Sellars on Compatibilism and the Consequence Argument

Jeremy Randel Koons

Georgetown University – Qatar

Philosophical Studies 179 (2022): pp. 2361–2389.

Abstract: No contemporary compatibilist account of free will can be complete unless it engages with the consequence argument. I will argue that Wilfrid Sellars offered an ingenious version of compatibilism that can be used to refute the consequence argument. Unfortunately, due to the significant difficulty of Sellars's writings on free will, his solution has been neglected. I will reconstruct his view here, demonstrating how it represents a powerful challenge to the consequence argument, and tying it to some recent developments in the compatibilist literature.

Keywords: Sellars, Wilfrid; List, Christian; free will; compatibilism; determinism; consequence argument

* * *

1. The Consequence Argument

"It seems to be generally agreed," writes van Inwagen, "that the concept of free will should be understood in terms of the *power* or *ability* of agents to act otherwise than they in fact do. To deny that men have free will is to assert that what a man *does* and what he *can* do

coincide" (1975, p. 188).[1] The consequence argument seeks to demonstrate that if determinism is true, then one can never do otherwise than one actually does—and hence, that one is not free.

Van Inwagen's (1989) formulation of the consequence argument was not the earliest, but has become the standard formulation. Van Inwagen introduces the following notation; I have updated van Inwagen's definition of 'NP' to reflect the revision to this he made (van Inwagen 2015) in response to a counterexample presented by McKay and Johnson (1996):

> □ = standard logical necessity
>
> 'P' stands for any true proposition
>
> 'NP' is a modal expression meaning "P and nothing that any human being is or ever has been able to do is such that if someone were to do it, that person's action might result (could possibly result) in its not being the case that P" (van Inwagen 2015, p. 19)

Borrowing from the resources of standard modal logic, van Inwagen then introduces the following two rules of inference:

> Rule Alpha: From □p deduce Np.
>
> Rule Beta: From Np and N(p ⊃ q) deduce Nq.

As the final stage to setting up his argument, van Inwagen introduces some additional notation and terminology:

> "'L' represent[s] the conjunction into a single proposition of all laws of nature" (van Inwagen 1989, p. 405)
>
> "'P_O' represent[s] a proposition that gives a complete and correct description of the whole world at some instant in the remote past—before there were any human beings" (van Inwagen 1989, p. 405)

[1] Van Inwagen is, of course, aware of Frankfurt's famous essay—published six years prior to this quoted piece—and mentions it in a footnote, but does not defend the 'Principle of Alternate Possibilities' in the quoted piece. He does, of course, address the PAP a few years later (van Inwagen 1978). However, the present piece is concerned not with moral responsibility, but instead with free will as involving the ability to do otherwise than one actually does, and so I will not discuss the PAP.

Given these tools, we can (according to van Inwagen) argue that all of our actions are "humanly unalterable", as van Inwagen (2015) puts it. For the sake of simplicity, I am presenting here a shortened version of the argument:[2]

1. □ ((P_O & L) ⊃ P)	assumption (truth of determinism)
2. N((P_O & L) ⊃ P)	1; Rule Alpha
3. N(P_O & L)	Premise
4. Therefore, NP	2, 3; Rule Beta

No contemporary compatibilist account of free will can fail to engage with the consequence argument; it has fundamentally reshaped the debate since it first appeared.

Interestingly, Sellars's first serious engagement with the problem of free will ("Fatalism and Determinism," hereafter FD) appeared in the same volume where Ginet offers what is generally considered to be the first formal version of the consequence argument. In FD, Sellars offers us a powerful set of tools that can be used to refute the consequence argument. I will unpack these tools and deploy them against the version of the argument best-known to readers, namely, van Inwagen's model consequence argument.

2. The Manifest Image and Intentional Explanation

The first key move Sellars makes is to bring to bear his famous distinction between the scientific image (SI) and the manifest image (MI). The MI, of course, is "roughly…the world as we know it to be in ordinary experience, supplemented by such inductive procedures as remain within the framework. The MI is, in particular, a framework in which the distinctive features of persons are conceptually irreducible to features of nonpersons, e.g., animals and

[2] The full version of the argument is presented in van Inwagen (1989, p. 405), but this shortened version gets all the essential premises on the table. The simplified version was suggested by an anonymous referee for *Philosophical Studies*. Also, as noted in the text, this argument employs a re-defined version of the operator N. McKay and Johnson (1996) presented a counterexample to van Inwagen's Principle Beta; van Inwagen (2015) avoids this counterexample by simply redefining N.

merely material things" (FD, p. 145). This latter claim will be essential to the argument to come. The SI, by contrast, is "man-in-the-world as we anticipate he would be conceived by a unified scientific account, which makes use of the familiar techniques of theory construction" (FD, p. 145).

One way of understanding the difference between the SI and the MI in relation to human behavior is as two orders of intelligibility. As Sellars (in)famously says, "in the dimension of describing and explaining the world, science is the measure of all things, of what is that it is, and of what is not that it is not" ("Empiricism and the Philosophy of Mind" [hereafter, EPM], §41/p. 173). The SI makes things intelligible in virtue of offering causal explanations and elaborating the alethic modal connections obtaining between elements of the causal order. Seen as such, persons are merely one more element in the ongoing swirl of microparticles, neither different nor special. The SI does *not* offer us normative explanations, or even tell us about the normative, because scientific methodology reveals no such objects or powers to us in the world.

The MI can (and does) offer causal explanations—although the MI must always defer to the SI in such matters. But when it comes to persons, a different kind of intelligibility is on offer—*rational* intelligibility. We can only seldom predict someone's actions (and Sellars comments that to say of someone that she is predictable "is not always a compliment" [FD, p. 146]). Rather, this kind of intelligibility is generally retrospective. Thus, even if Smith performs an action we could not predict, Smith's action can nevertheless "be intelligible in terms of being capable of explanation with reference to the practical reasoning and, ultimately, the volition of which it is the expression" (FD, p. 148). For example, although we might not *predict* that Smith would embezzle money from her firm, we might after the fact be able to make this action *rationally intelligible*. We might speculate that Smith was having financial difficulties; or observe that her partner or child had accrued unmanageable medical bills; or

note that she openly aspired to a higher standard of living; and so on. These rational explanations aspire at the same time to be causal explanations—Sellars, after all, holds that reasons can be causes[3]—but these are not the kind of causal explanations that populate the SI. They offer a fundamentally different kind of intelligibility.

MI explanations of human behavior are in terms of mental states—chiefly beliefs, intentions and volitions—and Sellars devotes several pages of FD to the discussion of volitions. This might seem like a puzzling digression, unless we understand that what Sellars is doing here is elaborating the framework within which human behavior is intelligible to us. Describing the role of these mental states in such explanations, Sellars writes, "The role of inner episodes in the manifest image can be compared to that of theoretical entities in the scientific image. They are, however, elements of an *autonomous* framework—not a speculative extension of microphysics—which carries the imprint of the specifically human observable behavior they are designed to explain" (FD, p. 148).

At least two things about this passage bear emphasizing. The first is the analogy Sellars draws between inner states and theoretical postulates; this analogy should be familiar to readers of EPM and the Myth of Jones. The second—and more important—point concerns the *autonomy* of the framework of inner episodes. Explanations of human behavior in terms of mental states cannot simply be replaced with some other kind of talk—say, talk about chemistry, or neurology. There are two related reasons for this. The first concerns multiple realizability. Even if you have a standard range of naturalist commitments about the mental—that mental states can be functionally defined, that the mental supervenes on the physical, etc.—the multiple realizability of the mental means that any description of the mental at the level of

[3] See, for example, AAE.

the physical will be wildly disjunctive. For this reason alone, the mental should be regarded is irreducible to the physical, as forming an autonomous level of discourse.[4]

The second reason supporting the autonomy of the framework of mental episodes relates to what we are doing when we offer explanations of human action: Again, we are trying to generate a certain type of intelligibility. But this type of intelligibility is lost if we abandon intentional explanation. As Anscombe famously notes, to describe an action as intentional is to say that it is the kind of action to which a certain kind of *why* question is appropriate and to which a certain kind of reasoning—namely, practical reasoning—applies. Thus, to use Anscombe's example, I might describe a person's action as follows: He is moving the pump handle up and down. In describing this action as intentional, I am placing it within a calculative order: I am saying that it is appropriate to ask *why*—to what *end* or *purpose*—he is moving the pump handle up and down. And we can nest this description within further descriptions that not only describe the action but also give the further purposes he intends in performing the action:

He is moving the pump handle up and down
Why? In order to pump water to the house
Why? In order to poison the inhabitants of the house
Why? In order to prevent the inhabitants from carrying out their evil plan
Etc.

Eventually, the nesting comes to an end because we arrive at an "answer [that] does not describe something he is doing in moving his arm but only something that he is moving his arm in order to do" (Wiseman 2016, p. 126). This kind of intelligibility—rational intelligibility—is conceptually tied to intentional explanation; to give an intentional explanation of an action *just is* to place an action within the *calculative order*. Thus, to abandon

[4] For a detailed Sellarsian argument for the irreducibility of MI concepts (including intentional concepts) to SI concepts, see Chapter 2 of Koons (2019).

one is to abandon the other; to abandon intentional explanation is to forego rational intelligibility.

I said above that the two reasons for the ineliminability of intentional discourse are interrelated; let me explain why. When we are talking about something that is functionally-defined, the specific mechanism through which this function is achieved is, for many purposes, irrelevant. If I need to converse with an auto parts dealer about what part needs replacing on my car, the specific mechanism by which (say) my fuel pump operates—whether it is mechanical or electrical, operates *via* a diaphragm or a plunger, etc.—is irrelevant to this conversation. What matters is the role the part plays in the functional economy of the automobile—it is the part that goes *here* and *functions* to move fuel from the fuel tank to the carburetor. Trying to understand the part through a complete technical description of its operation would, for starters, massively increase the computational load required for thought and communication. Functional descriptions both unify a number of otherwise disparate phenomena under a single heading, and reduce the computational load associated with understanding the items in question.[5] That is: Functional explanation makes rationally intelligible the operation of something whose design can be physically realized in a number of different ways; hence, the relation between rational intelligibility and multiple realizability.[6]

Otherwise stated, by defining the item in terms of the role it plays in the functional economy of the system, functional description makes the item intelligible to us in a way that a complex technical description would not. For starters, even trying to understand the part in

[5] In this connection, an important role that concepts play in cognition and communication is in aggregating a number of inferential proprieties under a single term, thereby allowing thought and communication to 'move' larger packets of information more cheaply. Connected to this point, and to the point made above in the text, it is worth quoting Sellars from "Volitions, Re-Affirmed": "To every explanation in mentalistic terms there corresponds *in principle* an explanation in non-mentalistic terms. I say 'in principle' because, in even the simplest cases, the complexity of such an explanation would be unmanageable" (VR §19/p. 51). Of course, these two explanations will fundamentally differ, one being a rational explanation, and the other one a purely causal explanation.

[6] Sellars makes the connection between (a) mental states being defined by normative functional roles and (b) multiple realizability in a number of places; see, e.g., SRLG, MP (esp. pp. 237ff).

terms of a technical description of its operation would likely not avoid reference to further functionally-defined items. As you can see from the diagram of a diaphragm-style fuel pump (Figure 1), the constituent parts of a fuel pump are themselves given functional labels (e.g., inlet, outlet, diaphragm, lever, etc.); and so you don't eliminate functional descriptions simply by going to a more technical level of description. (I suspect the same is true with manifest-image concepts, where any move from folk psychology to, say, neurology would just replace one set of functionally-defined items—such as belief and desires—with another—such as neurons and synapses.[7])

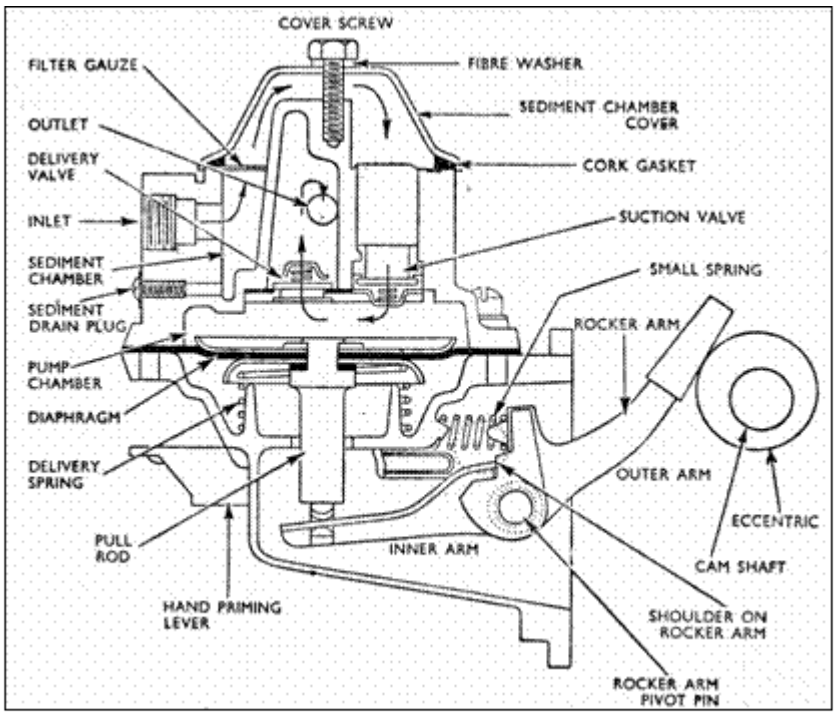

Figure 1: Diaphragm-style fuel pump[8]

[7] A reviewer has commented that it may not be obvious that "neurons and synapses" are functional terms. I would argue, however, that biological subsystems (including neurons and synapses) are functionally-defined, and therefore multiply-realizable: This is why (say) an artificial heart (or an artificial hip) is still a heart, or a hip. What matters is the job done; the physical constitution of what does the job is important only (!) insofar as in enables the doing of the job.

[8] Image source: <http://www.austin7.org/Technical%20Articles/AC%20Mechanical%20Fuel%20Pump/>. Image reused by kind permission of the Cornwall Austin Seven Club: <http://www.austin7.org/index.html>.

Further, the operation of this item is not intelligible in isolation from other parts of the car—the fuel tank, the carburetor, etc.; and so making the operation of the fuel pump intelligible would require a detailed technical description of these other items as well. Again, this is too cognitively burdensome; it is doubtful that the outcome would be intelligibility. And most crucially, the connection among these items—fuel, fuel tank, fuel pump, carburetor—*is precisely a functional relationship*. Thus, understanding the relation among them—and the operation of the fuel pump in relation to them—means understanding its function in the overall mechanical economy of the car. That is what it is to make the operation of the fuel pump intelligible—it is not separable from explaining its function. So not only would a detailed technical description of its function (along with the function of the systems with which it connects) be too cognitively burdensome to effect intelligibility in most people, it wouldn't be a description at the right level—or even of the right *phenomenon*—in the first place; it wouldn't be an account of the *function* of the pump. Thus, to describe the fuel pump in other than functional terms is *not to describe a fuel pump as such*, but merely to hope that the person receiving this description has sufficient automotive knowledge that they can make the functional inference and supply the missing conceptual information that your (non-functional) description lacks.

An analogous argument can be made with regard to mental states in the MI. Mental states (like fuel pumps) are functionally-defined; and this provides an argument for the autonomy of mentalistic discourse.[9] For example, to attribute a belief or commitment to

[9] Sellars reiterates his support for a functional notion of the mental—and the corresponding commitment to multiple realizability—in several places: "Concepts pertaining to mental acts are 'functional' in a way which leaves open the question as to the 'qualitative' or, as I prefer to say, contentual character of the items that function in such a way as to be the kind of mental act they are" ("Metaphysics and the Concept of a Person," p. 237); "the classification of thoughts, construed as classical mental episodes, permits of no such easy retreat to a non-functional level. Roughly, our classification of thoughts, construed as episodes which belong to a framework which *explains* the kaleidoscopic shifts of sayings and propensities to say, is almost purely functional. We have only the foggiest notion of what kinds of episodes, non-functionally described, perform the relevant functions" (AAE §34/p. 189). He does say, in the immediately preceding section (§33) of AAE that you can talk about the correlates of mental acts in non-functional terms. But to talk about such states, non-functionally described, is not to talk about conceptually-contentful mental acts.

someone is to attribute to her a range of counterfactual functionings: For example, she would maintain the commitment under *these* circumstances but not *those*; or given that she is committed to P, she would not assent to Q. Sellars puts the point bluntly:

> When we characterize a person's utterance by using a quotation, we are implying that the utterance is an instance of certain specific ways of functioning. For example, it would be absurd to say:
>
> Tom said (as contrasted with uttered the noises) 'It is not raining' but has no propensity to avoid saying 'it is raining and it is not raining.'
>
> Thus to characterize a person's utterances by quoting sentences containing logical words is to imply that the corresponding sounds function properly in the verbal behavior in question and to imply that the uniformities characteristic of these ways of functioning are present in his sayings and proximate dispositions to say. (Sellars, "Actions and Events" [hereafter, AAE], p. 187/§29)

Brandom states the point with a bit more elaboration:

> The hungry lioness would still chase the antelope if it were Tuesday or the beetle on the distant tree crawled slightly further up the branch, but not if the lioness's heart were to stop beating. The point is not that there is any particular set of such discriminations that one must be able to make in order to count as deploying the concepts involved. It is that if one can make *no* such practical assessments of the counterfactual robustness of material inferences involving those concepts, one could not count as having mastered them. (Brandom 2015, p. 142)

This supports the autonomy of the mental in multiple ways. First, the holistic interdependence of various commitment-states to each other means that an analysis of these in terms of underlying (say) neurological states will quickly become unmanageably complex. (It is a familiar point that similar worries about the holistic interdependence of mental states doomed behavioristic analyses of mental states.) Again, a point of concepts—and the point of explaining behavior in terms of mental states—is intelligibility and comprehensibility, and to attempt an explanation in terms of something too complicated for human cognition fails at this simple task. Just as importantly, as noted above, functional explanation makes rationally intelligible the operation of something whose design can be physically realized in a number of different ways; hence, again, the relation between rational intelligibility and multiple realizability.

A second way in which the Sellars/Brandom point argues in favor of the autonomy of the mental is that these functional characterizations are largely normative, and not merely descriptive. I don't merely *predict* that if Jones asserts that it is raining, he will refrain from asserting that it is not raining; I take it that he *ought* to refrain. I don't merely *predict* that if Smith believes the lioness's heart has stopped beating, then she will not chase the antelope; I take it that she *ought* to believe this. There is right and wrong with respect to the MI; there is error. To retreat to the framework of the SI is to lose sight of conceptual content altogether, because while it might be absurd for Smith to assert both P and not-P, it cannot be absurd (from a purely empirical standpoint) to have this set of neurons firing and also that set of neurons firing. Thus, a framework within which human behavior can be made intelligible (or at least within which we can explain its unintelligibility) is lost to us if we move to the SI.[10]

But most fundamentally, to explain human behavior in terms of the SI is *simply to explain the wrong phenomena*. It is like trying to explain fuel pumps in terms of microphysical particles. You are simply failing to capture the salient dynamic—namely, the *functional role* of the pump in the complex automotive system. The operation of a fuel pump is unintelligible independent of its functional role in the overall design economy of the automobile. Similarly, attempting to offer an SI explanation of human behavior is *to fail to explain human behavior as such*, which has an inherently intentional character and is explained *via* placement in a rational calculative order. To fail to understand the normative proprieties governing, for example, a belief—the way in which the belief commits one to other beliefs, is incompatible with others, the way it is connected to norms of assertion, and so on—is simply to fail to

[10] To head off an objection, I do not think that the assumption of normativity here begs the question against the incompatibilist (who after all is arguing that *moral* normativity *presupposes* free will). The question of free will is not obviously relevant to conceptual normativity, which is governed by ought-to-bes, not ought-to-dos. Sellars's considered opinion is that inferring is not under our voluntary control, and is not properly speaking an *action*. And again, the point here is intelligibility—Jones's behavior is intelligible because he conforms to certain normative proprieties, such as inferring 'Simba is a mammal' from 'Simba is a lion'. Smith's behavior is unintelligible precisely because she fails to conform to these proprieties—precisely because she is disposed, say, to assert both "It is raining" and "It is not raining" at the same time. The *normativity* of the MI and the sense in which it makes human behavior *intelligible* are not separate dimensions of this explanatory framework.

understand the state *as a belief*. It is for this reason that Sellars—like Brandom who follows him—thinks that the salient difference is between treating something as merely being *strictly* in the space of causes versus treating it as *also* being in the space of reasons.[11] To treat something as an *agent* is to treat it as being in the *space of reasons*. Thus, Brandom writes, "We treat some bit of behavior as the expression of a linguistic social practice rather than an objective process when we *translate* it, rather than offering a causal explanation of it" (Brandom 1979, p. 190). To describe a person in terms of chemical or neural processes is *not to offer an intentional description, and hence not to describe a person at all*. Again, failure to use functional/intentional language means you simply are failing even to talk about what you are aiming to talk about; it is to fail to even talk about action in the first place.

A retreat to the deterministic, SI picture of humans is thus, for Sellars, to make the radical move of *denying that there are actions as such*. Of course, the argument between the compatibilist and the incompatibilist isn't between whether there are or are not actions; it is between whether these actions are or are not free. Thus, the argument between compatibilists and incompatibilists must take place *within* the MI framework where the category of intentional explanation is ineliminable. This intermediate conclusion will be crucial to the future progress of the argument. Sellars's next step will be to argue that actions, by their nature, do not admit of deterministic explanation. But let us not get ahead of ourselves.

3. Determinism and Sellars's Two Images

There are key differences between the SI and the MI, differences relevant to the issue of free will. One important difference concerns physical determinism. Many philosophers treat physical determinism as a conclusion of scientific inquiry; Sellars, by contrast, holds that

[11] It is important that for Sellars the space of reasons is *within* the space of causes; he writes that "action in terms of *reasons* is a *special case of* explanation in terms of (occurrent) *causes*" (RD, p. 178). But to treat something as *merely* in the space of causes is not to treat it as an agent at all. To treat something as an agent is to offer a specific kind of causal explanation of its behavior—namely, *rational explanation*.

physical determinism is a presupposition—a "framework principle", as he calls it—of the SI. As for our place in the SI, "the scientific image contains a picture of man as part and parcel of a deterministic order" (FD, p. 145). By contrast, "the manifest image is not able, out of its own resources, to generate a deterministic picture" of persons in the world (FD, p. 147).[12]

This latter fact is only occasionally appreciated by philosophers working on the problem of free will. Van Inwagen, for example, in commenting on the 'laws of nature' as they appear in his version of the consequence argument, remarks that "conceivably, psychological laws, including laws (if such there be) about the voluntary behavior of rational agents, might be included under this term" (1975, p. 187). Ginet takes a more plausible view—and one more in line with Sellars's. In commenting on the appropriate language to describe the antecedent circumstances that "contingently necessitate" a person's actions, Ginet denies that such circumstances can be described in psychological terms—e.g., "desires, intentions, beliefs, and the like…because I suspect that such psychological factors, though commonly used in explaining behavior, can never be regarded as contingently necessitating that behavior" (1966, p. 95).

There is a principled reason for thinking that determinism, while true in the SI, is not true in the intentional/psychological discourse which finds its home in the MI. This reason is grounded in the multiple realizability of functionally-characterized states, which we have already discussed at great length. There is no unique neurological state that is correlated with my belief that-P; and hence no unique causal consequence of my belief that-P.

Consider the matter this way. Suppose we (*per impossibile*) could give a complete psychological characterization of a person at time t: Smith has, at time t, mental states $M_1 \ldots M_n$. Now if we suppose—as did Sellars—that some kind of token identity theory was true, then

[12] See also "Reply to Alan Donagan" (RD), p. 182: "Only the confusions of the vulgar determinist could lend plausibility to the idea that the conceptual resources of the (first level) manifest image are rich enough to generate (even in principle) a universal derivability of events, including human actions, from antecedent events."

these mental states are all token-identical with a set of brain states $B_1 \ldots B_n$. But, of course, the fact that I am in mental states $M_1 \ldots M_n$ underdetermines what brain states I currently possess. Thus, I might be in brain states $B_1 \ldots B_n$. Or I might be in brain states $B_\alpha \ldots B_\nu$. Or, etc. Of course, there is presumably a narrow range of brain states that can realize the corresponding mental state. But the *total* set of brain states that can realize my *total* set of mental states $M_1 \ldots M_n$ will vary greatly. Thus, the set of brain states I will be in at time $t + \Delta t$ will also vary widely—and so we can expect some variability as to the set of mental states I will possess at time $t + \Delta t$. Thus, when viewed *from the manifest image, at the level of the mental*, determinism will turn out to be false. From the fact that I am in mental states $M_1 \ldots M_n$ at time t, it in no way follows that at time $t + \Delta t$ I will be in mental states $M_\alpha \ldots M_\nu$.[13]

It is against this background that one must understand Sellars's elaboration of the common-sense concepts that define the space of voluntary action. Sellars gives his definitions in very technical form; I will give them more conversational form, with the exception of his first formulation of the principle of determinism, which I will state as he does. Here are the symbols he uses in defining the principle of determinism:

A = an action
t = a specific time
$t' = t - \Delta t$
N = absolute necessity
a = "the unspecified antecedent state of the universe relevant to x's doing A at t" (FD, p. 168)
x = an agent

The principle of determinism is stated as follows (FD, p. 168):

PD-I $\quad A(x, t) \rightarrow a(x, t') \bullet N[a(x,t') \rightarrow A(x,t)]$

Roughly: If x does A at t, then x participated in a state of the universe, a, at t', and it is necessary that (given x's participation in a at t', that x does A at t).

[13] For a contemporary and sophisticated version of this argument, see List (2019). I will discuss List's argument in section 5 below.

Is determinism compatible with the ability to do otherwise? Sellars thinks the answer is 'yes'.[14] Key to this is his analysis of involuntary action. Although Sellars discusses compulsion—and Alan Donagan accuses Sellars of attributing to contemporary libertarians the fallacy of conflating causation and compulsion—Sellars does not think that an analysis of compulsion is key to dissolving the problem of free will. Sellars writes, "To say of an action, A, that it is under the agent's voluntary control is to say that if, just before doing A, the agent had willed not to do A, he would not have done A… Even a compelled action must be voluntary in the sense defined" (FD, pp. 159-60). The converse of voluntary action, then, is not compelled action—*it is behavior which does not count as an action at all*, since it is not under the agent's voluntary control. Thus, writes Sellars, "To go out the window propelled by a team of professional wrestlers is not to *do* but to *suffer*, to be a patient rather than an agent" (FD, p. 160). Or, perhaps more to the point, Sellars writes, "Consider, for example, the case of the narcotics addict who succumbs to clear and present temptation. If we say that the action was not under his voluntary control, the concept of compulsion is misapplied, and we should rather say that the addict didn't *do* anything but rather, so to speak, blinked or twitched" (FD, p. 160). Sellars says that "the concept of compulsion is misapplied" because if we genuinely think this is a case of compulsion, then the resulting event isn't an *action*—so we cannot sensibly talk of its being voluntary or involuntary; these categories simply don't apply.

However, doesn't Sellars's point here merely sharpen our worry? That is, one might worry that Sellars's strong view—that non-voluntary behavior doesn't even count as action in

[14] This is how Sellars casts the debate, but it is not really a felicitous way of stating the problem. As we will see below, Sellars's argument ultimately amounts to the claim that there is no way to formulate the second conjunct of PD-I's consequent—$N[a(x,t') \rightarrow A(x,t)]$—so that it comes out true. Thus, it is more accurate to say that the principle of determinism cannot even be stated in a way that mentions actions *qua* events within the MI framework. The principle of determinism is true *within the SI*, but as such cannot mention agents, nor actions, nor contain any intentional elements. It is not entirely clear (to me, at least) whether Sellars recognized this consequence of his argument, but it follows naturally from his other commitments, and he should welcome it.

the first place—moves the determinist from the frying pan into the fire, insofar as determinism threatens to render our actions not merely unfree, but *non-actions*.

Again, though, Sellars is adamant that our analysis of action must proceed from within the MI, and vis-à-vis our actions, determinism is false within the MI: When viewed from within the MI, there is a range of actions that *really are* up to us. Like the libertarian, Sellars understands 'up to us' as consisting of both a both a negative condition ("that it is not causally pre-determined" ["Reply to Alan Donagan" (hereafter RD), p. 162]) and a positive condition (that it is caused by the self or agent). Unlike the libertarian, Sellars thinks that both of these conditions can be satisfied even if physical determinism is true—chiefly because (let us remind ourselves) physical determinism is a framework principle of the SI, whereas explanations of human action in terms of intentional behavior belong to the framework of the MI. I will only discuss the negative condition here, as that is the condition most relevant for the consequence argument; discussion of the positive condition will have to await another occasion.[15]

3.1. Sellars on the Negative Condition

The negative condition on free action is "that it is not causally pre-determined" (RD, p. 162). First, a note on what Sellars means by 'determinism': He distinguishes (FD, pp. 143-4) between two senses of 'predictability', epistemic vs. logical. Epistemic predictability is "predictability by a predictor in the system," so is "bound up with difficulties of the type explored by Gödel" (143). Logical predictability, on the other hand, "involves the derivability of a description of its state at a later time from a description of its state at an earlier time" (143-4). Sellars continues, "it does seem to me that this is what philosophers concerned with the free will and determinism have had in mind" (p. 144) noting, though, that to call this 'logical predictability' is "a misuse of the term 'predictability'" (p. 144). Thus, by 'determinism,'

[15] Koons (forthcoming).

Sellars means what contemporary philosophers of free will do, namely, the question of whether propositions about current states of the universe (including statements about human actions) are derivable from statements about antecedent laws and conditions. And a denial of causal pre-determination is simply a denial that there is such a derivation, in the case of a particular action. Thus, it is simply a denial of premise 1 of the consequence argument.

Recall Sellars's own formulation of the principle of determinism, stated earlier (from FD, p. 168):

PD-I $A(x, t) \rightarrow a(x,t') \bullet N[a(x,t') \rightarrow A(x,t)]$

Roughly: If x does A t, then x participated in a state of the universe, a, at t′, and it is necessary that (given x's participation in a at t′, that x does A at t).

I am simplifying Sellars's reply by unpacking it in conventional language and then reapplying it to a formal version of the consequence argument with which readers are more familiar: van Inwagen's modal version (1983). In the context of responding to considerations similar to those raised by familiar forms of the consequence argument, Sellars introduces various concepts surrounding voluntary action such as 'able', in the broad sense. Sellars formally defines it using some symbols introduced later in FD; to simplify, I will present it using the symbols defined earlier in section 3:

CAP $[A(x,t)] =_{def} \sim(\exists a)a(x,t') \bullet N[a(x,t') \rightarrow \sim A(x,t)]$[16]

Roughly: x is able to do A at t iff there is no antecedent state of the universe such that x participated in a state of the universe, a, at t′, and it is necessary that (given x's participation in a at t′, that x does not do A at t).

Now, examination of these two principles will make it clear that the definition of 'able' is simple the denial of the consequent of the principle of determinism. Thus, it is tempting, says Sellars, when looking at *any* action, A, taken by *any* person, to consider some action—

[16] Adapted from the version introduced at FD, p. 173.

A′—incompatible with A. We can then take PD-I and our definition of 'able' and substitute for 'circumstance' the total antecedent state of the universe at t′. Using these two principles, we can easily derive the conclusion that x was not able to do A′ at t—and prove that no one is ever able to do otherwise than she ever does. Doesn't this demonstrate the universal falsity of free will?

This gambit fails, for the simple reason that the incompatibilist argument treats *a*—the "unspecified antecedent state of the universe relevant to x's doing A at t"—as though it were a circumstance of action we may allowably substitute into the circumstance clause in our definition of 'able'. Sellars argues that this represents an illicit attempt to introduce elements from the SI framework into MI-framework explanations; *a* is not a circumstance, but rather a "pseudo-circumstance", as Sellars calls it, and is therefore not an allowable substituend into action-explanations. The antecedent of an action-explanation has to be couched in the language of the MI—in which case the explanation will be non-deterministic. Thus, Sellars will deny that we can legitimately make statements of the form:

$$N[a(x,t') \rightarrow \sim A(x,t)]$$

where the antecedent is couched in SI language and the consequent in the MI language of actions, intentions and volitions.

Thus, Sellars's gambit should be clear: Actions are not causally pre-determined, because 'action' is a concept belonging to the MI framework of persons, and determinism is not true with respect to this framework. One might do A in C without C depriving one of the capacity to do A′. As Sellars argues in RD,

> It is part of the logical grammar of capacity concepts that systems (macro- or micro-) do not lose at *t* their ability to *φ* simply by virtue of the fact that at *t* they are not *φ*ing. An automobile does not lose its capacity to go ten miles per hour when it is going one hundred miles per hour, however necessary the latter speed may be, relative to the antecedent state of the universe. In other words, we distinguish between those happenings by virtue of which a system fails to *φ* at *t without losing its capacity to φ*, and those

> other happenings by virtue of which it fails to φ at t *by being caused to lose, for a longer or shorter time, its capacity to* φ. (RD, p. 154)

Of course, this leaves open the possibility that some circumstances *do* cause a person to lose the capacity to do A′, or whatever. And Sellars certainly recognizes possible circumstances where "the hearing of an explosion causes a person to lose the capacity to deliberate, or the seeing of an available batch of heroin causes an addict to be unable to will to refrain from grabbing it" (RD, p. 154); but he denies that *all* circumstances bear this relation to action. *Even in the SI*, as he has argued, there is a distinction between circumstances that cause substances to lose a capacity—e.g., a circumstance which causes an elastic substance to lose temporarily its ability to respond to tension—and those that do not. Thus, it is a mistake to assimilate all action to the model of addiction (or even habit).

Thus, Sellars is able to argue that the negative condition is often satisfied with human action. *Real* circumstances (as opposed to pseudocircumstances) are compatible with an agent's doing—and indeed, willing—otherwise than she in fact does; and real circumstances do not prevent an agent from doing or willing otherwise than she actually does.

It is worth mentioning at this point that a consequence of Sellars's argument—one that it is not clear he recognized—is that various formulations of the principle of determinism offered in FD all turn out to be false. Sellars often casts the discussion in FD in terms of whether determinism is compatible with the ability to do otherwise; and he generally treats determinism as being captured by his various formulations such as PD-I. But PD-I cannot be true, for the simple reason that the second conjunct of the consequent—$N[a(x,t') \rightarrow A(x,t)]$—cannot be formulated so that it comes out true (for reasons rehearsed in the above section). The principle of determinism is true as a framework principle of the SI, but as such it mentions neither actions, nor agents, nor intentionality. As such, the principle of determinism *as true in the SI* will not take the form of PD-I. The principle of determinism cannot even be stated in a way that mentions actions *qua* events within the MI framework. Thus, Sellars's claim is more

accurately rendered as: Determinism (in the SI) is compatible with the ability to do otherwise (in the MI).

4. A Sellarsian Reply to the Consequence Argument

As just noted, Sellars thinks that the negative condition on freedom can be satisfied by his compatibilist account. (So can the positive condition, but for reasons of scope, I have omitted that discussion from this piece.) This, I believe, gives us the key to how Sellars would reply to contemporary formulations of the consequence argument. A clue is to be found in a cryptic comment Sellars makes in FD when discussing various possible deployments of the principle of determinism in incompatibilist arguments:

> Notice, however, the shift from
> . . . which it was possible *that* x do at *t*
> to
> . . . which it was possible *for x to do* at *t*
> This shift is a telltale symptom of the confusion which is being made. (FD, p. 169)

In short, the incompatibilist argues that because determinism entails x's participating in a state of the universe α at t means it was impossible *that* x do A at t′, x's participating in a state of the universe α at t means it was impossible *for* x to do A at t′. While initially puzzling, I believe what Sellars is driving at here is that claims about necessity and possibility have different truth values at different levels of description. When we say it was impossible *that* x do A at t′, we are making a claim that given a deterministic world view, the fact that x's participation in α at t necessitates certain subsequent behaviors. The 'that' signifies that we are taking the impersonal standpoint of descriptive explanation; and hence this is a claim that is being made from within the scientific image; for determinism is a principle of that framework, and no other. Thus, the consequence argument must itself be understood (by Sellars) as stated in the language of the scientific image; and as such, it simply isn't relevant to explanations of human action. When we explain human action (the clause signaled by the use of 'for'), we are doing so from

the framework of intentional explanation, where we attempt to make such action rationally intelligible.

This conclusion—that the consequence argument can only be made from within the SI, and is therefore irrelevant to discussions of action and freedom—may seem to fly in the face of reason, and will need further justification. But it is a conclusion that follows from what has already been said. Sellars does not deny that humans are part and parcel of the causal-deterministic order; but he denies that determinism belongs to the framework of actions and persons. So whatever we may say about humans as determined beings is true, but irrelevant to our understanding of human rational agency. An attempt—like the consequence argument—to paint human action as necessitated by prior events cannot be coherently formulated because it attempts to derive *actions* (which belong to the MI) from *pseudocircumstances* (which belong to the SI). This is precisely what the MI framework of action-language does not allow.

Consider the first premise of the consequence argument:

1. $\Box ((P_O \ \& \ L) \supset P)$

As already indicated, "the manifest image is not able, out of its own resources, to generate a deterministic picture" of persons in the world (FD, p. 147). Thus, the antecedent of this conditional must be stated in terms of the SI. And, indeed, this is how van Inwagen envisions this—as a conjunction of a proposition giving a complete description of a state of the world with a conjunction of all of the laws of nature. However, the consequent of the conditional will be in the language of the MI—it will be cast in terms of intentions and volitions.

The consequence argument, if it is to be of any interest at all, cannot only apply to one, or a few, actions. It is supposed to be a truth about actions generally. But then, Sellars would argue, premise 1 cannot be stated as a general truth, because it cannot be a truth about actions *per se*. For Sellars, again, it is a conceptual matter that actions are in the space of reasons—recall Sellars's earlier claim that "To go out the window propelled by a team of professional

wrestlers is not to *do* but to *suffer*, to be a patient rather than an agent" (FD, p. 160). Therefore, any system of behavior that is systematically described in *purely* causal-scientific terms (as opposed to being given a *causal-rational*—and thus *normative* and *intentional*—explanation) is thereby not a system of *actions*, at all. Thus, the advocate of the consequence argument faces a trilemma:

> (1) She can state premise 1 purely in terms of the MI, in which case it is false, since determinism is false within the MI.
> (2) She can state premise 1 in 'mixed' form, with an antecedent couched in SI terms and a consequence couched in MI terms. But then it is ill-formed, for reasons explained in the previous paragraph.
> (3) She can state premise 1 in purely SI terms. But then we are not talking about actions at all, and so Sellars will continue to claim that the MI framework of intentions, volitions, etc., is indeterministic.

No doubt, the advocate of the consequence argument will not be satisfied with this response; and will argue that given some basic assumptions (e.g., supervenience), the consequence argument can be shown to be fatal to Sellars's account.[17] Consider the matter this way: We can agree (by stipulation) that agency facts supervene upon scientific facts. Thus, a complete description of the state of the universe at a particular time will necessitate some particular set of facts about agency. We can define the 'N'-operator in a way that it will apply cross-framework—thus, it will apply to both the SI and the MI, as perhaps 'inevitability'—and conclude that since the physical facts are inevitable, then the agency facts are also inevitable. We can thus restate the consequence argument as follows, with the following terminological stipulations:

'P' represents the state of the entire universe at time t, stated in terms of the SI

'A' represents an action performed by an agent at time t, stated in terms of the MI

1. $\Box ((P_O \& L) \supset P)$	assumption (truth of determinism)
2. $N((P_O \& L) \supset P)$	1; Rule Alpha
3. $N(P_O \& L)$	Premise

[17] An anonymous referee for *Philosophical Studies* suggested something like the following objection.

4. Therefore, N(P) 2, 3; Rule Beta

5. □ (P ⊃ A) Supervenience of MI on SI

6. N(P ⊃ A) 5, Rule Alpha

7. Therefore, N(A) 4, 6; Rule Beta

Sellars will reject premise 6; and he has a principled reason for doing so. Premise 6 purports to deploy a modal operator ('inevitable') across separate frameworks. But Sellars will deny that one can meaningfully do so. Modal operators are only meaningfully defined *within* causal-explanatory frameworks, and cannot usefully be applied *across* frameworks. To do so is to be guilty of equivocation.

(Sellars will also deny premise 5. This might seem deeply implausible; it seems to represent a case of cross-framework modality that Sellars cannot deny without rejecting supervenience. I will return to this premise in section 5.1 below.)

Perhaps some analogies will help explain Sellars's point of view. Consider the following two arguments:

A1) 1. Pluto (the dwarf planet) is small.
2. King Kong is smaller than Pluto.
C. Therefore, King Kong is small.

A2) 1. You do not know you aren't a brain in a vat (in skeptical contexts).
2. If you do not know that you aren't a brain in a vat (in skeptical contexts), then you do not know that you are sitting at a desk right now (in an ordinary context).
C. Therefore, you do not know that you are sitting at a desk right now (in an ordinary context).

The first argument is clearly invalid, because it attempts to use the word 'small' in a context-invariant way. 'Small' only has meaning relative to a particular context, and it is not clear that it has *any* meaning that is *entirely* context invariant—that is, that can be used unrestrictedly across different contexts and frameworks. Contextualists have argued, too, that the second argument is guilty of a similar equivocation: Sentences involving 'knows' have

different truth values across contexts, and there simply is no use of the word 'knows' that can apply across both skeptical and non-skeptical contexts.

This is the argument we should understand Sellars as making with regard to modal vocabulary. Sellars spends a lot of time defining various modal operators (e.g., 'CAN', 'ABLE') from within the framework of agency. On the surface, his goal is to demonstrate that all of these can be defined such that agency is compatible with determinism (as defined in the SI). But on a deeper level, I believe his point is that various modal operators *can only be defined from within a particular framework*—e.g., the framework of agency, or the framework of scientific explanation.[18] To insist that we can define a modal term (such as 'inevitable') in a way that would make premise 6 true—that is, in a way that would allow it to take an SI proposition as its antecedent and an MI proposition as its consequent—is analogous to claiming that there *must be* a context-invariant meaning of 'small' (or 'knows'). And Sellars is free simply to deny that this is possible.

That this is Sellars's view is made most clear in a revised (and unpublished) version of FD. There, after a lengthy discussion of such practical modal operators such as 'CAN,' 'ABLE,' and 'PREVENT'—and the all-important question of whether scientific determinism has a bearing on the applicability of these operators—Sellars writes,

> It is only if we avoid confusing the specifically *practical* concepts of 'circumstances of action' and 'being prevented from doing something,' with the more generic and scientifically oriented concepts of 'circumstances' (i.e. relevant state of the universe) and 'physical impossibility,' that the temptation to suppose determinism to imply that the antecedent state of the universe 'makes it impossible for an agent to do anything other than he did,' (i.e. prevents him from doing it) can itself be avoided. (FD-revised, p. 26)

[18] Sellars is explicit about this toward the end of FD, where he writes (for example) that "our concept of *ability to do*…applies only to doings in the conduct sense" (FD p. 173). He goes on in the following pages to define a broader sense of 'ability' that applies not only to actions, but also to volitions (which for Sellars are *acts*, not *actions*); and he writes in the concluding paragraphs, "It is with reference to 'real' circumstances that abilities and hindrances are defined" (FD, p. 174). Sellars's larger point is, I hope, clear: Modal terms (like 'able') are defined relative to a particular framework, such as the framework of actions (or acts).

It is clear that for Sellars, certain concepts (such as 'circumstances of action' and the associated modal and counterfactual concepts such as 'able', 'prevented', etc.) are "specifically *practical*"; and must not therefore be confused with analogous scientific concepts (including scientific modal concepts). The implication of this is also clear—*antecedents couched in the language of the SI may not appear within the scope of agency-related modal predicates* (e.g., 'CAN' or 'possible', as related to actions). The scope of modal operators is limited according to framework.

This allows us to circle back and re-state our criticism of the original consequence argument, as laid out in section 1 above. The only way the original consequence argument can be sound is if premise 2 [$N((P_O \,\&\, L) \supset P)$] has as its antecedent a conjunction of SI claims, and an MI claim as its consequent. And for the reasons just explained, no meaning can be assigned to 'N' when premise 2 is formulated in this way. Thus, Sellars will argue that the consequence argument is guilty of equivocation, and premise 2 is the one where the equivocation takes place (where a supposedly 'context free' sense of N is smuggled in which moves us from SI modality to MI modality). (Premise 1 of the original consequence argument is ill-formed for similar reasons.)

I will consider one final objection against the Sellarsian account. It might seem as though on Sellars's view, determinism is really true, but we simply ignore it—or act like it isn't true—when we wish to talk about human action. Thus, we merely pretend to be free, even though we concede in principle that the consequence argument (and other related arguments) demonstrate that our actions are necessitated.[19]

But this worry overlooks the subtlety of Sellars's account. The MI framework really is the only framework within which action is intelligible. To try to explain action by appealing to SI antecedents is either (a) to change the subject—that is, not to talk about a state with

[19] Peter Olen raised this concern about Sellars's account.

intentional content at all, not to be talking about *action* any more at all—or (b) to exit the MI framework of rational intelligibility in favor of some other kind of project. I explained Sellars's case for this in section 2 ("The Manifest Image and Intentional Explanation"). While it might be possible to explain this or that action purely by reference to causal antecedents (rather than offering a rational explanation for the action), the category 'action' is constituted in part by the fact that it is conceptually connected to volitions, intentions, and to the calculative order that comprises the realm of intentions. (Again, it is no coincidence that Sellars has a long-ish section in FD—which might otherwise seem like a pointless digression—on the relations between intention, volition, and action.) To attempt to offer, systematically, a purely causal/scientific account of the causal antecedents of action is not to make your account of action more sophisticated—it is *to replace your account of action with some other account altogether* (e.g., the observable behavior of featherless bipeds). But such an account would not be an account of *actions*—the behavior, so described, would not be rationally intelligible as such, and the categories and concepts of this account would in no way mesh or match with the categories and concepts of our current theory of intentional explanation.

Thus, Sellars's strong claim is that we are only talking about action if we are using the MI framework of intentional explanation, within which such action is made rationally intelligible, using the framework of intentions and volitions—in short, the framework of reasons as causes. But further—and crucially—we don't merely *act* as though determinism is false with respect to action—*it really is false*. Action is only intelligible from within the MI; and determinism is not true from within the MI. Indeterminism isn't a pretend doctrine we adopt from within the MI. Rather, as Sellars puts the point, "'free-will', thus defined, belongs *neither* to the manifest image *nor* to the scientific image, for it is a higher order principle concerning the limits of the derivability of events from prior events. If there is a connection of free-will, thus defined, with the scientific image, it lies in the *fact* that only the confusions of

the vulgar determinist could lend plausibility to the idea that the conceptual resources of the (first level) manifest image are rich enough to generate (even in principle) a universal derivability of events, including human actions, from antecedent events" (RD, p. 182).

With this in mind, let us take a final look at the principle of determinism, as presented by Sellars:

PD-I $\quad A(x, t) \rightarrow a(x, t') \bullet N[a(x,t') \rightarrow A(x,t)]$

My translation of this was: If A does x at t, then x participated in a state of the universe, *a*, at t′, and it is necessary that (given x's participation in *a* at t′, that x does A at t).

Recall that the point of intentional explanation is to make human behavior rationally intelligible. As I explained in section 2, this limits the domain of allowable antecedents in such explanations. Thus, one may not just substitute anything for '*a*'. The only allowable substitutes are those that contribute to placing A within a rational, calculative order—that answer *a certain sort* of 'Why?' question, as it were. Thus, Sellars presents the incompatibilist with a dilemma: If we substitute for '*a*' states of affairs as describable within the MI, then PD-I is false; for the MI doesn't have the resources to support such claims of causal necessitation. But we cannot substitute for '*a*' states of affairs that are sufficiently fine-grained to make PD-I true, because then we have violated the restriction on permissible substitution—we would not be offering *action explanations* anymore, because we would not be appealing to real circumstances in light of which action is rationally intelligible. We would have abandoned the MI framework of rational explanation in favor of a different kind of explanation. We would be offering a scientific explanation of a set of physical behaviors by a featherless biped. But we wouldn't be explaining (say) *why Smith robbed a bank*. In principle, the latter type of explanation is unreachable from the SI; such explanations lie entirely within the domain of the MI, where the principle of determinism is factually false.

To conclude this section: I have been calling Sellars a compatibilist throughout. And in a sense, he is. But his view is more complex than this, in a way that defies easy labelling. Determinism is true, but not within the framework of intentional explanation. Within *that* framework, Sellars is an indeterminist. So is Sellars a compatibilist or an incompatibilist? Sellars often rejected what he saw as the false dichotomies of the received tradition: foundationalism vs. coherentism, rationalism vs. empiricism, and so on. And here again, we see Sellars trying to find a third way, one that reconciles determinism and indeterminism, compatibilism and incompatibilism. His repeated ability to find—and articulate—this third way is a testament to the depth of his philosophical insight.

5. Contemporary Versions of the Argument

I have argued that Sellars's version of compatibilism (as I will continue to call it, notwithstanding the considerations of the previous paragraph) offers a novel and powerful way of responding to the consequence argument. Unfortunately, his view received little uptake in the literature. FD is—even by the standards of Sellars articles—extremely difficult, and even seasoned Sellars scholars have often found it off-putting. This is a shame, because I think it represented a possibility to advance the free will debate, a possibility that was missed due to the customary neglect of Sellars's difficult work—neglect that is only recently being remedied.

Sellars's view—or something very much like it—has been independently advanced quite recently, by Christian List.[20] List—seemingly without knowledge of Sellars's earlier work—has developed a sophisticated response to the consequence argument that is remarkably similar to Sellars's earlier suggestion. List's argument consists of two key claims:

1) The consequence argument illicitly mixes descriptions from two separate 'levels'—the level of scientific accounts of the world, and the level at which we offer intentional explanations of agents' action.
2) At the level of intentional explanation, determinism is false.

[20] There is also some similarity between Sellars's view and that advanced by Kenny (1978), esp. Chapter 2.

Let us begin with the first claim. List argues that "the argument involves a category mistake: it conflates two different levels of description, namely, the physical level at which we describe the world from the perspective of fundamental physics and the agential level at which we describe agents and their actions" (List 2019, p. 253). Elaborating on this claim, List argues, "It is doubtful whether 'mixed' propositions such as Np_O, Nl, N($l \rightarrow p$), and N($p_O \rightarrow (l \rightarrow p)$) are well-formed at all, because N and p are agential-level expressions, while p_O and l are physical-level ones. It is especially doubtful whether it makes sense to put fundamental physical-level propositions such as p_O and l within the scope of an agential-level operator such as N. The operator N applies to agential-level propositions, not to fundamental physical-level ones" (List 2019, p. 258). This is essentially a sophisticated version of Sellars's argument—in explaining agent's actions, we cannot appeal to circumstances that are couched in the language of physics. This would be to appeal to Sellarsian 'pseudocircumstances' rather than to genuine circumstances of action. Only agential-level circumstances (and laws, if such there be) can be substituted into explanations of agents' actions.

Now to the second claim: List argues that determinism is false at the level of intentional explanation, and this is true because of multiple realizability. His argument is essentially the one I represented above and attributed to Sellars: The same intentional states can be represented by multiple different sets of physical states. Thus, the fact that a person is in intentional state S_1 at time t underdetermines what physical state she is in, and consequently underdetermines the physical state she will be in at t + Δt—and also, therefore, underdetermines the *intentional* state she will be in at t + Δt.

Various objections have been raised against List's project, but I believe that our examination of Sellars's project indicates how to respond to them. Recent authors have attempted to refute List by constructing the sort of cross-framework entailments that would prove that it is impossible (at the level of agency) for an individual ever to do otherwise than

what she actually does.[21] As we saw above, List doubts "whether 'mixed' propositions…are well-formed at all" (2019, p. 258); both he and Sellars have a principled reason for rejecting the possibility of such statements. On the final page of FD, Sellars makes some (again, regrettably obscure) comments about the laws governing human behavior; and concludes by saying that, "It is with reference to 'real' circumstances that abilities and hindrances are defined" (FD, p. 174). What is Sellars's point here? Laws are, for Sellars, metalinguistic statements which express inference licenses—say, that one claim may (or must) be inferred from another. The content of concepts is constituted by such law-like inferential proprieties (see "Concepts as Involving Laws and Inconceivable Without Them"). Inherent in the notion of a law is that it has modal force: It covers not only actual instances of B following from A (say, x's being a metal upon x's being copper), but a range of counterfactual instances as well. (Brandom has elaborated on this element of Sellars's view; see Brandom 2015, chapter 1, particularly his discussion of the "Kant-Sellars thesis about modality".) Thus, Sellars's argument in the penultimate paragraph of FD should be understood as *an argument about the allowable construction of laws within the SI and MI frameworks, and therefore as about the scope of various modal operators*. A law with an action as a consequent cannot take a 'pseudo-circumstance' as its antecedent because this would require cross-framework necessity; and (as I have argued) modal notions cannot be defined across frameworks in this way. Any attempt to construct a cross-framework argument will eventually have to rely on a premise formulated along the lines of premise 6 from our reconfigured consequence argument from section 4 above; and as I said there, Sellars (and List) have a principled reason for rejecting the coherence of such a premise.

To conclude this section: We can see Sellars's argument as concerning the *specific content* of the conclusion of the consequence argument. He insists that the *actual* conclusion

[21] See, for example, Birch (manuscript); Gebharter (2020).

will always differ from the *desired* conclusion, for the incompatibilist.[22] Sellars warns us toward the end of FD that we should not confuse

> it is *not possible that* x at *t* willed to do A′ [*because,* according to the principle of determinism, etc.]

with

> x *was not able* at *t* to will to do A′. (FD, p. 172)

Thus, when arguing from SI antecedents, the conclusion of the consequence argument is in fact the former, whereas the desired conclusion is the latter; since the latter, deploying action-appropriate modal vocabulary (such as 'able') and mentioning an *action* (A′)—all of which belong to the MI—cannot be the consequent of an argument with SI antecedents. But (Sellars must argue) if the former claim is the conclusion of the consequence argument, then the argument is quickly mired in incoherence. For even though the first formulation deploys SI-modality ("it is *not possible that*"), it attempts to smuggle in MI action-vocabulary (A′, which again names an action). Such content cannot appropriately be placed within the scope of an SI modal operator; the SI knows neither agents, nor actions, nor intentionality *per se*. Such an argument could properly describe an array of behavior by a featherless biped—but it couldn't make a statement about an *action*. This way of framing the argument—as an argument about the *content* of the consequence argument, and how its advocate must choose between falsity and incoherence—is perhaps a uniquely Sellarsian take on the issue, and one that distinguishes Sellars's approach from List's.

5.1. Supervenience

One final hurdle remains. The revised consequence argument presented in section 4 relied on the thesis of supervenience, which generates the (seemingly uncontroversial) premise:

[22] This way of framing Sellars's argument was suggested by an anonymous referee for *Philosophical Studies*.

5. □ (P ⊃ A)

Yet this premise seems to rely on cross-framework modality. Isn't this a counterexample to the Sellars/List claim that cross-framework modal claims are illegitimate? And indeed, criticism has focused on this aspect of List's account; Jonathan Birch, for example, writes that to deny the legitimacy of such claims, List "would be forced to conclude that any assertion of the supervenience of one level on another is also a category mistake. But the existence of such supervenience relations is a foundational assumption of List's approach" (manuscript, pp. 8-9).

List seems to hold that *no* kinds of modality apply in a cross-framework sort of way. What of supervenience claims, then? List writes,

> To respond to this objection, we must begin by noting a key desideratum that the mixed-level language would have to fulfil in order to express the consequence argument successfully. The language would have to enable us, not merely to *talk about* the argument from some external ('meta-linguistic') perspective, but to *assert* the argument— that is, to *use* its constituent propositions, in an 'object-language' way, not just to offer external commentary on them. Arguably, when we engage in supervenience talk, we often adopt an external perspective, for instance by stepping outside any particular level of description and then talking about how what is true at one level relates to what is true at another. (List 2015, p. 268).

Arguably, then, when we make claims such as premise 5 from the revised consequence argument, we are not as it were making a statement within the framework of the MI (or within the SI, or within both). Rather, we are making a meta-claim about the relation between the two frameworks.

It seems clear that Sellars must himself follow this type of response. Throughout FD, Sellars entertains principles (such as the principle of determinism) which seem to employ cross-framework modality; but (as I argued at the end of section 3.1) it is clear that he ultimately must reject the legitimacy of such statements, involving as they do an attempt to formulate conditionals with SI antecedents and MI consequents. Modal claims that have agency-propositions in the consequent position must have MI-framework statements in the antecedent

position. Thus, while supervenience might be true as a general statement of the *relationship* between two frameworks, it does not follow that this allows us to formulate *modal claims* with SI antecedents and MI consequents. As an imperfect analogy, while there is a sense in which biology supervenes upon particle physics, it doesn't follow that one could formulate laws with antecedents stated in the language of particle physics, and consequents stated in the language of biology. Sellars will continue to insist that entailments of any sort whose antecedents are formulated in the language of the SI cannot have consequents formulated in agency-language; in embedding behavior within a scientific entailment, such behavior is not denoted as *action* at all.[23]

The incompatibilist might reply that in premise (5):

$\Box (P \supset A)$

'$\supset$' can be read as the symbol for material implication.[24] The requirements for material implication to be well-formed are quite relaxed; premise (5) can in fact be well-formed, as long as both 'P' and 'A' are statements concerning the same world, *w*. Further, the conditional will be true if 'P' and 'A' are both true at *w*. (Or, of course, if the antecedent is false at *w*.) Thus (it is argued), premise (5) does represent a well-formed conditional employing cross-framework modality—so long as the conditional is interpreted as material implication.

Sellars will reject the above argument. Sellars, of course, famously argues that formally valid deduction—the kind embodied in material implication relations—is parasitic on, and merely expresses the normative proprieties implicit in, content-dependent materially valid

[23] It is also worth noting that token identity (which Sellars endorses) will not by itself secure the transmission of modal statuses across frameworks. Let us use the following notation:

m = a mental state, specified in the language of the MI
b = a brain state, specified in the language of the SI
M = a proposition asserting that m obtains
B = a proposition asserting that b obtains
P = a proposition representing the state of the entire universe at time t, stated in terms of the SI

Even if m = b (i.e., the two states are token identical), it is *not* the case that $\Box (P \supset B)$ entails $\Box (P \supset M)$.

[24] The following argument was suggested by Anjana Jacob.

inferences we deploy in our ordinary practices. But more to the point, the implication relation embodied in supervenience cannot be mere material implication, because supervenience must be an explanatory relation. If O′ supervenes upon O, it cannot just *happen to be the case that:*

$$\Box [O_{1\ldots n} \supset O'_1].$$

To offer an illustrative example, if we interpret '⊃' as material implication, the following is true:

$$\Box [(2 + 2 = 5) \supset \text{Grass is green}]$$

even though—in fact, largely because—the antecedent and consequent are entirely unrelated to each other. The supervenience relation *cannot be like that*. The array of facts denoted in the subvening base must somehow explain the fact denoted at the supervening level. The antecedent must somehow *explain* the consequent. But if the relation must be explanatory, then Sellars will again insist that it is part of the logical grammar of action-vocabulary that actions belong to an autonomous explanatory framework, and we cannot explain action by appeal to SI states of affairs. Thus, we have a principled reason for endorsing List's conclusion: From the fact that supervenience is true as a relation between frameworks, it doesn't follow that we can formulate necessary entailments involving statements from the two different frameworks in the antecedent and consequent positions.

5.2. Sellars and the Normativity of the Framework of Persons

However, I think that Sellars's rejection of the supervenience argument might take a more radical form. On a deeper level, the real issue is that Sellars does not intend for the MI to be an explanatory/descriptive framework in the first place. This may seem like an odd claim, given the focus on section 2 on intentional explanation. However, we can make sense of this. Note Sellars's language at the end of "Philosophy and the Scientific Image of Man", where he writes, "To recognize a featherless biped or dolphin or Martian as a person requires that one

think thoughts of the form, 'We (one) shall do (or abstain from doing) actions of kind A in circumstances of kind C'. To think thoughts of this kind is not to *classify* or *explain*, but to *rehearse an intention*" (PSIM, p. 40).

An analogy here might be useful. Sellars is, of course, an expressivist about moral discourse. If one sincerely utters a claim such as, "Physician-assisted suicide is immoral," or "Women have the right to control their reproductive lives," one is not *describing* anything. One is *expressing* a special kind of intention. Without necessarily committing ourselves specifically to Sellars's version of expressivism, we can note that the expressivist would have us move away from descriptivist accounts of moral discourse (and therefore away from the need to locate moral facts that must therefore supervene on anything in the first place).

Similarly, in ascribing a belief or desire to someone, one is not in the first instance describing that person. One is placing that individual in the space of reasons.[25] Obviously, the proper utterance of a moral claim—just like the proper attribution of a belief or a desire—will depend upon certain (vaguely specified) factual circumstances obtaining. But this is true for any speech act or mental state. As I have emphasized, the space of reasons is a subset of the space of causes; having a belief (or a desire, or a moral commitment, etc.) will, *ceteris paribus*, have various causal consequences. Nevertheless, for Sellars, the question of whether various expressive speech acts supervene on the natural should sound just as odd as the question of whether "Shut the door!" supervenes on the natural.[26] Surely, the latter has appropriate assertion conditions; surely, this does not entail that we can or should usefully talk about supervenience with respect to it.

The existence of these proper attribution conditions explain why we can use expressive language in rational explanations. As I indicated in section 2, the functional roles defining

[25] Brandom offers a superb articulation of this view in (Brandom 1979).
[26] I owe this way of putting the issue to Mark Lance, who has made this point to me a number of times over the years.

mental states are for Sellars defined normatively and not computationally (something that distinguishes Sellars from traditional functionalists in the philosophy of mind). However, Sellars is also committed to what James O'Shea calls Sellars's 'norm/nature meta-principle': "Espousal of principles is reflected in uniformities of performance" ("Truth and Correspondence," p. 216; cited in O'Shea 2007, p. 138). Thus, 'belief,' 'desire,' and other mental states are normatively defined; but proper attribution of a mental state to someone requires that they in fact instantiate in their behavior a certain number of these various norms; and this is what enables folk-psychological terms to play a role in intentional explanation. It does not follow from this that intentional vocabulary is primarily descriptive.

I will conclude this section by tying together these two Sellarsian rebuttals to the supervenience argument. Michael P. Wolf and I have argued[27] that for the Sellarsian, supervenience should not be seen as an ontological thesis that is a consequence of naturalism. Rather, it should be seen as a consequence of the demand for rational consistency, a consequence of the normativity of reason. If two worlds are indiscernible at the non-normative level, it would be odd if different normative or expressive claims were justified in these two worlds. And this is because our consideration of other possible worlds is guided by various normative commitments. For present purposes, the most central among these is the demand for rational consistency and parity of reasons: Roughly speaking, when we judge, we should judge in the same way where we have the same reasons. Conversely, if we judge differently in two different cases, there should be some reason justifying our different choice. If we stipulate that two worlds are identical in all of their non-normative properties, that gives us all of the same empirical facts to consider in making our expressive judgments; and so a difference in judgments would require different empirical facts or else it would become arbitrary and irrational. However, this is a normative commitment guiding us in how to make our expressive

[27] Wolf and Koons (2016), pp. 92-3.

judgments, not a discovery of some natural facts that constitute facts about the normative (or what have you). It is this normative requirement of consistency that pushes toward accepting supervenience, rather than supervenience itself being an ontological or metaphysical thesis that helps explain the nature of normativity.

However—and this is where we tie together the two Sellarsian arguments—whatever facts we appeal to in order to justify our expressive judgment must be framework-appropriate. Understanding supervenience in the way just outlined makes it clear why statements of the form

$$\Box (P \supset A)$$

are illegitimate. They are illegitimate because supervenience is not an ontological or metaphysical thesis, but rather a rational requirement on *justificatory* or *explanatory* claims. And if the above entailment is understood as a template for *explanatory* claims, then '⊃' *cannot* be read merely as the symbol for material implication. Thus, the above the modal operator is restricted in the ways Sellars claims; the entailment cannot be a cross-framework one.

Thus, we can *either* think of supervenience as a relation between two frameworks; but then it does not allow the formulation of modal claims between propositions in these two frameworks. *Or* we can understand supervenience as a requirement imposed by the norm of rational consistency, in which case it allows us to assert inferential explanatory or justificatory relations between descriptive and expressive facts, but only intra-framework facts. Thus, we can explain agency facts, but only by appeal to MI facts.[28]

[28] There is a sense in which SI facts can legitimately be appealed to in rational explanations, e.g.: "Why is Pierre celebrating? Because the collision in the particle accelerator released a particle with the mass of 125GeV/c^2." This is a claim from within the SI, but it is not being used as a deterministic pseudocircumstance to explain Pierre's action. Rather, it is being situated within the MI framework of agency—we must also take it as read that Pierre *knows* of these results, *knows* that they confirm the existence of the Higgs boson, *understands* the importance of this result, etc., etc. What is disallowed, for Sellars, is a modal statement in which a pure SI proposition (with no background premises, enablers, etc., from the MI) entails an agency claim from the MI. That would rather be like an 'is' entailing an 'ought.'

This second feature of Sellars's account of intentional language—that it is primarily expressive rather than descriptive—represents one way in which Sellars's account diverges from List's. It may also represent an advantage of Sellars's account, in that (as I argued) demands to satisfy supervenience requirements are more pressing for descriptivist accounts; with respect to other kind of speech act, demands for an account of how they supervene on the natural seem out of place. Of course, one can jettison this expressivist element of the Sellarsian project and still retain the basic insight recognized by Sellars and List, but I would urge that it nevertheless gives Sellars additional argumentative resources for responding to the incompatibilist.

6. Conclusion

Allen Wood famously wrote, "When we consider all Kant's views together, it is tempting to say that he wants to show not only the compatibility of freedom and determinism, but also the compatibility of compatibilism and incompatibilism" (Wood 1984, p. 74). *Qua* noumenal creatures, we must think of ourselves as subject only to necessitation of reason; although as phenomenal creatures, we are subject to physical laws. Sellars takes this formula, and precisely inverts it: Insofar as we belong to the SI (which is Sellars's epistemological take on the noumenal realm), we are genuinely subject to physical determination. But insofar as we are rational animals, belonging to the MI, we are not subject to determinism. So Sellars has sought not merely "to show… the compatibility of compatibilism and incompatibilism"; he has sought to show the compatibility of determinism and indeterminism. I argue that he has made a plausible case.

Space has prevented me from discussing every element of Sellars's account. For example, I only discussed his treatment of the negative condition on free action, and not the positive condition. I also gave only the most cursory treatment of List's contemporary version

of the argument. I encourage readers to look at List's splendid argument; it is impossible for me to do justice to it in the space available. It develops, in rich detail, the essential insight recognized by Sellars—that determinism is false in the MI, and that you cannot explain intentional behavior (which belongs to the MI framework) by appeal to descriptions belonging to the SI framework (the framework where determinism belongs). It is only unfortunate that an adequate development had to wait so long, due to the neglect of Sellars's original articulation of this point. This is merely another example of philosophical insight being delayed for decades by neglect of important historical figures; consider, for example, Mark Schroeder's (2008) independent development of what was essentially Sellars's expressive logic—again, decades after Sellars's work.[29] Perhaps closer attention to the work of great figures like Sellars would hasten our philosophical progress.[30]

[29] I make this comparison in more detail in Koons (2019, Chapter 5).

[30] I am grateful to Pete Olen for providing me with helpful comments on an early draft of this essay. I also had helpful conversations with Anjana Jacob and Mark Lance concerning some of the issues discussed here. Finally, two anonymous referees for *Philosophical Studies* provided a good deal of feedback that lead to the substantial improvement of the essay.

Works Cited

Birch, Jonathan. Manuscript. "Free Will and the Cross-Level Consequence Argument." <http://philsci-archive.pitt.edu/18413/>. Accessed 14 April 2021. Cited by permission of author.

Brandom, Robert B. 1979. "Freedom and Constraint by Norms." *American Philosophical Quarterly* 16, no. 3 (July): 187-96.

— 2015. *From Empiricism to Expressivism: Brandom Reads Sellars* (Cambridge, MA: Harvard University Press).

Gebharter, Alexander 2020. "Free Will as a Higher-Level Phenomenon." *Thought* 9, pp. 177-187.

Ginet, Carl 1966. "Might We Have No Choice?" in Keith Lehrer (ed.), *Freedom and Determinism* (New York, Random House): 87-104.

Kenny, Anthony 1978. *Freewill and Responsibility* (London and New York: Routledge and Kegan Paul).

Koons, Jeremy Randel forthcoming. ""To show the compatibility of compatibilism and incompatibilism': Sellars's reinvention of Kant's conception of free will," in Luz C. Seiberth and Mahdi Ranaee (eds.), *Reading Kant With Sellars*.

— 2019. *The Ethics of Wilfrid Sellars* (New York: Routledge).

List, Christian 2019. "What's Wrong with the Consequence Argument," *Proceedings of the Aristotelian Society* 119:3 (October): 253-274.

McKay, Thomas J. and David Johnson. "A Reconsideration of an Argument Against Compatibilism," *Philosophical Topics* 24(2): 113-122.

O'Shea, James R. 2007. *Wilfrid Sellars: Naturalism with a Normative Turn.* Cambridge, UK: Polity Press.

Pereboom, Derk 2001. *Living Without Free Will* (Cambridge: Cambridge University Press).

Schroeder, Mark 2008. *Being For: Evaluating the Semantic Program of Expressivism*. Oxford: Oxford University Press.

Sellars, Wilfrid 1980. "Concepts as Involving Laws and Inconceivable Without Them," in Wilfrid Sellars, *Pure Pragmatics and Possible Worlds: The Early Essays of Wilfrid Sellars*, edited by Jeffrey Sicha Atascadero, CA: Ridgeview Publishing Company.

— 1962 (TC). "Truth and 'Correspondence'." *The Journal of Philosophy* 59, no. 2 (January 18): 29-56.

— 1963a (PSIM). "Philosophy and the Scientific Image of Man," in Wilfrid Sellars, *Science, Perception and Reality* (Atascadero, California: Ridgeview Publishing Company): 1-40.

— 1963b (EPM). "Empiricism and the Philosophy of Mind," in Wilfrid Sellars, *Science, Perception and Reality* (Atascadero, California: Ridgeview Publishing Company): 127-196

— 1963c (SRLG). "Some Reflections on Language Games" in Wilfrid Sellars, *Science, Perception and Reality* (Atascadero, California: Ridgeview Publishing Company): 321-358.

— 1966 (FD). "Fatalism and Determinism," in Keith Lehrer (ed.), *Freedom and Determinism* (New York, Random House): 141-174.

— ND. (FD-revised). "Fatalism and Determinism." Revised unpublished version of Sellars 1966. <https://digital.library.pitt.edu/islandora/object/pitt:31735062218957>. Accessed 13 May 2021.

— 1973 (AAE). "Actions and Events," *Noûs* 7:2 (May): 179-202.

— 1974 (MP). “Metaphysics and the Concept of a Person,” in Wilfrid Sellars, *Essays in Philosophy and Its History* (Dordrecht: D. Reidel Publishing Company): pp. 214-241.
— 1975 (RD). “Reply to Alan Donagan,” *Philosophical Studies* 27: 149-184.
— 1976 (VR). “Volitions Re-Affirmed.” In Myles Brand and Douglas Walton (eds.), *Action Theory* (Dordrecht: D. Reidel Publishing Company): 47-66.
van Inwagen, Peter 1975. “The Incompatibility of Free Will and Determinism,” *Philosophical Studies* 27:3 (March): 285-199.
— 1978. “Ability and Responsibility,” *The Philosophical Review* 87:2 (April), pp. 201-224.
— 1983. *An Essay on Free Will*. Oxford: Clarendon Press.
— 1989. “When is the Will Free?” *Philosophical Perspectives Volume 3: Philosophy of Mind and Action Theory*: 399-422.
Wiseman, Rachael 2016. *Routledge Philosophy Guidebook to Anscombe’s Intention*. New York: Routledge.
Wolf, Michael P. and Jeremy Koons. 2016. *The Normative and the Natural*. London: Palgrave Macmillan.
Wood, Allen W. (ed) 1984. *Self and Nature in Kant’s Philosophy*. Ithaca, New York: Cornell University Press.