable *X* will alter *X*'s microphysical "supervenience base" *Z*. The supervenience base *Z* may be causally relevant to some effect *Y* of *X*, even though *Z* falls on no causal path that includes *X*. We should not deny on this basis that our manipulation of *X* is an intervention. Else, we will conclude that interventions almost never occur. How can our change in *X* count as an intervention, given that it changes *Z*? Because *X* and *Z* are not independently manipulable. The causal relation between supervenient variable *X* and effect *Y* is not independent from the causal relation between subvenient variable *Z* and *Y*, because the supervenient and subvenient variables are intimately linked.

The soccer game: Suppose that the crowd will riot just in case Team A beats Team B in the soccer game. Intuitively speaking,

Whether *A* or *B* wins the game is causally relevant to whether the crowd riots.

In the absence of ties, a soccer game has two periods. The team with the higher second period score wins the game. So we also want to say that

The second period score is causally relevant to whether the crowd riots.

We can capture these two reactions by introducing two variables GAME and SCORE. GAME reflects whether A or B wins the game. SCORE reflects the second period score. An intervention on GAME may alter SCORE. For instance, we might intervene on GAME by drugging Team A so that it scores fewer points. This manipulation is a legitimate intervention on GAME, because GAME and SCORE are not independently manipulable. The game's outcome and the second period score are not independent causal influences on whether the crowd riots.

Notably, neither GAME nor SCORE supervenes upon the other. A change in SCORE may yield no change in GAME. Conversely, SCORE leaves GAME undetermined in case of a tie. In that case, an extra tie-breaking round determines the winner. (The precise details vary, but they are irrelevant to us.) Which team wins the game is intimately entangled with, although not always determined by, the second period score. Thus, two causal variables may not be independently manipulable even though neither supervenes upon the other.

Under what conditions are X and Z "independently manipulable"? The intuitive idea is that X and Z vary arbitrarily from one another. A natural formalization of this idea is that, for any value x of Z and any value z of Z, it is metaphysically possible that $X=x$ and $Z=z$. My formalization may require some tinkering. The topic is an important one, but I cannot pursue it here. What matters for us is that any satisfactory development of interventionism requires some notion of "independent manipulability" in this vicinity.

§6. Interventionism and mental causation

How does interventionism apply to mental causation?

We routinely manipulate one another's propositional attitudes, thereby influencing behavior. Most such manipulations are not interventions, since they simultaneously alter numerous distinct mental states. In theory, though, we can imagine "targeted interventions" that alter individual mental states through brainwashing, subliminal advertising, or even direct physiological manipulation of the neural substrate. For instance, we might intervene through subliminal advertising to make Oscar desire water rather than OJ. As a result of this targeted intervention, Oscar may walk to the refrigerator rather than the sink (assuming an appropriate background of additional beliefs and desires). In general, it seems plausible that many targeted interventions on individual mental states yield determinate changes in behavior. So it seems plausible that a detailed interventionist analysis would depict mental content as causally relevant to behavior (Woodward, 2008).

In manipulating the content of Oscar's desire from water to OJ, we alter various neural properties of Oscar's brain states. Can our manipulation still count as an intervention on content? The key question here is whether mental content is independently manipulable from the neural or physical substrate (Campbell, 2008). If mental content is *not* independently manipulable, then an intervention on mental content may legitimately alter the neural or physical substrate.[5]

Philosophical discussions of mental causation frequently assume a strong failure of independently manipulability. Specifically, many discussions assume that mental properties supervene upon whatever neural or physical properties are causally relevant to behavior (Kim, 2005). A high-level variable is not independently manipulable from its supervenience base.

Other philosophers deny that mental properties supervene upon neural properties. *Content externalists* hold that mental content is "wide" rather than "narrow," in the sense that it

[5] Brad Weslake and James Woodward both emphasize independent manipulability in unpublished drafts that discuss mental causation.

does not supervene upon neurophysiology (Burge, 2007). For instance, suppose that neural state type *N* is currently associated with Oscar's desire to drink water. According to externalists, we can change the content of Oscar's desire without altering the neural substrate. *N* could instead have been associated with a desire to drink twater (the substance on Twin Earth) had Oscar been differently embedded in his environment. On this view, one can change mental content in *certain* ways without changing the neural substrate.

It does not follow that mental content and the neural substrate are independently manipulable. The question is not whether *N* could have had *different* semantic properties. The question is whether *N* have had *arbitrarily different* semantic properties. Could *N* have corresponded to a desire to eat breakfast? Or a desire to jump off a cliff? Or a desire to see Rome before one dies? Externalism provides no reason to suppose that *N*'s semantics can vary so wildly. Externalism provides no reason to suppose that we can change content *arbitrarily* with respect to the neural substrate. By analogy, we can change VELOCITY in *certain* ways while holding THRESH fixed, but we cannot change VELOCITY *arbitrarily* while holding THRESH fixed. Thus, THRESH and VELOCITY are not independently manipulable. For an analogy not involving supervenience, consider GAME and SCORE. We can change SCORE without changing GAME, and we can change GAME without changing SCORE (when the second period score is tied). Nevertheless, GAME and SCORE are not independently manipulable.

Some readers may protest that we can vary mental content arbitrarily with respect to the neural substrate if we change not only "external" factors but also the brain's internal wiring. Couldn't the very same neural state *N* have occupied a radically different functional role, thereby enjoying an arbitrarily different content? Perhaps. Perhaps not. We currently understand very little about the neural underpinnings of mental activity. Given our present state of ignorance, independent manipulability of mental content and the neural substrate is not an uncontroversial premise that epiphenomenalists can assume without extensive argument.

A compelling interventionist rebuttal to epiphenomenalism would argue convincingly that mental content and the neural substrate are *not* independently manipulable. Ultimately, such an argument should address neuroscientific details in great detail. Thus, I do not claim to have provided a definitive interventionist argument against epiphenomenalism. My goal is simply to highlight the central role that independent manipulability must play in any adequate interventionist argument *for or against* epiphenomenalism.

§6.1 The exclusion argument

In this connection, let us consider Kim's widely discussed "exclusion argument." The argument hinges upon the "Principle of Causal Exclusion" (2005, p. 17):

> 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 case of genuine overdetermination).

An example of genuine overdetermination is two snipers who simultaneously shoot some victim, with each shot sufficient for death. Such cases are rare. Most philosophers agree that mental causation does not involve overdetermination in anything like the way the snipers overdetermine the victim's death. Given that mental causation does not involve overdetermination, and given that physical properties are causally sufficient for behavior, mental properties can achieve causal relevance (according to the exclusion principle) only if they are identical with physical properties. Kim concludes that only reductive physicalists can avoid epiphenomenalism.

Kim's exclusion argument has sparked a vast literature. Numerous philosophers have complained that there is no compelling reason to believe the exclusion principle. As Burge notes, arguments for the exclusion principle tend to presuppose a “hydraulic” model of causation, on which only “so much energy is needed to get the job done” (2007, p. 380). The picture is that physical properties do all the “causal work,” leaving no “work” left over for mental properties. There is no clear reason why we should accept these metaphorical appeals to “causal work.” There is no clear reason to believe that a sufficient physical cause for some behavior precludes mental properties from causal relevance. Mental and physical properties do not somehow “compete” for causal relevance. They are both relevant.

As Bennett (2007) notes, most attacks on the exclusion principle hold that *C* and *C** do not compete for causal relevance if they bear a suitably “intimate” relation to one another, such as constitution, realization, supervenience, determination of a determinable, etc. The basic idea behind these accounts is that, because a mental and physical cause are so “intimately” related, they are not truly independent in the way that the two snipers are independent. Thus, we can say that the mental and physical causes are both causally relevant without positing anything resembling the overdetermination engendered by the two snipers.

Interventionism provides an attractive framework for developing this basic idea. Interventionists can gloss the requisite “intimacy” between *C* and *C** as follows:

C and *C** are not independently manipulable.

If two variables satisfy this criterion of “intimacy,” then they can both be causally relevant to a single effect, absent overdetermination. I offered three examples in the previous section: baseball velocity (VELOCITY and THRESH); a “high-level” variable and its supervenience base; and the soccer game (GAME and SCORE). Each case involves a pair of variables that are intimately entangled with one another. The two variables are not independently manipulable, so they do not compete for causal relevance. Both variables can be causally relevant to some effect without engendering “causal overdetermination” in any ordinary sense. Thus, so long as mental content is not independently manipulable from the neural substrate, both mental and neural variables can be causally relevant to behavior.

As we have seen, Kim assumes a strong failure of independent manipulability. He assumes that mental content supervenes upon those physical properties causally relevant to behavior. Hence, from an interventionist perspective, Kim’s exclusion argument poses no serious challenge to the causal relevance of mental content. Various interventionists have defended this conclusion in considerable detail (Campbell, 2008), (List and Menzies, 2009), (Shapiro, 2010), (Shapiro and Sober, 2007), (Woodward, 2008).[6]

Notably, Kim appears to concede as much. He emphasizes that the exclusion principle concerns “causation as generation, or effective production and determination,” rather than causation elucidated counterfactually (2005, p. 18). He claims that the former conception, rather than the latter, “is fundamentally involved in the problem of mental causation” (p. 18). As List and Menzies (2009) observe, neither Kim nor anyone else has satisfactorily elucidated “causation as generation” in non-counterfactual terms. In any event, I assume that interventionism captures *one* fruitful notion of causal relevance. I want to explore whether mental content is causally relevant in *that* sense.

A convincing interventionist version of the exclusion argument must show that mental content and the neural substrate are independently manipulable. More precisely, it must isolate a neural variable *N* that is causally relevant to some bodily motion and that is *also* independently

[6] The exclusion principle concerns token-causation, while interventionism as I formulated it concerns type-causation. Another wrinkle is that the exclusion principle discusses “causal sufficiency” rather than “causal relevance.” A more careful discussion would take these subtleties into account. But I do not think the basic moral would change.

manipulable from the pertinent mental content variable. Since Kim assumes supervenience, he does not even begin to provide such an argument.

§6.2 Content externalism and interventionism

Another popular epiphenomenalist argument targets content externalism. Various philosophers hold that wide content is not causally efficacious, since only properties that supervene on the thinker's "intrinsic" or "local" physical properties are causally relevant to behavior (Fodor, 1987). For instance, suppose we transform Oscar's water-thoughts into twater-thoughts while holding fixed his neural properties, perhaps by surreptitiously transporting him to Twin Earth as a child. The same bodily motion will result from Oscar's desire for twater as from his desire for water. Doesn't this show that the causally relevant factor is really the neural state, or perhaps some narrow content that supervenes on the neural state?

Many authors, including Burge (2007, pp. 316-343) and Yablo (1997), have criticized such arguments.[7] From an interventionist perspective, the arguments seem particularly dubious. Consider a variable DESIRE that includes each possible desire in Oscar's "cognitive repertoire," along with each possible desire in Twin Oscar's "cognitive repertoire." Each value of DESIRE reflects a content towards which either Oscar or Twin Oscar, as currently configured, can bear the attitude of desire. Quite plausibly, there are distinct values p, p^* of DESIRE such that intervening to set DESIRE=p versus DESIRE=p^* reliably yields determinate changes in Oscar's bodily motion. If we intervene through subliminal advertising to make Oscar desire OJ rather than water, then he walks to the refrigerator, not the sink. Apparently, then, DESIRE is causally relevant to Oscar's motor gestures. Admittedly, an intervention from *desiring water* to *desiring*

[7] Fodor (1994) eventually abandons arguments along these lines. See (Yablo, 1997) for citations to the extensive literature on externalism and mental causation.

twater yields no determinate change in Oscar's motor gestures. But that does not show that DESIRE is causally irrelevant to the motor gestures. Causal relevance of *X* to *Y* does not require that *every* change in *X* yield a change in *Y*. It requires only that *some* change in *X* yield a change in *Y* (Woodward, 2003, pp. 65-70). The baseball's velocity is causally relevant to whether the window breaks, even though a slight change in the baseball's velocity would not alter whether the window breaks. Causal relevance of DESIRE requires that *some* change in DESIRE yield distinct outcomes, not that *all* changes in DESIRE yield distinct outcomes.

Consider now a variable DESIRE* with only two possible values: *desiring water* and *desiring twater*. DESIRE* is causally irrelevant to Oscar's motor gestures, because an intervention on DESIRE* yields no determinate change in those gestures:

> Whether Oscar desires water or twater is causally irrelevant to which motor gesture he performs.

The contrast between water-thoughts and twater-thoughts is causally irrelevant to bodily motion. It would be fallacious to conclude that wide content is causally irrelevant to bodily motion. DESIRE and DESIRE* are different variables, with the potential to enter into different causal relations. Given the contrastive nature of causation, the following two theses are consistent:

> Whether one entertains one propositional attitude rather than some other wide attitude with the same "narrow content" is causally irrelevant to behavior.
>
> Which wide propositional attitude one entertains is causally relevant to behavior.

In effect, many arguments for the causal impotence of wide content slide fallaciously from the first thesis to the negation of the second. The arguments establish the causal irrelevance of "external" differences that leave "internal" factors fixed (e.g. *water* versus *twater* desires). They do not thereby establish the causal irrelevance of differences type-identified partly with respect to "external" factors (e.g. *water* versus *OJ* desires).

I have been arguing that interventionism can help defuse various well-known epiphenomenalist worries. A complete discussion would explore these issues much more fully. But I now want to shift attention back towards my central topic: CTM. As I argued above, CTM faces a particularly vexing epiphenomenalist worry. My primary goal is to dissolve that worry by deploying interventionism.

§7. Interventionism and CTM

The main idea behind my solution is to reject PRIM-SYN. The main problem facing my solution is that PRIM-SYN seems extremely plausible. At first blush, it is difficult to see how semantic properties of a primitive processor's input can exert any causal influence upon the processor's syntactic output. To highlight the worry, Block (1990, pp. 149) imagines

> a computational device (say an *and* gate) to which a {"1", "0"} pair --- representing the numbers 1 and 0 --- is input, yielding a "0" output, representing 0. Now the {"1", "0"} input pair would have had just the same effect, viz., the production of "0" as output, even if the "1" and "0" had represented truth and falsity instead of 1 and 0, or even if these symbols had represented black and white, or even if we hadn't been using them to represent anything at all… In common philosophical parlance, the syntax of the representation "screens off" the meaning from having any causal relevance to the output.

This argument looks extremely compelling.

To make matters worse, we can apparently translate Block's argument into a rigorous interventionist argument for PRIM-SYN. The translated argument assumes two widely accepted premises. The first premise, which strikes me as indisputable, runs as follows:

> **SYNTACTIC RULES:** We can specify the mechanical rules governing a primitive processor in syntactic terms, without explicitly mentioning semantic properties such as meaning, truth, or reference.

For instance, we can summarize the rules governing an AND gate as

"1", "1" ⇒ "1"

"1", "0" ⇒ "0"

"0", "1" ⇒ "0"

"0", "0" ⇒ "0"

Similarly, a behavioral transducer conforms to a mechanical rule of the form: given certain syntactic inputs, output certain muscle activations (or muscle contractions, etc.).

The second premise, which seems implicit in Block's discussion, is also widely accepted among proponents of CTM:

> **SEMANTIC NEUTRALITY:** The connection between a syntactic entity and its meaning is arbitrary. A syntactic entity could have had any other meaning, or no meaning at all, without any change in its underlying nature, essence, or identity.

The "0"s and "1"s manipulated by an AND gate could have denoted the numbers 0 and 1, or black and white, or anything else, or nothing at all. Virtually all proponents of CTM regard SEMANTIC NEUTRALITY as a near-truism. For instance, Haugeland (1985) defines a computer as "a *symbol-manipulating machine*" (p. 106), where "the meanings of symbols (e.g. words) are *arbitrary*… in the sense that there is no intrinsic reason for them to be one way rather than another" (p. 91). Similarly, Gallistel and King hold that the symbols manipulated by Turing machines "are to be regarded as purely arbitrary symbols (really data), having no more intrinsic reference than magnetic patterns" (2009, p. 107).[8]

SYNTACTIC RULES and SEMANTIC NEUTRALITY suggest a natural interventionist argument for PRIM-SYN. Let SYN-IN be a variable whose values are possible syntactic inputs to primitive processor *G*. Let variable SYN-OUT have *G*'s possible syntactic outputs as values. SYNTACTIC RULES yields counterfactuals regarding how SYN-OUT would change if we intervened on SYN-IN. For instance, the rules for an AND gate support the following interventionist counterfactuals:

An intervention sets SYN-IN = "1", "1" □→ SYN-OUT = "1"

An intervention sets SYN-IN = "1", "0" □→ SYN-OUT = "0"

These counterfactuals obtain against any background in which *G* instantiates the desired syntactic rule. Thus, *G*'s syntactic input is causally relevant to its syntactic output, as expected.

Now let SEM-IN be a variable that describes possible inputs to *G in semantic terms*. For instance, if *G* is an AND gate, then there is a value of SEM-IN corresponding to a situation where we provide *G* with syntactic inputs whose meanings are *m* and *n*, respectively. Is SEM-IN causally relevant to SYN-OUT? For an interventionist, this is equivalent to asking whether there are distinct values *d*, *d** of SEM-IN and *y*, *y** of SYN-OUT such that

(Δ) An intervention sets SEM-IN=*d* □→ SYN-OUT=*y*

An intervention sets SEM-IN=*d** □→ SYN-OUT=*y**

[8] A few authors, such as Cummins (1996), question SEMANTIC NEUTRALITY. If you share Cummins's skepticism, then rest assured that I do as well. Indeed, (Rescorla, forthcoming b) develops a theory of computation that rejects SEMANTIC NEUTRALITY. However, I think that something like the solution proposed in §§7-11 must inform even those views that reject SEMANTIC NEUTRALITY. One can read this paper as showing how to avoid epiphenomenalism while conceding SEMANTIC NEUTRALITY.

By SEMANTIC NEUTRALITY, we can alter SEM-IN as we please without changing SYN-IN. For any values *d*, *d** of SEM-IN, there are interventions *I*, *I** that set SEM-IN to *d*, *d**, respectively, without changing SYN-IN. By SYNTACTIC RULES, syntactic input determines syntactic output. Thus, precisely the same syntactic output results from *I* and *I**. But then setting SEM-IN to *d* versus *d** yields no determinate change in SYN-OUT. The desired interventionist counterfactuals (Δ) are false. SEM-IN is causally irrelevant to SYN-OUT.

By analogy, suppose my grandmother prefers certain syntactic items to others. Intuitively, her syntactic preferences are causally irrelevant to *G*'s output. To capture this intuition within an interventionist framework, let GRAND-IN be a variable whose values reflect grandmother-preference-properties of *G*'s input (e.g. *receiving as input my grandmother's two favorite syntactic items*). Interventions on GRAND-IN yield no determinate change in SYN-OUT, because one can alter my grandmother's preferences (e.g. by brainwashing) while leaving SYN-IN and SYN-OUT fixed. Thus, GRAND-IN is causally irrelevant to SYN-OUT.

Please distinguish the foregoing argument for PRIM-SYN from the following quite different argument:

> Syntactic input is "causally sufficient" for syntactic output. Thus, semantic input is causally irrelevant to syntactic output (on pain of massive causal overdetermination).

This argument presupposes something like Kim's exclusion principle. In contrast, my interventionist argument does not presuppose anything like the exclusion principle. My argument does *not* maintain that syntax leaves no "causal work" left over for semantics. My argument simply applies interventionism to two doctrines widely accepted by advocates of CTM: SYNTACTIC RULES and SEMANTIC NEUTRALITY.

We can easily extend my interventionist argument from syntactic processors to motor processors. In other words, we can easily construct an interventionist argument for PRIM-MOT. The argument recapitulates my argument for PRIM-SYN, replacing SYN-OUT with a variable that reflects the transducer's possible motor outputs.

In summary, there apparently exist compelling interventionist arguments for PRIM-SYN and PRIM-MOT. How, then, can I reject PRIM-SYN? How can I claim that a primitive processor's semantic input is sometimes causally relevant to its syntactic output?

§8. Indigenous meanings and primitive processors

The key here is to consider a primitive processor not *in isolation* but *as embedded in an overall computational system*. When a primitive processor is embedded in a suitable computational network, §7's argument for PRIM-SYN collapses.

To see why, consider a network containing primitive processors G_1, … G_n wired together in various ways. Assume that time is discrete and indexed by the natural numbers, with 0 representing the moment at which computation begins. A processor's syntactic input at time t determines its syntactic output at time t+1. Each processor fires repeatedly over time, so we must employ variables indexed by time. If G_i is purely syntactic, let $\text{SYN-IN}_{i,\,t}$ and $\text{SYN-OUT}_{i,\,t}$ be variables reflecting its syntactic input and output at time t, and let $\text{SEM-IN}_{i,\,t}$ reflect possible semantic inputs to G_i at time t. Thus, $\text{SYN-IN}_{i,\,t}$ is causally relevant to $\text{SYN-OUT}_{i,\,t+1}$. Is $\text{SEM-IN}_{i,\,t}$ causally relevant to $\text{SYN-OUT}_{i,\,t+1}$?

In many cases, §7 provides a compelling argument against causal relevance. Consider five logic gates G_1, …, G_5 wired together in a simple pattern. The only possible meanings of G_i's syntactic inputs are inherited, assigned by an observer through explicit stipulation or tacit convention. External stipulations and conventions vary independently of the system's syntactic machinations. We can arbitrarily change G_i's semantic input without changing the system's syntactic profile. We can intervene as we please on SEM-IN$_{i,\ t}$ without changing SYN-IN$_{i,\ t}$, and hence without altering SYN-OUT$_{i,\ t+1}$ or any other subsequent syntactic/motor developments. Thus, G_i's semantic input is causally irrelevant to those developments.

I conclude that semantics is causally irrelevant to any computational system that only has inherited meanings, including pocket calculators and probably even desktop computers. Inherited meanings are no more causally relevant than grandmother-preferences.

The situation changes when we consider a computational system that generates indigenous meanings, such as a sophisticated robot or (according to classical cognitive science) the human mind. We now construe SEM-IN$_{i,\ t}$ as a variable whose values are *indigenous* semantic inputs. We construe SEM-IN$_{i,\ t}$=d as signifying that G_i receives at time t a syntactic input with indigenous meaning d. As I argued in §4, we cannot arbitrarily vary indigenous semantic interpretation while preserving the system's syntactic profile. We cannot choose interventions I, I^* that set SEM-IN$_{i,\ t}$ to arbitrary values d, d^* while leaving SYN-IN$_{i,\ t}$ fixed. We cannot arbitrarily alter indigenous semantic input to an embedded processor while holding fixed that processor's syntactic input. In this context, §7's argument seems far less compelling.

Say that a primitive processor is *embedded* iff it is connected to other primitive processors, forming a larger computational system. Say that a primitive processor is *isolated* iff it is not embedded. Distinguish two doctrines:

> **PRIM-SYN$_{\textit{isolated}}$:** Semantic properties of a symbol received as input by an isolated syntactic primitive processor are causally irrelevant to syntactic properties of the processor's output.

> **PRIM-SYN*embedded*:** Semantic properties of a symbol received as input by an embedded syntactic primitive processor are causally irrelevant to syntactic properties of the processor's output.

I endorse PRIM-SYN*isolated*. For an isolated gate, meanings are inherited. We can vary them however we please without changing SYN-IN, and hence without changing SYN-OUT. But PRIM-SYN*isolated* does not entail PRIM-SYN*embedded*. Moreover, WHOL-SYN/MOT entails epiphenomenalism only when combined with PRIM-SYN*embedded*. Block's epiphenomenalist argument conflates PRIM-SYN*isolated* with PRIM*embedded*. The sleight of hand behind the argument is that it first invites us to consider an isolated primitive processor, then surreptitiously shifts attention to an embedded primitive processor.

More precisely, say that a primitive processor is "sensitive" to semantics if semantic input is causally relevant to its syntactic/motor output. We must distinguish two claims:

(a) If a computational network is composed of primitive processors that are individually insensitive to semantics *while embedded in the network*, then semantics is causally irrelevant to the network's syntactic/motor activity.

(b) If we form a computational network from various primitive processors, each of which is individually insensitive to semantics *before we form the network*, then semantics is causally irrelevant to syntactic/motor activity in the resulting network.

(a) is basically a restatement of WHOL-SYN/MOT, which I accept. But (a) does not entail (b). In effect, Block's epiphenomenalist argument slides from (a) to (b). It thereby overlooks the possibility that, by appropriately connecting primitive processors that are insensitive to semantics *in isolation from one another*, we transform some of those processors so that they become individually sensitive to semantics. The transformation reflects the contrast between inherited meanings, which are external to the system's causal structure, and indigenous meanings, which are intimately entangled with that causal structure.

§9. Indigenous meanings as causally relevant

In the previous section, I critiqued §7's argument for PRIM-SYN. I claimed that §7 establishes PRIM-SYN$_{isolated}$ but fails to establish PRIM$_{embedded}$. I now want to argue that, under natural assumptions, PRIM$_{embedded}$ is false. Under natural assumptions, an embedded gate's semantic input is causally relevant to its syntactic output.

I want to capture the following constraint: holding fixed the system's current syntactic profile *except gate G_i's* input, the only way to supply G_i with semantic input d is to supply it with syntactic input x, and the only way to supply it with semantic input d^* is to supply it with syntactic input x^*. Define an *instantaneous description* of a computational network to be a complete description of the network's current state, including the wiring connections between gates and the syntactic, perceptual, or motor inputs and outputs to every gate. Let D_i be a instantaneous description of the system, *minus a description of gate G_i's current input*. Thus, D_i specifies the system's current syntactic, perceptual, and motor profile *excluding G_i's current input*. Let x and x^* be distinct primitive syntactic types. Suppose that necessarily

(1) If a physical system P satisfies D_i, and if P uses syntactic types t and t^* to express indigenous meanings m and m^*, respectively, then $x=t$ and $x^*=t^*$.

Suppose furthermore that

(2) G_i yields a different syntactic output when it receives x as input than when it receives x^* as input (holding fixed any other inputs to G_i).

According to (1), D_i constrains semantic interpretation so that m and m^* are expressed, if at all, only by x and x^*. Perhaps x can mean m, n, or o, but it cannot mean m^*. Perhaps x^* can mean m^*, n^*, or o^*, but it cannot mean m. Only x can mean m, and only x^* can mean m^*. For instance, x might mean either *water* or *twater*, but it cannot mean *OJ*. Similarly, x^* might mean either *OJ* or *Twin-OJ*, but it cannot mean *water*. Only x can mean *water*, and only x^* can mean *OJ*. The system's current syntactic profile constrains indigenous meanings without determining them.

Assume a background in which D_i is constant. Given (1), an intervention that supplies G_i with semantic input m must supply it with syntactic input x, and an intervention that supplies G_i with semantic input m^* must supply it with syntactic input x^*. Given (2), this difference in syntactic input ensures a determinate difference in syntactic output. In other words, assuming a background in which we hold fixed D_i and any other syntactic inputs to G_i:

An intervention supplies G_i with indigenous semantic input m $\Box\!\!\rightarrow$ G_i yields syntactic output y

An intervention supplies G_i with indigenous semantic input m^* $\Box\!\!\rightarrow$ G_i yields syntactic output y^*,

where $y \neq y^*$. Hence, G_i's semantic input is causally relevant to its syntactic output.[9]

To illustrate, let variable DESIRE include all desires in Oscar's and Twin-Oscar's cognitive repertoires. Holding fixed the computational model instantiated by Oscar, it is plausible that there are distinct values p, p^* of DESIRE such that a change from p to p^* requires a change in syntactic properties of Oscar's states. For instance, suppose we intervene through subliminal advertising to make Oscar desire OJ rather than water. Quite plausibly, our

[9] We could retain a version of this argument while altering (1) along the following lines: if a physical system P satisfies D_i, and if P uses syntactic types t and t^* to express indigenous meanings m and m^* *when those types serve as inputs to gate* G_i, then $x=t$ and $x^*=t^*$. This altered formulation allows the indigenous meaning expressed by a syntactic type to vary with location in the network of primitive processors.

intervention alters Oscar's "desire box" by replacing a mental symbol that means *water* with another mental symbol that means *OJ*. This syntactic difference will ramify through Oscar's mental computation. Depending on his other beliefs and desires, he may walk to the refrigerator rather than the sink. Plausibly, then, our intervention on DESIRE yields a determinate change in Oscar's behavior. *Other* interventions on DESIRE, such as a change from *water*-desires to *twater*-desires, may yield no such change. It does not follow that the content of Oscar's desire is causally irrelevant to his behavior. Causal relevance of X to Y does not require that *every* change in X yield a change in Y. It requires only that *some* change in X yield a change in Y. Causal relevance of DESIRE requires only that *some* values of DESIRE yield determinate distinct behaviors, not that *all* values of DESIRE yield determinate distinct behaviors.

The intuitive idea behind my solution is that, under certain conditions, some change in SEM-IN$_{i,\,t}$ requires a determinate change in SYN-IN$_{i,\,t}$. Clause (1) captures this intuitive idea more formally. It is not obvious that (1) is satisfiable. What matters for this paper that (1) is not obviously unsatisfiable. Assuming (1) and (2), the interventionist theory presented in §5 entails that SEM-IN$_{i,\,t}$ is causally relevant to SYN-OUT$_{i,\,t+1}$. Those who seek to infer epiphenomenalism from CTM have the burden of rebutting (1).[10]

How compelling my solution seems will depend partly on one's background theory of mental content. To illustrate, suppose one adopts the Russellian view of content espoused by Fodor (1994): content is individuated by denotation but not "mode of presentation." Fodor urges that Frege cases (such as Hesperus-Phosphorus) involve syntactically distinct mental symbols with identical (Russellian) meanings. Suppose Fodor is correct. Then distinct syntactic types x

[10] My solution shares an underlying idea with Block's: if a computational system generates indigenous meanings, then it instantiates counterfactual patterns that secure content's causal relevance to syntactic/ behavioral activity. Block develops this idea by targeting WHOL, whereas I develop it by targeting PRIM. Also, Block presupposes a functional role theory of meaning, while I do not. Despite these differences, the debt my treatment owes to Block will be evident to anyone familiar with his discussion.

and x^*, as used by the thinker, will often express a single meaning *m*. In that case, providing a processor with input *m* may have no determinate implications for syntactic output, since our intervention may involve either x or x^*. Thus, a solution that depends on (1) appears to have limited, and perhaps even non-existent, scope.

In contrast, suppose we individuate contents in a more fine-grained way. For instance, suppose we follow Frege in individuating content partly by "mode of presentation." Although Fodor dislikes this Fregean approach, it is consistent with CTM. On a Fregean conception, my solution becomes more compelling. Since Fregean senses are so fine-grained, it seems plausible that a sufficiently detailed syntactic description highly constrains Fregean senses expressed by syntactic types, so that a given sense is expressible by at most one such type. More explicitly, suppose that $x_1, \ldots, x_n$ are the primitive syntactic types manipulated by the computational system. Then it is plausible that D_i highly constrains, even if it does not determine, which Fregean senses each type $x_1, \ldots, x_n$ can indigenously express. D_i may not determine whether x_k expresses *water* or *twater*, but it may still mandate that x_k is the only type that can indigenously express either *water* or *twater*. Given such a picture, (1) has widespread applicability.

A complete discussion would more thoroughly investigate these complex issues about mental content. But recall our goal: to investigate whether CTM *in itself* entails epiphenomenalism. Suppose we grant that epiphenomenalism results from combining CTM with some theory of content, such as Fodor's version of Russellianism. In that case, the culprit may be our theory of content, rather than CTM. *In itself*, CTM leaves ample room for (1). To show that CTM avoids epiphenomenalism, we need not show that CTM avoids epiphenomenalism when combined with each possible theory of content.

§10. But isn't syntax doing all the work?

A lingering sense may persist that syntax rather than semantics does the "causal work" in computation. Isn't semantics just "along for the ride," resting atop a purely syntactic process? How can a primitive processor's syntactic input and its semantic input *both* be causally relevant to its syntactic output?

Such questions evoke the exclusion principle. We should regard them warily. According to interventionism, causal relevance requires appropriate counterfactual patterns. We need not apportion some fixed amount of "causal work" between syntax and semantics. Both syntax *and* semantics can be causally relevant to a syntactic output, provided that appropriate counterfactual patterns prevail. Mere intuitions about "causal work" cannot undermine this analysis. The burden falls on those who hold such intuitions to develop them into a systematic objection.

As I suggested in §6, the key issue here is independent manipulability. Two causal variables may both be causally relevant, absent overdetermination, if the two variables are not independently manipulable. We have seen several examples:

THRESH and VELOCITY, reflecting different ways of describing baseball velocity

A "higher-level variable" and its microphysical supervenience base

GAME and SCORE, reflecting different ways of describing a soccer game

In each case, the two variables are not independent causal influences upon the relevant effect, so they can both be relevant without engendering overdetermination. An intervention on the first variable may legitimately alter the second variable.

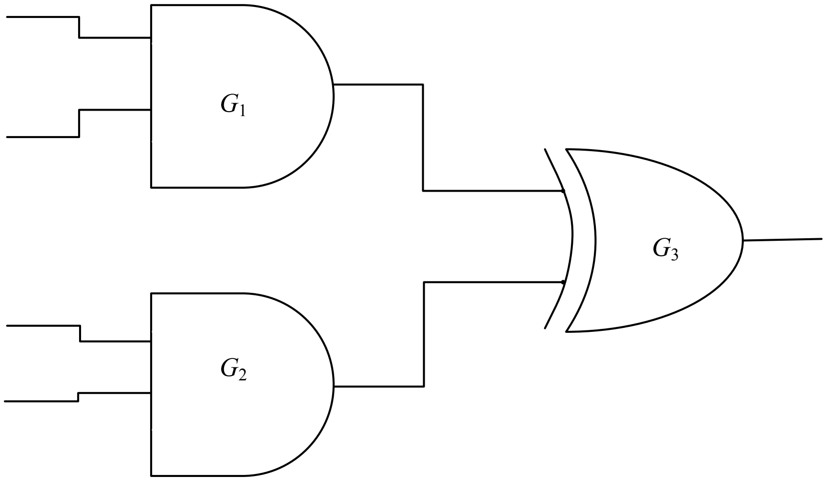
A parallel analysis applies to the relation between syntax and semantics. If a processor is isolated, then SEM-IN and SYN-IN are independently manipulable, because we can arbitrarily change either without changing the other. If a processor G_i is embedded in a suitable computational model, then we may be able to vary SEM-IN$_{i,\ t}$ in *certain* ways without altering SYN-IN$_{i,\ t}$. For instance, we can change a syntactic type's meaning from *water* to *twater* without changing the system's syntactic profile. However, we cannot *arbitrarily* change SEM-IN$_{i,\ t}$ while holding SYN-IN$_{i,\ t}$ fixed. Thus, SEM-IN$_{i,\ t}$ and SYN-IN$_{i,\ t}$ are not independently manipulable. Since the two variables are so intimately linked, an intervention on one variable may legitimately alter the other. Both variables can both be relevant to SYN-OUT$_{i,\ t+1}$, absent overdetermination. The two variables do not compete for causal relevance, any more than mental content and the neural substrate compete. We should no more conclude that syntactic properties do all the "causal work" than we should conclude that neural properties do all the "causal work."

In both the syntactic and neural cases, we block epiphenomenalist worries by denying that semantics varies arbitrarily with respect to the underlying syntactic/neural level. One might therefore describe the mistake behind §7's argument as follows: the argument slides from SEMANTIC NEUTRALITY to the quite different thesis that semantic and syntactic input to an embedded processor are independently manipulable. Even assuming that a syntactic item could have had an arbitrarily different meaning (SEMANTIC NEUTRALITY), it does not follow that the syntactic item *as used by the current computational network* could have had an arbitrarily different meaning. We are studying the causal structure of a fixed computational system, so what matters is the slack between syntax and semantics *as they figure in the fixed computational network*. Assuming that the fixed computational network helps generate indigenous meanings, syntactic and semantic input to the embedded gate do not float free of each other.

§11. The contrast between isolated and embedded gates

A critic might object that syntactic and semantic input are independently manipulable even for an embedded gate, since they float free of each other if we extract the gate from the embedding system. We can embed the gate in a new network, thereby associating it with any new indigenous meanings generated by that network. So syntax and semantics vary arbitrarily with respect to one another, after all.

The proposed objection recapitulates a mistake diagnosed in §8: it studies a gate in isolation, not as embedded in a larger computational network. When studying causation within a computational network, we hold the network itself fixed. We investigate how changes to states of the computational network propagate causal influence through the fixed network. Accordingly, we regiment causal claims through variables that reflect possible internal states of the fixed network. To illustrate, consider a simple network composed of two AND gates (G_1 and G_2) that provide their outputs to an XOR gate (G_3):



Each variable SYN-IN$_{i,\,t}$ reflects the syntactic input to some particular node *within the above wiring diagram*. For instance, the claim

SYN-IN$_{1,\,t}$ is causally relevant to SYN-OUT$_{3,\,t+2}$

corresponds to the English paraphrase:

> Which syntactic input the top AND gate in the wiring diagram receives at time *t* is causally relevant to which syntactic output the XOR gate in the wiring diagram produces at time *t*+2.

(The AND gates receive their input at *t*; they yield their outputs at *t*+1; and the XOR gate yields its output at *t*+2.) Our causal variables reflect possible states *of the fixed computational network*.

In general, a proper interventionist treatment of a computational network should introduce causal variables individuated relative to the network's fixed wiring. Each variable reflects possible syntactic, semantic, perceptual, or motor properties of a node *in the fixed computational network*. SYN-IN$_{i,\,t}$ reflects possible syntactic inputs to gate G_i *as appropriately embedded in the fixed network*. SYN-IN$_{i,\,t}$=*x* just in case:

> G_i is appropriately embedded in the overall computational network.
>
> At time *t*, G_i receives syntactic input *x*.

Similarly, SEM-IN$_{i,\,t}$ reflects possible semantic inputs to gate G_i *as appropriately embedded in the fixed network*. When the network generates its own indigenous meanings, we furthermore demand that those semantic inputs be indigenous. Thus, SEM-IN$_{i,\,t}$=*d* just in case:

> G_i is appropriately embedded in the overall computational network.
>
> At time *t*, G_i receives a syntactic input whose indigenous meaning, as used by the network, is *d*.

We individuate SYN-IN$_{i,\,t}$, SEM-IN$_{i,\,t}$, and other variables partly by relations to the embedding network. This regimentation reflects our interest in the network's fixed causal structure.

Understood in this way, SEM-IN$_{i,\,t}$ and SYN-IN$_{i,\,t}$ are not independently manipulable. Indigenous semantic input to a fixed node in a fixed computational network does not vary arbitrarily from syntactic input to that same node.

We can also regiment causal claims through variables that ignore the embedding context. We can introduce variables SYN-IN$^{isolated}_{i,\,t}$, SYN-OUT$^{isolated}_{i,\,t}$, and so on, that reflect states of gate G_i, whether or not G_i is appropriately embedded in the desired computational network. SYN-IN$^{isolated}_{i,\,t}$ and SEM-IN$^{isolated}_{i,\,t}$ are independently manipulable, since we can vary them arbitrarily by detaching a gate from its present computational network. It follows that SEM-IN$^{isolated}_{i,\,t}$ is causally irrelevant to SYN-OUT$^{isolated}_{i,\,t+1}$. Nevertheless, SEM-IN$_{i,\,t}$ is causally relevant to SYN-OUT$_{i,\,t+1}$, under appropriate assumptions. This discrepancy reflects the contrastive nature of causation. The regimentation

(1) SEM-IN$^{isolated}_{i,\,t}$ is causally relevant to SYN-OUT$^{isolated}_{i,\,t+1}$

corresponds to the English paraphrase

> Which semantic input G_i receives at time t is causally relevant to which syntactic output G_i provides at time t+1

while the regimentation

(2) SEM-IN$_{i,\,t}$ is causally relevant to SYN-OUT$_{i,\,t+1}$

corresponds to the English paraphrase

> Which indigenous semantic input G_i receives at time t while appropriately embedded in the overall computational network is causally relevant to which syntactic output G_i yields at time t+1 while still so embedded.

(1) is false, but (2) is true under appropriate assumptions.

The falsity of (1) provides no support for epiphenomenalism. As I have just urged, the variables cited by (1) are not suitable for studying causal transactions within a computational network. Our focus is the causal structure of a network that contains G_i as a mere cog. We want to investigate semantic changes to the network's internal states, not semantic changes to an isolated gate. Does some semantic change in the network's internal state yield a determinate change in syntactic/motor aspects of subsequent computation? The appropriate variable SEM-IN$_{i,\, t}$ for regimenting this question is individuated partly by relations to the fixed embedding network, because the manipulations that concern us are manipulations relative to that fixed network. By establishing SEM-IN$_{i,\, t}$'s causal relevance to subsequent syntactic effects, we vindicate our desired intuitive thesis: manipulating indigenous semantic properties of a system's internal state is a way of manipulating subsequent syntactic/motor developments. Causal irrelevance of SEM-IN$^{isolated}_{i,\, t}$ casts no doubt upon the desired intuitive thesis.

At this point, readers may worry that a parallel maneuver applies to isolated gates. Let *Int* be a semantic interpretation for syntactic items manipulated by isolated gate *G*. Introduce a new variable SEM-IN^{linguistic}, where SEM-IN^{linguistic}=d iff

G is "appropriately connected" to a linguistic environment that confers *Int* on *G*.

G receives an input whose inherited meaning from that environment is *d*.

We individuate SEM-IN^{linguistic} partly by the surrounding linguistic environment, thereby ensuring that SEM-IN^{linguistic} and SYN-IN are not independently manipulable. Similarly, we can introduce a variable GRAND-IN^{pref} individuated partly by facts about my grandmother's preferences over syntactic items, thereby ensuring that GRAND-IN^{pref} and SYN-IN are not independently manipulable. Should we conclude that inherited meanings and grandmother preferences are causally relevant to syntactic manipulation after all?

We should not. I concede that SEM-IN^{linguistic} and GRAND-IN^{pref} are causally relevant to SYN-OUT. I do not thereby concede that semantic input or grandmother-preference-input is causally relevant to an isolated gate's syntactic output. The reason is that regimentations through SEM-IN^{linguistic} and GRAND-IN^{pref} do not faithfully reflect our intuitive starting point.

What are we asking when we ask whether semantic input to an isolated gate is causally relevant to the gate's syntactic output? From an interventionist perspective, we want to know whether manipulating the gate's semantic input is a way of manipulating its syntactic output. Does some change in *G*'s semantic input yield a determinate change in its syntactic output? No, because we can manipulate semantic properties of *G*'s input by altering the linguistic environment while holding syntactic input fixed. We cannot properly describe such a manipulation through SEM-IN^{linguistic}, whose values presuppose that syntactic items are associated with a fixed interpretation. So SEM-IN^{linguistic} is an inappropriate variable for regimenting intuitive claims about the causal relevance of semantic input. The appropriate variable is SEM-IN^{isolated}, which places no restriction on the surrounding linguistic environment. For any meanings *d*, *d**, there are relevant ways of altering *G*'s semantic input from *d* to *d** that are captured by SEM-IN^{isolated} but not SEM-IN^{linguistic}. Some of those alterations yield no change in SYN-OUT. Thus, manipulating *G*'s semantic input is not a way of manipulating *G*'s syntactic output, even if SEM-IN^{linguistic} is causally relevant to SYN-OUT. Similarly, manipulating grandmother-preference input is not a way of manipulating syntactic output, even if GRAND-IN^{pref} is causally relevant to SYN-OUT.

The situation changes dramatically for an embedded gate G_i. Here, our intuitive starting point crucially involves the fixed computational network. We want to investigate how syntactic and semantic changes to computational states *of the fixed network* propagate causal influence *through the fixed network*. We seek to investigate the fixed network's causal structure. In particular, we seek to answer the following question: is manipulating semantic properties of the network's computational states a way of manipulating subsequent syntactic developments? To answer this question, we choose variables individuated partly with regard to the fixed computational network.

Hence, despite superficial formal similarities between SEM-IN$_{i,\,t}$ and SEM-IN^{linguistic}, these are two quite different variables. Only the former variable figures in suitable regimentations of intuitive claims about causal relevance.[11]

From a sufficiently lofty perspective, the contrast between isolated versus embedded gates may seem negligible. One might argue that, in both cases, a gate inherits meanings from an outside source: either an external observer, or else the computational system in which it is embedded. Either way, meaning is irrelevant to the gate's own causal structure.

I have argued that this lofty perspective obscures vital distinctions. We must sharply differentiate an isolated processor, whose intentionality derives entirely from an external source, and a processor causally immersed in a system that generates its own intentionality. For an isolated processor, the relevant intuitive notion of *manipulating semantic input* includes changes to the surrounding linguistic environment, which yield no change in syntactic output. For an embedded processor, the relevant intuitive notion of *manipulating semantic input* presupposes a background in which the embedding network is constant. Under natural assumptions, a semantic change conducted against that background yields determinate syntactic/motor changes in

[11] One might embrace a more concessive analysis: claims about the causal relevance of an isolated gate's semantic input to its syntactic output are ambiguous between regimentations involving SEM-IN^{isolated} versus SEM-IN^{linguistic}. On the first disambiguation, but not on the second, semantic input is causally irrelevant. I myself doubt that the alleged ambiguity obtains. Most importantly, however, I think that no such ambiguity holds for an embedded processor. In the embedded case, the only proper regimentation involves variables individuated by the overall embedding context, because our concern is the fixed network's causal structure.

subsequent computation. Admittedly, it yields subsequent syntactic/motor changes only because it requires a determinate change in syntactic input. We should not conclude that the change in syntactic input does all the "causal work." Instead, we should conclude that changes in syntax and semantics are sometimes so intimately linked that they do not constitute independent channels of causal influence. By conferring indigenous meanings on its own states, the network renders syntax and semantics sufficiently intertwined that they do not compete for causal relevance.

Proponents of CTM have erred by insisting that only syntax, not semantics, influences elementary syntactic manipulation. Although it can sound like merest common sense to say that primitive processors are "sensitive to syntax, not semantics," that slogan is at best highly misleading. It promotes a fundamentally flawed picture of causal relations in computational systems that help generate their own intentionality. It thereby invites epiphenomenalism. The remedy is to allow that, when a computational system has indigenous meanings, syntax and semantics are *both* relevant to its elementary syntactic manipulations.

Some readers may still feel convinced that syntax rather semantics occupies the "driver's seat," causally speaking. I leave these readers with a two-pronged challenge. Can you articulate a systematic theory of causal relevance as compelling as interventionism? And can you deploy the theory to argue convincingly that content is causally irrelevant to computation's elementary syntactic operations? The current literature does not even begin to meet that challenge.

Works Cited

Bennett, K. 2007. "Mental Causation." *Philosophy Compass* 2: pp. 316-337.

Block, N. 1990. "Can the Mind Change the World?" In *Meaning and Method: Essays in Honor of Hilary Putnam*, ed. G. Boolos. Cambridge: Cambridge University Press.

Burge, T. 2007. *Foundations of Mind*. Oxford: Oxford University Press.

Campbell, J. 2008. "Comment: Psychological Causation without Physical

Causation." In *Philosophical Issues in Psychiatry*, eds. K. Kendler and J. Parnas. Baltimore: Johns Hopkins University Press.

Cummins, R. 1996. *Representations, Targets, and Attitudes*. Cambridge, MA: MIT Press.

Dretske, F. 1988. *Explaining Behavior: Reasons in a World of Causes*. Cambridge: MIT Press.

Egan, F. 2010. "A Modest Role for Content." *Studies in History and Philosophy of Science* 41: pp. 253-259.

Fodor, J. 1981. *Representations*. Cambridge: MIT Press.

---. 1987. *Psychosemantics*. Cambridge: MIT Press.

---. 1990. *A Theory of Content and Other Essays*. Cambridge: MIT Press.

---. 1991. "Replies." In *Meaning in Mind*, eds. B. Loewer and G. Rey. Cambridge: Blackwell.

---. 1994: *The Elm and the Expert*. Cambridge: MIT Press.

---. 2008. *LOT2*. Oxford: Clarendon Press.

Gallistel, C. R. and King, A. 2009. *Memory and the Computational Brain*. Malden: Wiley Blackwell.

Haugeland, J. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

Kazez, J. 1995. "Computationalism and the Causal Role of Content." *Philosophical Studies* 75: pp. 231-260.

Kim, J. 2005. *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

Lewis, D. 2000. "Causation as Influence." *Journal of Philosophy* 97: pp. 182-197.

List, C., and Menzies, P. 2009. "Nonreductive Physicalism and the Limits of the Exclusion Principle." *Journal of Philosophy* 106: pp. 475-502.

Parisien, C. and Thagard, P. 2008. "Robosemantics." *Minds and Machines* 18: pp. 169-178.

Peacocke, C. 1994. "Content, Computation, and Externalism." *Mind and Language* 9: pp. 303-335.

Pearl, J. 2000. *Causation*. Cambridge: Cambridge University Press.

Rescorla, M. Forthcoming a. "Are Computational Transitions Sensitive to Semantics?". *Australasian Journal of Philosophy*.

---. Forthcoming b. "How to Integrate Representation Into Computational Modeling, and Why We Should." *The Journal of Cognitive Science*.

Schaffer, J. 2005. "Contrastive Causation." *Philosophical Review* 114: pp. 297-328.

Searle, J. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3: pp. 417-424.

Shapiro, L. 2010. "Lessons from Causal Exclusion." *Philosophy and Phenomenological Research* 81: pp. 594-604.

Shapiro, L. and Sober, E. 2007. "Epiphenomenalism --- the Dos and the Don'ts." In *Thinking about Causes: From Greek Philosophy to Modern Physics*, eds. G. Wolters and P. Machamer. Pittsburgh: University of Pittsburgh Press.

Spirtes, P., Glymour, C., and Scheines. R. 1993. *Causation, Prediction and Search*. New York: Springer-Verlag.

Woodward, J. 2003. *Making Things Happen*. Oxford: Oxford University Press.

---. 2008. "Mental Causation and Neural Mechanisms." In *Being Reduced*, eds. J. Hohwy and J. Kallestrup. Oxford: Oxford University Press.

Woodward, J. and Hitchcock, C. 2003. "Explanatory Generalizations, Part 1: A Counterfactual Account." *Nous* 37: pp. 1-24

Yablo, S. 1997. "Wide Causation." *Philosophical Perspectives* 11: pp. 251-281.