Global Supervenience Without Reducibility*

Stephan Leuenberger

Many debates in philosophy are concerned with how properties in a given class—mental properties, say, or ethical properties—are related to certain other properties, which are supposed to be metaphysically less puzzling—such as the physical, the descriptive, or the naturalistic properties. Specifically, it is often asked whether the former are reducible to the latter.

In such debates, the concept of supervenience has played a central role. A recent paper starts by observing that "[w]hen it comes to evaluating reductive hypotheses in metaphysics, supervenience arguments are the tools of the trade."[1] As is familiar, supervenience is modal co-variation: it rules out differences in one respect without differences in another respect. The theses that mental properties and ethical properties supervene on physical properties are widely accepted. But there is disagreement about what they imply. On one side, so-called "non-reductive physicalists" in the philosophy of mind endorse a supervenience thesis to distinguish themselves from dualists, while also opposing reductionism. As Jaegwon Kim put it, "many philosophers saw in [supervenience] the promise of a new type of dependency relation that seemed just right, neither too strong nor too weak, allowing us to navigate between reductionism and outright dualism."[2] Likewise, metaethicists who repudiate reductive naturalism typically still accept that the ethical supervenes on the natural. On the other side, a number of philosophers—notably Kim himself in the mental and Frank Jackson in the ethical case— have argued that such an intermediate position is untenable, since supervenience entails reduction.[3]

Who is right? Before answering the question, we need to recall a distinction

*Thanks for helpful feedback to audiences in Glasgow, Durham, Oslo, Hamburg, and Edinburgh, in particular Michael Clark, Cian Dorr, Peter Fritz, Bryan Pickel, Stefan Roski, Jeff Russell, Gabriel Uzquiano, Alan Weir, Nathan Wildman, and Timothy Williamson; and also to a number of anonymous referees. My research on this paper was prompted by conversations with Campbell Brown—special thanks to him, as well as to Bruno Whittle, whose detailed comments and suggestions greatly improved the paper. This work was supported by the Templeton Foundation, via the Durham Emergence Project, and by the Arts and Humanities Research Council [grant number AH/M009610/1].

[1] Johannes Schmitt and Mark Schroeder, "Supervenience Arguments under Relaxed Assumptions," *Philosphical Studies*, CLV, 133–162.

[2] Jaegwon Kim, "Supervenience as a Philosophical Concept", in *Supervenience and Mind: Selected Philosophical Essays* (Cambridge: Cambridge University Press, 1993), pp. 131–60, at p. 147.

[3] Kim, "Supervenience as a Philosophical Concept," *op. cit.*; Frank Jackson, *From Metaphysics to Ethics: A Defence of Conceptual Analysis*, (Oxford: Oxford University Press, 1998).

between two kinds of supervenience claims: local and global. The former concern co-variation among properties of possible individuals, the latter co-variation among property distributions of possible worlds. In this paper, my interest is mainly in global supervenience claims. The result that certain local supervenience claims entail reduction would be significant only to the extent that the claims themselves are plausible. However, in many areas where reductionism is debated, local supervenience theses look untenable. Broadly physical properties arguably do not locally supervene on microphysical ones, because of the phenomenon of quantum entanglement. Furthermore, externalism about mental content and mental states gives us reason to deny the local supervenience of the mental on broadly physical properties. Likewise, many ethical theories seem to imply that ethical properties need not be locally supervenient on ordinary descriptive ones. The supervenience claims that are widely accepted in these domains are global ones.[4]

I shall argue that global supervenience does not entail reductionism—that what I call a "reduction principle" linking them is false. In a nutshell, my argument is this: from the metaphysics of possible worlds, we learn that there is no upper bound on the size of possible worlds; and from the model theory of infinitary logics, we learn that if there is no such upper bound, then some globally supervening properties are not reducible.

Roadmap: After presenting some technical background material in section 1, I briefly survey the debate about whether global supervenience entails reductionism, and present what I call the "Jackson–Stalnaker argument" for a positive answer (section 2). I then identify a hidden premise in the argument, which I argue should be rejected (section 3). In section 4, I sketch a counterexample to the reduction principle, drawing on the model theory of infinitary logic. Before concluding, I discuss a number of responses (sections 5 and 6).

1 Reduction and Definability

The term "reduction" has no settled meaning in philosophy. In philosophy of science, reduction is often taken to be a relation between theories. My concern here is with reduction among properties and relations instead. According to a standard usage of the philosophy of mind, a mental property is reducible to physical properties just in case it is identical with a physical property. Generalizing that usage, we could take the claim that a class of properties A is reducible to a class of properties B to be simply the claim that every property in A is also in B—or equivalently, that A is a sub-class of B.[5]

While this use of "reduction" is partly stipulative, it is well-motivated. It is widely accepted that A's being a sub-class of B is sufficient for reduction.

[4] A popular gambit to safeguard local supervenience claims from such apparent counterexamples is to include extrinsic properties in the base. I will discuss this in sections 2 and 7.

[5] I will typically use 'properties' in a broad sense, applying to include relations as well. Context will make it clear when the term is to apply to monadic properties only.

Our question would then become: does the global supervenience of A on B entail that A is a sub-class of B? If properties are individuated very finely, the answer to that question is bound to be negative. Being an equilateral triangle modally co-varies with being an equiangular triangle, so supervenience does not entail reduction if these are distinct properties. For the purposes of this paper, I shall make the assumption that no two properties can be co-intensional. That is, if F and G are instantiated by exactly the same things in all possible worlds, then $F = G$. Without that assumption, the reduction principle that I am going to criticize is a non-starter.

But even granted a coarse-grained individuation of properties, there seems to be a simple argument to show that global supervenience does not entail reductionism. Let F and G be such that it is possible for something to have one but not the other, and suppose that H is the disjunction of F and G, in the sense that in all possible worlds, an individual has H iff it has F or G. Obviously, $\{H\}$ supervenes on $\{F, G\}$—on any reasonable way of defining supervenience—even though H is not a member of $\{F, G\}$.

Defenders of the link between supervenience and reduction can respond by denying that membership is necessary for reduction. Intuitively, the disjunctive H should count as reducible to F and G. There is a more general strategy available, which does not merely deal with the case of disjunction: to say that global supervenience implies *definability*, which in turn implies reducibility.[6]

We know what it is for a predicate to be definable from other predicates. This gives rise to a corresponding relation of definability among the properties expressed by these predicates. Thus for a given class of putative base properties we can consider the properties definable from that class. Obviously, H as described above is definable from F and G, using disjunction.

The claim that definability implies reducibility is plausible and well-motivated, at least if B is one of the usual suspects, such as the the physical, descriptive, or naturalistic properties. Kim put the point well when he argued that putative base properties do not need to satisfy what he calls the "resemblance criterion"—that sharing of a property must ensure resemblance in some respect:

> [I]t seems that we allow, and ought to allow, freedom to combine and recombine the basic theoretical predicates and functors by the usual logical and mathematical operations available in the underlying language, without checking each step with something like the resemblance criterion; that would work havoc with free and creative scientific theorizing.[7]

The idea is clear: when a physicist exercises her freedom and chooses to define properties from a given stock of physical properties, she does not, at any point in the process, stop doing physics. Likewise, nobody ceases to do science by defining further properties in terms of naturalistic properties.

[6] An alternative response is to restrict the reduction principle: for any B closed under definability, if A globally supervenes on B, then A reduces to B.

[7] Kim, "Supervenience as a Philosophical Concept," *op.cit*, at p. 153.

Kim's claim needs to be qualified, however. The property of being a unit set of an electron is definable from the language of physics using mathematical operations, but it is hardly a physical property, nor reducible to physical properties. The difference between logical and mathematical operations matters in this context. The relevant properties are those definable from B using such operations as negation, disjunction, conjunction, and quantification, as well as the identity predicate. This observation shows that we need to make the concept of definability more precise, before we can properly assess whether it can serve as a fulcrum between global supervenience and reduction.

Definition 1. *Given a language L whose vocabulary includes all members of B, an n-ary relational predicate F is* L-definable *from predicates in class B iff there is an L-formula Φ, with no non-logical vocabulary except perhaps members of B, such that the following is true in all possible worlds w:*

$$\forall x_1 \ldots \forall x_n (F(x_1, \ldots, x_n) \leftrightarrow \Phi(x_1, \ldots, x_n))$$

Definability thus understood is relative to a language, here taken to be determined not just by a class of sentences, but also by a division of the vocabulary into logical and non-logical terms.

Definition 1 is in broad agreement with definitions of explicit definability of predicates given in logic textbooks, except that it quantifies over possible worlds rather than over models.[8] The purpose of these modifications is to ensure that the theses of supervenience and of definability are stated in a common idiom. If one thesis was about possible worlds and the other about models, nothing interesting would follow from them, and in particular, no reductionist conclusions would follow.

For the same reason, we need the notion of definability to apply to classes of relations as well as predicates. I shall here assume that any formula Φ of L expresses a relation, which is instantiated by exactly those things that satisfy the formula.[9] Given the assumption that no two properties are co-intensional, there is one and only one property expressed by every predicate. In light of this, no harm will be done by letting a predicate symbol F do double duty as a name of the corresponding property, and by applying the term "definable" not just to predicates, but also to the relations they express.

The fact that definability is relative to a language prompts the question of which language is relevant, if we are interested in reduction. Those who claim

[8] For a representative example, see George S. Boolos, John P. Burgess, and Richard C. Jeffrey, *Computability and Logic*, 4th edition (Cambridge: Cambridge University Press, 2002), at p. 266.

For a discussion of logical issues concerning the relationship between non-global (local) supervenience and definability, see Lloyd Humberstone, "Note on Supervenience and Definability", *Notre Dame Journal of Formal Logic*, XXXIX, 2: 243–252.

Philosophers have considered more demanding notions of definition. See Gideon Rosen, "Real Definition," *Analytic Philosophy* LVI: 189–209.

[9] This amounts to a so-called "abstraction principle" for relations. For languages that do not allow terms for such abstracts—such as $\mathcal{L}$, to be introduced shortly—this principle will be predicative and not susceptible to Russell's paradox and its cognates.

that supervenience entails reductionism have urged that we should be quite liberal. What we are interested in is reducibility in principle, not reducibility given some contingent limitations. In particular, we should allow the background logical language to have infinitary resources.

To be able to discuss this move, I shall now introduce an infinitary language $\mathcal{L}$, which will play a major role in the following.[10] $\mathcal{L}$ differs from the standard language of first-order logic in two ways. First, instead of combining with a single variable, as in '$\exists x$' or '$\forall x$', an existential or universal quantifier may be attached to any expression standing for a set of variables, as in '$\exists X$' or '$\forall X$'. Just like in the familiar case of first-order logic, truth-conditions for formulas in which they occur are stated relative to a variable assignment, or valuation. For a world w, a *w-valuation* is any function that maps variables to things that exist in w.[11] A w-valuation g' is an *X-variant* of a w-valuation g iff it differs at most on X, that is, if $g'(y) = g(y)$ for every variable y that is not a member of X. Then for every world w:

- $\forall Xp$ is true relative to w-valuation g iff for all X-variants g' of g, p is true relative to g'.

Existential quantification is defined in terms of universal quantification and negation in the usual way.

Second, $\mathcal{L}$ has two new symbols $\bigvee$ and $\bigwedge$ instead of $\vee$ and $\wedge$, which may be attached to any expression standing for a set of formulas, to form respectively the disjunction and conjunction of its members. The truth-conditions for conjunction are (again, disjunction is defined in the usual way):

- $\bigwedge S$ is true relative to g iff every member of S is true relative to g.

For readability, I shall sometimes use the more familiar symbols, writing $\forall x$ instead of $\forall\{x\}$, and $p \wedge q$ instad of $\bigwedge\{p, q\}$, for example.

In $\mathcal{L}$, $\bigvee$ and $\bigwedge$ apply to sets of formulas of any cardinality, and $\exists$ and $\forall$ to sets of variables of any cardinality.

With definability thus specified, the two theses that jointly entail the reduction principle can be stated more precisely:

DefRed If A is $\mathcal{L}$-definable from B, then A reduces to B.

SupDef If A globally supervenes on B, then A is $\mathcal{L}$-definable from B.

I shall grant DefRed to the proponent of the reduction principle. But in subsequent sections, I shall challenge SupDef.

[10] My presentation of infinitary logic aims for accessibility to readers familiar with first-order prediate logic. Certain details that are not relevant for the philosophical point I wish to make will be omitted.

[11] Such valuations will be "class functions." For convenience, I will pretend that there are such functions. A more rigorous discussion would work with valuations that are defined only "locally", for the variables in formulas of a relevant set.

It is worth observing that the language $\mathcal{L}$ has more expressive power than any of the other standard infinitary languages.[12] Hence if a property is not definable in $\mathcal{L}$, it is not definable in any of these other languages either. The decision to focus on $\mathcal{L}$, and to assume that putative base properties are closed under $\mathcal{L}$-definability, is thus friendly to the proponent of the reduction principle.

2 The Jackson–Stalnaker Argument

Some philosophers have affirmed what is effectively the reduction principle without argument. Thus David Lewis:

> [I]f supervenience fails, then no scheme of translation can be correct and we needn't go on Chisholming away in search of one. If supervenience succeeds, on the other hand, then some correct scheme must exist; the remaining question is whether there exists a correct scheme that is less than infinitely complex.[13]

Lewis did not specify what kind of supervenience he was talking about. Other philosophers recognized that the answer to the question may be sensitive to what variety of supervenience is at issue, and that at any rate an argument is required. Kim established that local supervenience entails reducibility. More specifically, he proved this for the variety of local supervenience known as "strong supervenience", which holds between A and B iff for all individuals x in w and x' in w', if x in w and x' in w' have the same B-properties, then they have the same A-properties.[14]

While Kim's result does not speak directly to the question of this paper, it plays a pivotal role in the pertinent arguments. For this reason, it is worth stating here (it is proved in appendix A). In fact, the version here is more general than Kim's, because it applies to relations as well as monadic properties.[15]

Proposition 1. *For all sets of relations A and B, if A strongly supervenes on B, then A is $\mathcal{L}$-definable from B.*

[12] In the literature (for example M. A. Dickmann, *Large Infinitary Languages. Model Theory* (Amsterdam: North Holland Publishing Company, 1975), what I call $\mathcal{L}$ is usually called $\mathcal{L}_{\infty\infty}$.

[13] David Lewis, "New Work for a Theory of Universals," in *Papers in Metaphysics and Epistemology* (Cambridge: Cambridge University Press, 1999), pp. 8–55, at pp. 29–30.

Lewis speaks of 'translation' rather than 'reduction', but this makes little difference. A translation would map a predicate to one that is co-intensional, and thus expresses the same property, given our assumption that properties are coarse-grained. Translation suffices for reduction.

[14] Kim, "Supervenience as a Philosophical Concept," *op. cit.*.

[15] The generalisation of the usual definition of strong supervenience to cover relations, which I relegate to the appendix, is taken from Stephan Leuenberger, "Supervenience in Metaphysics," *Philosophy Compass*, III/IV, pp. 749–762. In the context of the present paper, it does not matter whether that definition provides an adequate explication of strong supervenience for relations. (I propose a different explication, which I take to be superior, in Stephan Leuenberger, "Supervenience among Classes of Relations", in Miguel Hoeltje, Benjamin Schnieder, and Alex Steinberg, eds., *Varieties of Dependence* (Munich: Philosophia, 2013), pp. 325–346.) The concept only plays an auxiliary role in the argument to be discussed here, and does not appear in either the premises or the conclusion.

Kim considered the question whether global supervenience entails reduction on a number of occasions.[16] In one place, he wrote that "it seems a plausible conjecture that if extrinsic properties are included in ... [the subvenient set]—in particular, if, along with the usual Boolean operations, identity and quantification are allowed for property composition—... the equivalence will obtain."[17]

Kim's conjecture was supported by Frank Jackson and Robert Stalnaker.[18] Even though their arguments were advanced independently, the key idea is the same, such that it is appropriate to speak of the "Jackson–Stalnaker argument". While Stalnaker's version is an important point of reference in the specialized literature on supervenience, Jackson's has been more widely influential, especially in contemporary metaethics. The version I present here is closer to Stalnaker's, which provides more formal detail.[19] It is more general than either Stalnaker's or Jackson's by applying to relations of any (finite) adicity, rather than just to monadic properties.[20]

The key idea behind the argument is the one adumbrated by Kim in the conjecture quoted above: to move from a claim of global supervenience to one of local supervenience—specifically, strong supervenience—on a base expanded to include extrinsic properties, and then appeal to Proposition 1 to infer a definability claim. Given a putative base B, the associated class that includes the relevant extrinsic properties—e.g. *being such that something is* F, where F belongs to B—can be taken to be the class of all properties that are $\mathcal{L}$-definable from B.[21] I shall call this class $\mathcal{L}(B)$. It is the so-called "closure" of B under $\mathcal{L}$-definability.

Suppose that we hold that the property of thinking of gold fails to strongly supervene on microphysical properties: someone might duplicate me with respect to microphysical properties, but live in a possible world where there is schmold rather than gold, and hence fail to share my thought about gold. Then the strategy is to find extrinsic microphysical properties that I have and my duplicate lacks. In our example, it is easy to come up with examples of such extrinsic properties, like *being such that something has a nucleus with 79 protons*. But in other cases, finding examples might be harder, whence the need

[16] Kim was influenced by the insightful discussion of the reduction principle in Bradford Petrie, "Global Supervenience and Reduction," *Philosophy and Phenomenological Research*, XLVIII: 119–30.

[17] Jaegwon Kim, "Postscripts on Supervenience," in Kim, *Supervenience and Mind*, *op. cit*, at p. 170.

[18] Jackson, *op. cit*, Robert C. Stalnaker, "Varieties of Supervenience," *Philosophical Perspectives* X: 221–241.

[19] For technical objections to Jackson's specific version of the argument, see Timothy Williamson, "Ethics, Supervenience and Ramsey Sentences," *Philosophy and Phenomenological Research*, LXII: 625–630. Like Stalnaker's, the version presented here is not vulnerable to these objections.

[20] Mark Moyer, "Weak and Global Supervenience are Strong," *Philosophical Studies*, CXXXVIII: 125–30, at pp. 148–49, also indicates how Stalnaker's argument can be generalized to relations.

In many philosophical contexts, the putative global supervenience bases that are relevant contain relations, such as spatiotemporal distance. See also footnote 36.

[21] This is no substantive claim about extrinsicality. The word "extrinsic" plays a purely heuristic role in my presentation of the argument.

for a general licence to move from a global supervenience claim to a strong supervenience claim with extrinsic base.

A proof that we have such a general licence requires a precise explication of global supervenience. I shall here adopt Stalnaker's, which is called "strong global supervenience" in the literature, and which I will simply call "global supervenience" here.[22] ("Strong global supervenience" is not to be confused with "strong supervenience", which is a species of local supervenience.)

Global supervenience is defined in terms of certain mappings between the domains of possible worlds. Say that a function f from the domain $D(w)$ of world w to the domain $D(v)$ of world v is a *domain-isomorphism* iff it is one-one and onto; and say that it preserves A iff for all $R \in A$ and $x_1, \ldots, x_n \in D(w)$, $\langle x_1, \ldots, x_n \rangle$ instantiates R in w iff $\langle f(x_1), \ldots, f(x_n) \rangle$ instantiates R in v. An A-isomophism is a domain-isomorphism that preserves A. Finally:

Definition 2. *A class of relations A globally supervenes on a class of relations B iff every B-isomorphism is also an A-isomorphism.*[23]

The following theorem, proved in Appendix A, is a generalized version of a result that Stalnaker proved for the case of monadic properties:

Proposition 2. *For all classes of relations A and B, if A globally supervenes on B, then A strongly supervenes on $\mathcal{L}(B)$.*

Armed with these two propositions, we can argue for SupDef as follows. Suppose that A globally supervenes on B. By Proposition 2, A strongly supervenes on $\mathcal{L}(B)$. By Proposition 1, A is $\mathcal{L}$-definable from $\mathcal{L}(B)$. But $\mathcal{L}(B)$ is itself closed under $\mathcal{L}$-definability: what is $\mathcal{L}$-definable from it is already in it.[24] It follows that A is in $\mathcal{L}(B)$, that is, that A is $\mathcal{L}$-definable from B.

It appears that SupDef has been proved! Given that I have granted DefRed, the reduction principle seems to be vindicated.

Here is how Stalnaker describes the result himself:

> [A] global supervenience thesis is in fact quite strong: Kim has shown that if A strongly supervenes on B, then every A-property is necessarily equivalent to a property definable in terms of B-properties.

[22] With the main rival explications of global supervenience, the result to be proved, Proposition 2, does not hold. Following Williamson, *op. cit*, let a property be *uniform* iff in every possible world, it applies to everything or to nothing, and let B be the class of all uniform properties. It is easily seen that $\mathcal{L}(B) = B$, and that for every class of properties A, A stands in both the relations of intermediate global supervenience and weak global supervenience to B. (See Karen Bennett and Brian McLaughlin, "Supervenience," in Edward N. Zalta, ed., *Stanford Encyclopedia of Philosophy*, (Spring 2018 Edition), URL = ¡https://plato.stanford.edu/archives/spr2018/entries/supervenience/¿. for these varieties of global supervenience.) However, the class of physical properties, like many other classes, obviously fails to strongly supervene on $\mathcal{L}(B)$.

[23] If A is a unit set $\{F\}$, I shall also say that F globally supervenes on B to mean that $\{F\}$ does.

[24] The idea behind this is that replacing a term occurring in a definition with its own definition produces another definition. Proving this from Definition 1 is straightforward but tedious.

So our result implies that it is also true that if A *globally* supervenes on B, then every A-property is necessarily equivalent to a property definable in terms of B-properties. If necessary equivalence is enough for property identity, then we can say that if A globally supervenes on B, the A-properties all *are* properties definable in terms of B-properties.[25]

3 Are world-sizes bounded?

Prima facie, Stalnaker's version of the Jackson–Stalnaker argument is watertight. On closer inspection, however, it relies on a hidden extra premise: that the class $\mathcal{L}(B)$ is a *set*. Without that premise, Proposition 1, which generalizes over sets, does not apply.

How are we to decide whether that premise is justified? If the class B is not a set—if it is a proper class—then since $B \subseteq \mathcal{L}(B)$, $\mathcal{L}(B)$ is not a set either. The interesting question is this: under what conditions is $\mathcal{L}(B)$ a set given that B is? The following result, proved in Appendix A, gives an answer:

Proposition 3. *For any set of relations B, $\mathcal{L}(B)$ is a set iff there is a cardinality κ such that for every possible world w, the cardinality of the domain of w is less than κ.*

We might have thought that the question whether there is such an upper bound on the sizes of worlds would be of little interest beyond specialized debates in the metaphysics of possible worlds. But if it had an affirmative answer, then the gap in the Jackson–Stalnaker argument could be closed, in light of Proposition 3. So the question turns out to be highly relevant for the fate of non-reductive physicalism and analogous views in metaethics and other areas.

Recherché or not, it seems clear how the question should be answered. Drawing on work by Daniel Nolan, I shall offer two considerations against the idea that world sizes are bounded.

First, a plausible theory of possible worlds includes a *principle of recombination*. This is a principle telling us what worlds there are in modal space. Intuitively, it states that if you cut, copy, and paste portions of possible worlds, the result is again a possible world. As we all know from our text processors, you can paste a given item multiple times. In one canonical formulation, the principle reads as follows: "for any objects in any worlds, there exists a world that contains any number of duplicates of those objects."[26] In this version, it immediately entails that there is no upper bound on the size of worlds. For any such putative bound λ, there will be a number $\kappa > \lambda$, and, by the principle, a world containing κ intrinsic duplicates of David Lewis. It then follows that λ is

[25] Stalnaker, "Varieties of Supervenience", *op. cit.*, at p. 228.

[26] Daniel Nolan, "Recombination Unbound," *Philosophical Studies*, LXXXIV: 239–62, at p. 239.

not an upper bound on the size of worlds. Modifying the principle to limit the number of duplicates would seem to be *ad hoc*.[27]

Second, consider a statement of the form "there are fewer than κ individuals". For some κ, this will be true in all possible worlds, if there is an upper bound on world sizes. Given the link between necessity and possible worlds, "there are fewer than κ individuals" will be necessary. However, such a statement would hardly be a priori true. In this post-Kripkean age, we have become used to some claims being necessary a posteriori. But the generally accepted examples of such claims have certain features in common: most obviously, they are stated using proper names or natural kind terms. But "there are fewer than κ individuals" does not have these features. It would be a brute necessity of a non-Kripkean kind.[28] Those of us who share Leibniz' view that "there is always a presumption on the side of possibility, that is, everything is held to be possible unless it is proven to be impossible"[29] will be reluctant to accept it.[30]

4 Global supervenience without $\mathcal{L}$-definability

We have seen that the argument for the reduction principle based on Kim's and Stalnaker's result is not sound unless there is an upper bound on the size of possible worlds. Moreover, I have argued on metaphysical grounds that there is no such upper bound. Of course, it does not follow that the reduction principle cannot be established, or is even false.[31]

In this section, I shall sketch an example of a relation S that globally supervenes on a set of relations B, but is not $\mathcal{L}$-definable from them. This is a

[27] In *On the Plurality of Worlds* (Oxford: Blackwell, 1986), at p. 103, David Lewis claimed he needed such a restriction to avoid a *reductio ad absurdum* of the principle of recombination. However, Nolan, "Recombination Unbound," *op. cit*, showed that the alleged *reductio* was fallacious. In David Lewis, "Tensing the Copula," *Mind*, CXI: 1–13, at p. 8, Lewis acknowleged that "Nolan ... has made a fairly persuasive case that there are more possibilia than I used to think, in fact proper-class many."

[28] When Kripke considered the issue himself, he wrote "it seems to me to be reasonable to suppose ... that for every cardinality κ it is possible that there are exactly κ individuals" (Saul Kripke, "A Puzzle about Time and Thought," in *Philosophical Troubles. Collected Papers, volume I*, pp. 373–379, at 378.

[29] G. W. Leibniz, *Philosophical Essays*, Roger Ariew and Daniel Garber, eds. (Indianapolis: Hackett Publishing, 1989), at p. 238.

[30] The Jackson–Stalnaker argument, and in particular Proposition 2, rests on another assumption that I shall not challenge here: that the domain of every possible world is set-sized. Given that the quantifiers of $\mathcal{L}$ range over everything in the world, the assumption can only hold if sets are not in the relevant sense "in the world". I grant that assumption to the proponent of the reduction principle because it is standardly made in the debate about global supervenience.

[31] Stalnaker himself briefly considered the question whether his result could be established in a different way. Responding to Michael Glanzberg, "Supervenience and Infinitary Logic," *Noûs*, XXXV: 419–439, he wrote that "Michael Glanzberg has shown that a more sophisticated argument establishes that a language that has infinite Boolean combinations but only finite strings of quantifiers will suffice." ("Postscript added in 2002", in *Ways a World Might Be. Metaphysical and Anti-Metaphysical Essays* (Oxford: Oxford University Press, 2003), pp. 105–108, at p. 105.) However, Glanzberg's argument only applies if no possible world has an uncountable domain, and hence does not avoid the objection of the last section.

counterexample to the reduction principle, granted the following bridge principle:

RedDef If A reduces to B, then A is $\mathcal{L}$-definable from B.

For the time being, I shall take this principle—whose converse DefRed was introduced in section 1—for granted. I will consider challenges to it in section 6.

To understand why SupDef fails, it is helpful to consider a simpler example first. Suppose that our non-logical predicates are F and K, meaning "is a fork on the table" and "is a knife on the table", and we wish to express in the language of first-order predicate logic with identity that there as many knifes and forks on the table. In this language, we can say very specific things that entail that the knifes and forks are equinumerous—that there are 17 knifes and 17 forks, say. We can also say some slightly less specific things that still entail it—that either there are 17 knifes and 17 forks, or 18 knifes and 18 forks, say. It is not tempting to form the disjunction of all these claims, for every natural number. But since there are infinitely many natural numbers, and since the language only allows us to form finite disjunctions, this attempt expressing the equinumerosity claim fails.

In the example to be presented, it will also be tempting to form a disjunction of the claims in a certain class—but since the class is not a set, no such disjunction can be formed. So the obvious way of providing a definition fails. Showing that no non-obvious way can succeed requires considerable work. The counterexample to SupDef to be presented is an adaptation to our metaphysical context of Jerome Malitz' example to the effect that $\mathcal{L}$ does not have what is called the "Beth definability property."[32]

Let B consist of the monadic properties F and G and binary relations $<^F$ and $<^G$.[33] A *B-world* is one that satisfies the following two conditions:[34]

- Everything is an F or a G, but not both.
- If $x <^F y$, then Fx and Fy, and if $x <^G y$, then Gx and Gy.

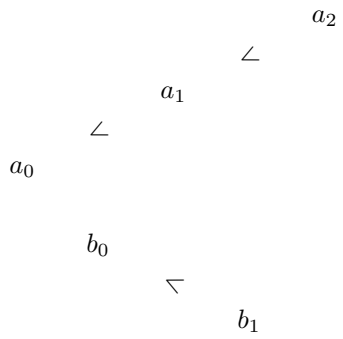
Thus a B-world is one that is in a certain sense compartmentalized by the members of B: its domain is partitioned by the monadic properties in B, and its relations never relate things from different cells of that partition.

The diagram below represents a B-world. The Fs are represented in the top half of the drawing, diagonally upwards. The Gs are represented in the bottom half, diagonally downwards.

[32] Jerome Malitz, "Infinitary Analogs of Theorems of First Order Model Theory," *The Journal of Symbolic Logic*, XXXVI: 216–228.

[33] I shall use the letters A, B, F, G, etc sometimes as names of the properties or sets of properties involved in my counterexample to the reduction principle, and sometimes schematically. Context will disambiguate.

[34] I tacitly assume that worlds have non-empty domains.



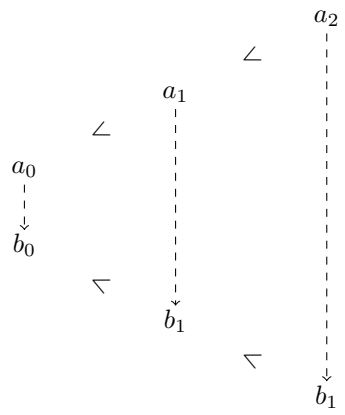
Further, suppose that S is a binary relation that is instantiated in all and only B-worlds such that $<^F$ and $<^G$ form well-orderings that are isomorphic to each other—in all and only B_S-worlds, for short. More formally, say that s is an *FG-isomorphism* in w iff w is a B-world, and s is a one-one mapping from the set of Fs onto the set of Gs in w such that $x <^F y$ iff $s(x) <^G s(y)$. Then we can capture the distinctive principles governing the distribution of S in a possible world w as follows:

(1) For all w, if there is an FG-isomorphism in w, then S is instantiated in w.

(2) For all w, if S is instantiated in w, then there is an FG-isomorphism s such that $S(x, y)$ iff $s(x) = y$.

From a standard set-theoretic result, it follows that if there is an isomorphism between the well-ordered sets of the Fs and the Gs, it is unique. Given a B_S-world w, I can thus write $S_w(x)$ for the s whose existence is guaranteed by (2).

In the world depicted above, the well-orderings of the Fs is not isomorphic to the well-ordering of the Gs, since there are three Fs and only two Gs. But in the world below, these well-orderings are isomorphic. This world is hence a B_S-world. A dashed arrow from a_i to b_i represents that $S_w(a_i) = b_i$.

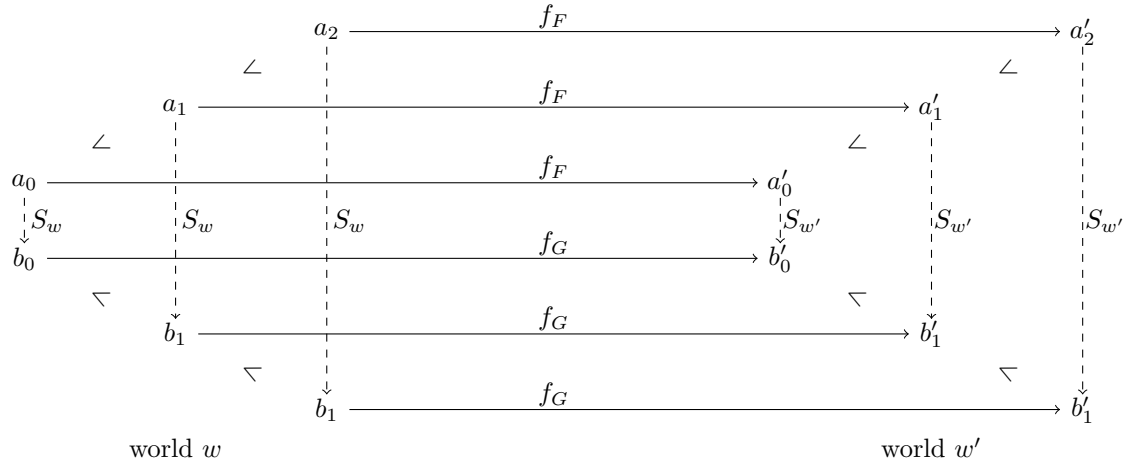


The principles (1) and (2) are purely structural. They do not tell us much about what S and the members of B are like. A plethora of properties and relations with very different intrinsic natures will satisfy (1) and (2). To fix ideas, imagine that all members of the domain of w are particles, and that they fall into two kinds: F-particles and G-particles. If x and y are F-particles such that $x <^F y$, then x comes into existence earlier than y; and likewise for G-particles. Since the Fs and the Gs are well-ordered, there is a first F-particle, and a first G-particle. We can think of S as a relation of simultaneity: S relates x to y if and only if x is an F-particle and y a G-particle that come into existence at the same time. Furthermore, (1) and (2) may be taken as expressing laws of natures holding in virtue of the natures of F-particles and G-particles: if the Fs and the Gs are isomorphic, then there are distinct but simultaneous particle births, given (1); otherwise there are not, given (2).

Recall that I am constructing a counterexample to the reduction principle: a case of a globally supervening but irreducible property. We are now in a position to show that S has the first of the two requisite features.

Proposition 4. *If (1) and (2) are true, then S globally supervenes on B.*

This is shown by verifying that for all worlds w and w' and all B-isomorphisms f between them, $S(x, y)$ holds in w iff $S(fx, fy)$ holds in w'. For the left-to-right direction, suppose that $S(x, y)$ holds in w. Then S_w is an FG-isomorphism in w. Let f_F be the restriction of f to the Fs in w, and f_G its restriction to the Gs in w. Then it is easy to verify—the picture below should help—that $f_G S_w f_F^{-1}$ is an isomorphism between $<^F$ in w' and $<^G$ in w'.



By (1), S is instantiated in w'. By (2), $S_{w'}$ is an isomorphism from the set of Fs and the set of Gs in w'. Since there is only one such isomorphism, as noted earlier, $S_{w'} = f_G S_w f_F^{-1}$. It follows that $S_{w'}(fx) = f_G S_w f_F^{-1}(fx) = f_G S_w(x) = f_G(y) = fy$, and hence $S(fx, fy)$ in w'. The right-to-left direction is shown in the same way, *mutatis mutandis*. Since f was chosen arbitrarily, we can conclude that S strongly globally supervenes on B.

While our assumptions (1) and (2) let us derive that S globally supervenes on B, they do not guarantee the other feature of our desired counterexample, namely, indefinability. It is easy to see that (1) and (2) are compatible with there being no B_S-worlds, in which case S would be $\mathcal{L}$-definable (e.g. by $\neg(x = x)$) from any class. To show that S is not $\mathcal{L}$-definable from B, we need some principles of a broadly recombinatorial character, giving us many different B_S-worlds:

(3) For every order type, there is a B-world where $<^F$ has that order type.[35]

(4) For every order type, there is a B-world where $<^G$ has that order type.

(5) For any B-worlds w and w', there is a B-world w'' such that $<^F$ in w'' is isomorphic to $<^F$ in w, and $<^G$ in w'' is isomorphic to $<^G$ in w'.

These assumptions allow us to prove the indefinability claim:

Proposition 5. *If (1)–(5) are true, then S is not $\mathcal{L}$-definable from B.*

The main result of this paper is now immediate:

Corollary 1. *If (1)–(5) are true, then global supervenience does not entail $\mathcal{L}$-definability.*

[35] Order types are abstractions from well-orderings under the relation of order isomorphism—two well-orderings have the same order type iff they are isomorphic.

The rest of this section is devoted to sketching a proof of Proposition 5. Details of the proof will be filled in Appendix B. Sections 5–7 of the paper do not presuppose this technical material.

To see why Proposition 5 holds, recall what it would take for S to be in $\mathcal{L}(B)$, according to Definition 1: that there is a B-formula Φ of $\mathcal{L}$ —i.e. a formula with no non-logical vocabulary except perhaps predicates from B—such that the following is true in all possible worlds:

$$\forall x \forall y (Sxy \leftrightarrow \Phi(x, y))$$

Adapting Malitz' strategy, I shall show that there is no formula $\Phi(x, y)$ satisfying this condition.

Let $\Phi(x, y)$ be any such candidate definiens. I shall show that there are worlds w_Φ and w_Φ^- (pronounced "w-phi-minus") such that:

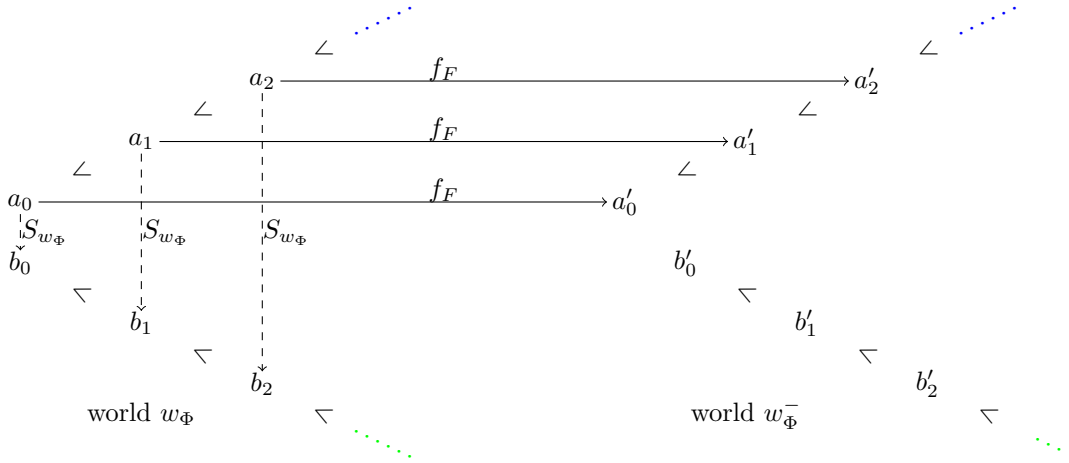
(i) $\exists x \exists y Sxy$ is true in w_Φ.

(ii) $\exists x \exists y Sxy$ is not true in w_Φ^-.

(iii) $\exists x \exists y \Phi(x, y)$ is true in w_Φ iff $\exists x \exists y \Phi(x, y)$ is true in w_Φ^-.

From this, it follows that $\forall x \forall y (Sxy \leftrightarrow \Phi(x, y))$ is false in w_Φ or in w_Φ^-. In either case, it is not true in all possible worlds, and Φ does not define S. Since Φ was chosen arbitrarily, we can conclude that S is not $\mathcal{L}$-definable from B.

We can picture the worlds w_Φ and w_Φ^- as follows: the Fs of w_Φ are isomorphic with the Gs of w_Φ and the Fs of w_Φ^-, but not with the Gs of w_Φ^-.



In proving that given a candidate definition ϕ, there are indeed such worlds w_Φ and w_Φ^-, the key move is to consider sentences in which all quantifiers are restricted to the Fs, and sentences in which all quantifiers are restricted to the Gs. We can choose a world w_Φ that is sufficiently large that the truth-values of

such restricted quantifications determine the truth-value of $\exists x \exists y \Phi(x, y)$ (Theorem 2, below). Roughly speaking: since B contains no relations that link Fs with Gs, giving separate B-descriptions of the Fs and the Gs is enough to give a B-description of w_Φ. Moreover, by a version of the Löwenheim–Skolem Theorem (Theorem 1), the relevant sentences in which quantification is restricted to the Gs fail to distinguish w_Φ from world w_Φ^-, whose Gs are non-isomorphic to those in w_Φ.

To make this reasoning rigorous, we need to be able to keep track of various sub-languages of $\mathcal{L}(B)$. For this purpose, we shall use a piece of standard terminology for infinitary languages: for cardinals κ and λ, $\mathcal{L}_{\kappa\lambda}$ denotes a language just like $\mathcal{L}$, except that $\bigvee$ and $\bigwedge$ can only attach to sets of formulas of cardinality less than κ, and quantifiers can only attach to sets of variables of cardinality less than λ. If $\kappa \leq \kappa'$ and $\lambda \leq \lambda'$, then every formula of $\mathcal{L}_{\kappa\lambda}$ also belongs to $\mathcal{L}_{\kappa'\lambda'}$. Our $\mathcal{L}$ is the union of the languages $\mathcal{L}_{\kappa\lambda}$, for all κ and λ. Hence the candidate definiens Φ is in $\mathcal{L}_{\kappa\kappa}$, for some κ.

It will turn out to be useful for me to also introduce another, not standardly used measure of complexity of formulas. A candidate definiens ϕ has a certain *depth*—an ordinal indicating how far it embeds atomic formulas. Specifically, let the depth d be 0 for an atomic formula of $\mathcal{L}$, and set $d(\neg\psi) = d(\psi)$, $d(\bigwedge \Gamma) = sup\{d(\psi) : \psi \in \Gamma\} + 1$, and $d(\forall X \Gamma) = d(\Gamma) + 1$ (with $\bigvee$ and $\exists$ treated like $\bigwedge$ and $\forall$, respectively).

For a fixed infinite κ, define $c(\alpha)$ recursively as follows: $c(0) = 2^{2^\kappa}$; $c(\alpha+1) = 2^{2^{c(\alpha)}}$ for successor ordinals $\alpha + 1$; and $c(\alpha)$ as the sum of $\{c(\beta) : \beta < \alpha\}$ for limit ordinals. (This is a "superexponentiation" function.) The function c is extended to formulas of $\mathcal{L}_{\kappa\kappa}$, with $c(\phi) =_{df} c(d(\phi))$.

Suppose the candidate definiens ϕ is in $\mathcal{L}_{\kappa\kappa}$. Given an order type of cardinality $2^{2^{c(\phi)}}$, let w_Φ be a world—whose existence is guaranteed by (3), (4), and (5)—in which the Fs and the Gs both have that order type. Clearly, the Fs and the Gs in w_Φ are isomorphic, and using (1), we can infer that (i) is true.

For a predicate F and a formula ϕ of $\mathcal{L}$, define ϕ^F to be the same formula as ϕ except that all quantifiers are restricted to F. More formally: $\phi^F = \phi$ for atomic ϕ, $(\neg\phi)^F = \neg\phi^F$, $(\bigwedge_{i\in I} \phi_i)^F = \bigwedge_{i\in I} \phi^F$, and $(\forall X \phi)^F = \forall X(\bigwedge_{x\in X} Fx \rightarrow \phi^F)$.

We can now appeal to a version of the Downward Löwenheim–Skolem Theorem, which is proved in the appendix:

Theorem 1. *Let ν be any cardinal, and H a predicate. Suppose that the size of H_w is at least 2^ν, and that there are at most ν non-logical terms in $\mathcal{L}_{\nu\nu}$. Then there is $D \subset H_w$, of size 2^ν, such that if $H'_w = D$, then for all sentences ϕ of $\mathcal{L}_{\nu\nu}$, ϕ^H is true in w iff $\phi^{H'}$ is true in w.*

If we set $\nu = 2^{c(\phi)}$, Theorem 1 guarantees that there is a subset of G_{w_Φ}, of cardinality $2^{c(\phi)}$, which can be taken as the extension of the new predicate G' to make (iv) true:

(iv) For all sentences ϕ of $\mathcal{L}_{\nu\nu}$, ϕ^G is true in w_Φ iff $\phi^{G'}$ is true in w_Φ.

Since $\nu = c(\phi) \geq 2^{2^\kappa}$, and since the relation of being a well-ordering is expressible in $\mathcal{L}_{\omega_1\omega_1}$, the G's will also be well-ordered by $<^G$, and have an order

type. By (4) and (5), there is a world which can serve as our w_Φ^-: its Gs have the order type as the G's in w_Φ, and its Fs are isomorphic to the Fs in w_Φ. Since $2^{2^{c(\phi)}} \neq 2^{c(\phi)}$ by Cantor's Theorem, there is no FG-ismorphism in w_Φ^-, and by (2), S is not instantiated in w_Φ^-. Hence (ii) is true as well.

It remains to show that our choice of w_Φ and w_Φ^- makes (iii) true. Let a B-sentence be a B-formula that is a sentence—i.e. a sentence in which every non-logical term expresses a member of B. Then by a straightforward induction on the complexity of formulas, we can show:

(v) For all B-sentences ϕ of $\mathcal{L}_{\nu\nu}$, $\phi^{G'}$ is true in w_Φ iff ϕ^G is true in w_Φ^-.

(vi) For all B-sentences ϕ of $\mathcal{L}_{\nu\nu}$, ϕ^F is true in w_Φ iff ϕ^F is true in w_Φ^-.

From (iv) and (v), we get:

(vii) For all B-sentences ϕ of $\mathcal{L}_{\nu\nu}$, ϕ^G is true in w_Φ iff ϕ^G is true in w_Φ^-.

Theorem 2, below, will allow us to infer the desired (iii) from (vi) and (vii). In preparation for stating the theorem, I give a more general definition of the condition on the distribution of B that is satisfied by the B-worlds of our example.

Say that a relation R is *restricted to* ϕ in a class of worlds W iff $\forall x_1 \ldots x_n(Rx_1 \ldots x_n \rightarrow \bigwedge_{1\leq i\leq n} \phi(x_i))$ is true in $w \in W$.

Definition 3. *A class of relations B, with $F, G \in B$, is* compartmentalized *by F and G in W iff every member of B is either restricted to F in all $w \in W$, or restricted to G in all $w \in W$, and $\forall x(Fx \leftrightarrow \neg Gx)$ is true in all $w \in W$.*

In light of this definition, the B-worlds introduced earlier form a class of worlds in which B is compartmentalized by F and G.

The following theorem is proved in Appendix B:

Theorem 2. *Let B be compartmentalized by F and G in W, and ϕ be a B-sentence of $\mathcal{L}_{\kappa\kappa}$. Then if for all $w, w' \in W$ and B-sentences ψ of $\mathcal{L}_{\nu\kappa}$, where $\nu = c(\phi)$, ψ^F and ψ^G have the same truth-value in w as they have in w', so does ϕ.*

By construction, w_Φ and w_Φ^- are B-worlds, and as noted, B is compartmentalized by F and G in all B-worlds. It follows from (vi) and (vii) that for all B-sentences ψ of $\mathcal{L}_{\nu\nu}$—and *a fortiori* of $\mathcal{L}_{\nu\kappa}$, since $\kappa < \nu$—ψ^F and ψ^G have the same truth-value in w_Φ as they have in w_Φ^-. Theorem 2 then entails that ϕ has the same truth-value in w as it does in w'. This shows that (iii) is also true, and completes the proof of Proposition 5.[36]

[36] While S is a binary relation, the argument also shows that the monadic property expressed by $\exists y Sxy$ globally supervenes on B without being $\mathcal{L}$-definable from it.(See also footnote 20.)

5 Are there real counterexamples?

I have shown that if there are properties and relations satisfing all of (1)–(5), then there is an instance of global supervenience without $\mathcal{L}$-definability. But are (1)–(5) jointly satisfied? After all, my counterexample was quite fanciful, involving properties and relations not dreamt of in contemporary science.

My response to this turns on David Lewis' influential distinction between *sparse* properties—roughly, properties whose sharing makes for objective resemblance—and *abundant* properties. Perhaps we do not have a reason to think that there are sparse properties corresponding to S or the members of B. However, in a context in which supervenience and reduction are under discussion, we typically quantify over abundant as well as sparse properties. One of the chief uses of the notion of supervenience is exactly in articulating the relationship between sparse and abundant ones. It is true that in the last section, I pretended that F, G, $<^F$ and $<^G$ are sparse. But this was just to make the example vivid. Its target, the reduction principle, is supposed to apply to abundant properties just as much. The members of B may be gerrymandered and disjunctive. Once we realize this, there is no immediate reason to think that no actually instantiated properties satisfy (1)–(5).

An objector may try to limit the range of potential counterexamples: she might argue that properties whose reducibility we are interested in can only be instantiated in a certain restricted class of worlds, with bounded size. We can easily adapt the Jackson–Stalnaker argument to show that such properties will be $\mathcal{L}$-definable from a set B if they globally supervene on B.[37] But what would the restricted class of worlds be? Suppose that the actual laws of nature limit the cardinality of the world—they might say that everything is a fusion of spacetime points, say, and that there are continuum many such points. Then the suggestion would be that since the relevant properties are only instantiated in worlds where the actual laws hold, the reduction principle holds for them.

I am happy to grant that *if* we have reason to think that members of A are only instantiated in worlds with the same laws as ours, then A is $\mathcal{L}$-definable from every class upon which it globally supervenes. But I deny that many interesting classes A are of that kind. I submit that the mental and certainly the moral properties are not like that. Suppose we have reason to think that F is only instantiated in worlds with our laws. Then it is these reasons, and not merely the global supervenience claim, that support reductionism. It would be highly misleading to claim that the global supervenience claim by itself entails reductionism.

[37] If A globally supervenes on a set B, and no member of A is instantiated in a world with more than κ individuals, then there is a cardinal λ such that A strongly supervenes on the properties that are $\mathcal{L}_{\lambda\lambda}$-definable from B. These properties form a set, and Proposition 1 applies.

6 Definitions in a more powerful language?

My counterexample to the reduction principle relied on there being properties and relations that satisfy (1)–(5). In the last section, I considered challenges to that existence claim. Here, I consider responses that grant that claim, but deny that it refutes the reduction principle.

In theory, one could adopt a more demanding notion of global supervenience. However, this is not attractive: if anything, the formulation we used seems stronger, not weaker, than the informal notion it explicates. An initially more promising option is to reject the principle RedDef, introduced in section 4, and take a notion of definability that is less demanding than $\mathcal{L}$-definability as sufficient for reduction.

The reason why the Jackson–Stalnaker argument failed is that $\mathcal{L}$ cannot form the disjunction of proper class many disjoints. So we could consider a more powerful language which can. This takes us beyond standard infinitary logic into largely uncharted territory.

There is more than one way to implement such a proposal. The simplest is to change the formation rules for $\bigvee$ and $\bigwedge$ in such a way that they apply to every class to form a new formula. But then a version of Russell's paradox threatens: let ϕ be $\bigvee\{\psi : \psi$ is a formula of the language that does not contain itself as a disjunct$\}$, and ask whether ϕ contains itself as a disjunct.

A more cautious approach is to define the formulas of a language $\mathcal{L}'$ in such a way that a class can be "collected together" once, as it were, but not repeatedly:

- Every $\mathcal{L}$-formula is a formula.
- If S is a class of $\mathcal{L}$-formulas, then $\bigvee S$ and $\bigwedge S$ are formulas.
- If ϕ and ψ are formulas and x a variable, then $\neg\phi$, $\phi \wedge \psi$, $\phi \vee \psi$ and $\forall x\phi$ are formulas.

Now the above route to Russell's paradox is blocked, since the formation rules do not guarantee that $\bigvee\{\psi : \psi$ is a formula of $\mathcal{L}'$ that does not contain itself as a disjunct$\}$ is a formula of $\mathcal{L}'$. Inspecting the proofs of Propositions 1 and 2, we can verify that if A globally supervenes on B, then A is $\mathcal{L}'$-definable from B. But I would argue that the new required bridging premise is implausible:

DefRed* If A is $\mathcal{L}'$-definable from B, then A reduces to B.

This is because $\mathcal{L}'$-definabilty does not really deserve to be called "definability." To put it differently: in the context of a language like $\mathcal{L}'$, Definition 1, which says what it is for a predicate to be definable in a language, should be rejected. The reason is that definitions should be *eliminable*: the definiendum can be replaced by the definiens in any context, without loss of well-formedness or truth. But in $\mathcal{L}'$, this is not guaranteed. Suppose that for a proper class S of $\mathcal{L}$-formulas, $\forall x(Fx \leftrightarrow \bigvee S)$ is a $\mathcal{L}'$-sentence true in all possible worlds. If S' is a proper class of $\mathcal{L}$-formulas, then $\bigvee(S' \cup \{Fx\})$ is a $\mathcal{L}'$-formula, but

$\bigvee(S' \cup \{\bigvee S\})$ is not. Replacing F with its supposed definiens has resulted in the loss of well-formedness!

A third implementation would iterate the above move. Just as $\mathcal{L}'$ was defined from $\mathcal{L}$, it defines $\mathcal{L}''$ from $\mathcal{L}'$, and likewise for every finite number of superscripted primes. The language $\mathcal{L}^+$ is then the union of all these languages in the hierarchy.

This approach avoids paradox, and guarantees the eliminability of definientia. But it introduces a new hierarchy of proper classes. By a generalization of Cantor's argument, there will be more classes at every stage in the hierarchy. I would argue that on this picture, the cardinals do not exhaust the sizes—there are many proper class sizes above them. The considerations from section 3 then suggest that there should be possible worlds with these sizes. But if there are, it is no longer guaranteed that $\mathcal{L}^+$ provides a definition for every globally supervening property.

Thus I do not see how the strategy of allowing disjunctions and conjunctions of proper class size could be implemented in a way that saves the reduction principle.[38]

The infinitary language $\mathcal{L}$ is a first-order language, allowing quantification into name position only. A further response to the counterexample is to allow quantification into predicate position. In the language $\mathcal{L}^2$, the quantifiers then range over classes. It is routine to give a second-order definition of S in terms of B. For we can express that X is a bijection from the Fs to the Gs such that if Xxy and $Xx'y$, then $x <^F y$ iff $x' <^G y'$. Let this claim be abbreviated as $ISO_{F,G}(X)$. Then $\forall xy(Sxy \leftrightarrow \exists X(ISO_{F,G}(X) \wedge Xxy))$ is a $\mathcal{L}^2$-definition of S in terms of B.

Unlike with $\mathcal{L}^+$, I do not see how we could use $\mathcal{L}^2$ to give a general argument from global supervenience to definability. Be that as it may, I think the strategy fails anyway. Once again, the second premise is implausible:

DefRed** If A is $\mathcal{L}^2$-definable from B, then A reduces to B.

The problem here is that the proposed $\mathcal{L}^2$-definition is impredicative: S is defined by existentially quantifying over classes, including relational classes, and S itself is one of those. I do not think that there is anything wrong with impredicative definitions. But I would insist that they do not serve the reductionist's purpose. They may help fix the reference of a new predicate, but only because we antecedently had a variable ranging over the referent.[39]

[38] Alan Weir, "Naive Set Theory, Paraconsistency and Indeterminacy: Part II," *Logique & Analyse*, CLXVII–CLXVIII: 283–340, discusses infinitary languages all of whose sentences can be conjoined. But he is working in a non-standard set-theoretical framework that requires a revision of classical logic.

[39] Billy Dunaway, "Supervenience Arguments and Normative Non-naturalism," *Philosophy and Phenomenological Research*, 91: 627–655, objects to the Jackson–Stalnaker argument on the grounds that definability is too cheap, and does not even require global supervenience. He is in effect considering definability in a second-order language where the range of the second-order quantifiers is restricted. My response would be that it is only *impredicative* definability that is too cheap in the debate about the reduction principle.

7 Conclusion

I have shown that global supervenience does not entail reducibility, not even given the very generous assumption that definability in an infinitary language, with no cardinality restrictions, suffices for reducibility. The Jackson–Stalnaker argument for the opposite conclusion relies on the hidden premise that world sizes are bounded—a claim we have good reason to reject.

I would like to finish by mapping out the relationship between three concepts that played a role in this paper: strong supervenience (SS), global supervenience (GS), and $\mathcal{L}$-definability (Df). Notice that Definition 1 endows this last relation with many of the characteristic features of supervenience relations, such as being reflexive, transitive, and monotonic.

It is immediate from their respective definitions (Definition 2 and Definition 4, given in Appendix A) that *SS(A,B)* entails *GS(A,B)*—that is, that strong supervenience claims entail the corresponding global supervenience claims. Moreover, that *Df(A,B)* entails *GS(A,B)* can be shown by induction on the complexity of B-formulas of $\mathcal{L}$. Hence global supervenience is entailed by both strong supervenience and by $\mathcal{L}$-definability. Without further assumptions, the converse entailments are not guaranteed, and strong supervenience does not entail $\mathcal{L}$-definability, nor vice versa.

However, we have seen that certain conditions secure further entailments.

Given a B that is closed under $\mathcal{L}$ ($B = \mathcal{L}(B)$), we can infer with Proposition 2 that *GS(A,B)* entails *SS(A,B)*, and hence that global supervenience and strong supervenience are equivalent in such a case. Since *Df(A,B)* entails *GS(A,B)*, it also follows that $\mathcal{L}$-definability entails strong supervenience.

Given a B that is a set, *SS(A,B)* entails *Df(A,B)*, by Proposition 1.

So if B satisfies both the condition of being a set and of being closed under $\mathcal{L}$-definability, then exactly the same classes A will be strongly supervenient on, globally supervenient on, and $\mathcal{L}$-definable from B. The three notions collapse. But given Proposition 3, no B satisfies both these conditions unless there is an upper bound on the size of possible worlds. As we have seen in section 3, it is highly implausible that there is such an upper bound.

The results of this paper suggest avenues for further work on both sides of the debate on reductionism.

The paper has brought good news for those who wish to endorse supervenience claims while eschewing reductionism: the most formidable argument against their combination of views, the Jackson–Stalnaker argument, does not work. They now have a foot in the door. For them, the challenge is to push it open further, by supplying a story that makes it plausible that moral or mental properties, say, fall in the same category as our relation S from section 4.

The paper has brought bad news for those who wish to squeeze out a reductionist thesis from a supervenience claim. But the failure of the Jackson–Stalnaker argument does not imply that they need entirely new tools if they wish to close the door on supervenience without reduction again.

I recommend that they work with the concept of strong supervenience. As we have seen, if the base B is a set, then strong supervenience, but not global

supervenience, entails the $\mathcal{L}$-definability of A in terms of B. This is because the pertinent argument does not appeal to the closure of B under $\mathcal{L}$-definability—the step where bounds on worlds sizes become relevant. When formulating their strong supervenience claim, aspiring reductionists need to take care that their specification of a base for A ensures that it is a set. Such a base may well contain many highly extrinsic properties. But there are limits: it needs to be possible for objects to be indiscernible with respect to all the base properties even though they are in worlds with different domain sizes. Specifically, the base cannot be closed under $\mathcal{L}$-definability.

Strong supervenience may thus afford a route to reductionism that goes via supervenience, despite the negative result of this paper. But in many philosophical contexts, that road is arduous, since it is the global supervenience claims that are immediately obvious. In so far as we have intuitions about whether strong supervenience claims involving highly extrinsic properties and relations hold, they are typically derived from intuitions about global supervenience claims. I conclude that when evaluating reductive hypotheses in metaphysics, we can use strong supervenience as a tool—but we need to specify the putative base, and the extent to which it includes extrinsic properties, much more carefully than we have been accustomed to.

A Supervenience, sets, and classes: proofs

A well-known result due to Kim states that if a set A of monadic properties strongly supervenes on a set of monadic properties B, then each member of A is definable from B. To generalize this to relations, as Proposition 1 does, I will use a definition of strong supervenience that applies to sets of relations, unlike the standard one.[40]

For worlds w and w', a *partial domain-isomorphism* is a bijection from $X \subseteq D(w)$ to $Y \subseteq D(w')$. A partial domain-isomorphism f from w to w' *preserves* R just in case $\langle x_1, \ldots, x_n \rangle$ has R in w iff $\langle f(x_1), \ldots, f(x_n) \rangle$ has R in w'; it preserves A—is a partial A-isomorphism—iff it preserves every $R \in A$.

Definition 4. *A class of relations A* strongly supervenes *on a class of relations B iff every partial B-isomorphism is a partial A-isomorphism.*

The following is then a generalization of Kim's result.

Proposition 1. *For all sets of relations A and B, if A strongly supervenes on B, then A is $\mathcal{L}$-definable from B.*

Before proving this, I introduce some auxiliary notions.

I shall write '$\phi[g]$ is true in w' to mean that ϕ is true in w under w-valuation g(where a w-valuation maps variables only to elements of the domain of w). Let V be a set of variables, and let $at(B, V)$ be the class of atomic formulas all whose non-logical predicates are in B, and all whose variables in V. For a world w and a w-valuation g, we define:

- $\mathrm{POS}^{B,V}_{w,g} = \bigwedge\{\phi : \phi \in at(B, V)$, and $\phi[g]$ is true in $w\}$
- $\mathrm{NEG}^{B,V}_{w,g} = \bigwedge\{\neg\phi : \phi \in at(B, V)$, and $\phi[g]$ is not true in $w\}$
- $\mathrm{LOC}^{B,V}_{w,g} = \mathrm{POS}^{B,V}_{w,g} \wedge \mathrm{NEG}^{B,V}_{w,\sigma}$

Heuristically, we can think $\mathrm{POS}^{B,V}_{w,g}$ as giving a complete *positive* description of w with respect to the things in the range of g, and $\mathrm{NEG}^{B,V}_{w,g}$ a complete *negative* description of w with respect to those things—the features they do not have. $\mathrm{LOC}^{B,V}_{w,g}$ is then a *local* description in the sense of only concerning the things in $g[V]$, which may be a subset of the domain of w.

If g maps x_i to a_i, for $1 \leq i \leq n$, we can think of $\mathrm{LOC}^{B,V}_{w,g}$ as expressing the n-ary relation that holds of n things if they have the same B-profile as the a_i's have in w. Clearly, $\mathrm{LOC}^{B,V}_{w,g}[g]$ is true in w.

If B and V are sets, so is $at(B, V)$. Hence $\mathrm{LOC}^{B,V}_{w,g}$ is indeed a sentence of $\mathcal{L}$. Since every sentence $\mathrm{LOC}^{B,V}_{w,g}$ corresponds to a subclass of $at(B, V)$, the class of all sentences $\mathrm{LOC}^{B,V}_{w,g}$, for any world w and function g from V to the domain of w, also forms a set.

[40] Leuenberger, "Supervenience in Metaphysics," *op. cit.*

Lemma 1. *Suppose that for some w' and w'-valuation g, $LOC^{B,V}_{w',g'}[g]$ is true in w. Then there is a partial B-isomorphism f from w to w' such that $f(g(x)) = g'(x)$ for all $x \in V$.*

Proof. To show that $f(g(x)) = g'(x)$ defines a function from $g[V]$ to $g'[V]$, we need to verify that for all x and y in V, if $g(x) = g(y)$, then $g'(x) = g'(y)$. Suppose $g(x) = g(y)$. Then $(x = y)[g]$ is true in w. Since $\mathrm{LOC}^{B,V}_{w',g'}[g]$ is true in w, and $(x = y) \in at(B, V)$, $(x = y)$ is a conjunct in $\mathrm{POS}^{B,V}_{w',g'}$. Hence $(x = y)[g']$ is true in w', and thus $g'(x) = g(y)$. The same argument, *mutatis mutandis*, shows that if $g'(x) = g'(y)$, then $g(x) = g(y)$, and hence that f is one-one. To show that f is onto, suppose that $a \in g'[V]$. Then for some variable x in V, $g'(x) = a$, and thus $a = fb$ for $b = g(x)$.

To show that f preserves B, suppose that $Fx_{j(1)} \ldots x_{j(n)}[g]$ is true in w, where $x_{j(i)} \in V$ for $1 \leq i \leq n$. Since $Fx_{j(1)} \ldots x_{j(n)} \in at(B, V)$, and since $\mathrm{LOC}^{B,V}_{w',g'}[g]$ is true in w, $Fx_{j(1)} \ldots x_{j(n)}$ is a conjunct in $\mathrm{POS}^{B,V}_{w',g'}$. Hence $Fx_{j(1)} \ldots x_{j(n)}[g']$ is true in w'. The argument from the falsity of $Fx_{j(1)} \ldots x_{j(n)}[g]$ in w to the falsity of $Fx_{j(1)} \ldots x_{j(n)}[g']$ in w' is similar. □

Proof of Proposition 1. Pick n-ary $F \in A$. Let V be $\{x_1, \ldots, x_n\}$. Let B_F be $\{\mathrm{LOC}^{B,V}_{w,g}$: $Fx_1 \ldots x_n[g]$ is true in $w\}$. Then $\bigvee B_F \in \mathcal{L}(B)$. To establish the Proposition, it thus suffices to show for every world w and w-valuation g, $Fx_1 \ldots x_n[g]$ is true in w iff $\bigvee B_F[g]$ is true in w.

⇒: Suppose that $Fx_1 \ldots x_n[g]$ is true in w. Then $\mathrm{LOC}^{B,V}_{w,g} \in B_F$. Since $\mathrm{LOC}^{B,V}_{w,g}[g]$ is true in w, $\bigvee B_F[g]$ is true in w.

⇐: Suppose that $\bigvee B_F[g]$ is true in w. Then for some w' and w'-valuation g', $\mathrm{LOC}^{B,V}_{w',g'} \in B_F$ and $\mathrm{LOC}^{B,V}_{w',g'}[g]$ is true in w. Hence $Fx_1 \ldots x_n[g']$ is true in w'. By Lemma 1, there is a partial B-isomorphism f from w to w' such that $f(g(x)) = g'(x)$ for all $x \in V$. Since $F \in A$ and A strongly supervenes on B, f preserves F. Hence $Fx_1 \ldots x_n[g]$ is true in w. □

Remark: Since all relations in B are of finite adicity, A is in fact $\mathcal{L}_{\infty\omega}$-definable from B given the assumptions of Proposition 1.

In preparation for proving Proposition 2, some further definitions:

- $\mathrm{COM}^V_w = \forall y(\bigvee\{x = y : x \in V\})$
- $\mathrm{GLO}^{B,V}_{w,g} = \mathrm{LOC}^{B,V}_{w,g} \wedge \mathrm{COM}_{w,g}$

Heuristically, we can think of COM_w as asserting that the values of the variables in V provide a *complete* inventory of the domain of the world w, and of $\mathrm{GLO}^{B,V}_{w,g}$ as a *global* description of w, covering everything in the domain of w.

Proposition 2. *For all classes of relations A and B, if A globally supervenes on B, then A strongly supervenes on $\mathcal{L}(B)$.*

Proof. Suppose that A globally supervenes on B, and let f be a partial $\mathcal{L}(B)$-isomorphism from w to w'. Pick n-ary $R \in A$. In light of Definition 4, we need to show that f preserves R.

Suppose $\langle a_1, \ldots, a_n \rangle$ instantiate R, with $a_1, \ldots, a_n$ in the domain of f. Pick a class of variables V equinumerous with $D(w)$, and a w-valuation g whose restriction to V is a bijection from V onto $D(w)$. Then $\mathrm{GLO}^{B,V}_{w,g}[g]$ is true in w.

Let V_g be $\{x : g(x) \in \{a_1, \ldots, a_n\}\}$, and $V' = V \setminus V_g$. Then the formula $\exists V'$ $\mathrm{GLO}^{B,V}_{w,g}$ expresses a relation in $\mathcal{L}(B)$. Since f preserves $\mathcal{L}(B)$, $\exists V'$ $\mathrm{GLO}^{B,V}_{w,g}[g']$ is true in w', where $g'(y) = f(g(y))$ whenever $y \in V_g$. By the evaluation clause for the existential quantifier, $\mathrm{GLO}^{B,V}_{w,g}[g'']$ is true in w', for some V'-variant g'' of g'. By Lemma 1, there is a partial B-isomorphism f^* from w' to w such that $f^*(g''(x)) = g(x)$ for all $x \in V$. To show that f^* is onto $D(w)$, suppose that $a \in D(w)$. Since g is onto $D(w)$, $a = g(x)$ for some $x \in V$. Then $f^*(g''(x)) = a$. Moreover, since COM^V_w is true, $g''[V]$ is $D(w')$. Hence f^* is a B-isomorphism from w' to w. Then $f' = f^{*-1}$ is a B-isomorphism from w to w'. By Definition 2, and the hypothesis that A globally supervenes on B, f' is an A-isomorphism. Since g' and g'' agree on $\{y : g(y) \in \{a_1, \ldots, a_n\}\}$, f' extends f, and hence f also preserves A, and in particular R. □

Proposition 3. *For any set of relations B, $\mathcal{L}(B)$ is a set iff there is a cardinality κ such that for every possible world w, the cardinality of the domain of w is less than κ.*

Proof. ⇒: Suppose that for any set of relations B, $\mathcal{L}(B)$ is a set. Then in particular, $\mathcal{L}(\emptyset)$ is a set. For a given κ that is a cardinality of some world, let V be a set of variables of size κ. Then the sentence $\exists V(\bigwedge\{\neg(x = y) : x, y \text{ distinct variables in } V\} \wedge \forall y(\bigvee\{x = y : x \in V\}))$ is true in w iff the cardinality of w is κ. If there is a world with κ things, we can take the sentence to express the property "being such that there are exactly κ things", and otherwise a property that is necessarily uninstantiated. Since $\mathcal{L}(\emptyset)$ is a set, there is a set λ of such cardinals κ such that the property "being such that there are exactly κ things" is expressible in $\mathcal{L}(\emptyset)$.

By the axiom of union, the union of λ is a set. Hence it has a cardinality ν. It follows that no member of λ has a cardinality higher than ν. Since the cardinalities of worlds are the members of λ, ν is an upper bound for the cardinality of worlds.

⇐: Suppose that κ is an upper bound for the size of worlds, and let B have cardinality λ. Then $\nu = 2^{\lambda\kappa^\omega}$ is an upper bound for how many different worlds there are, up to B-isomorphism. (Since $\mathcal{L}(B)$-formulas that have the same extension in a world w also have the same extension in every w' that is B-isomorphic to w, we can take equivalence classes of B-isomorphic worlds, in effect.) Further, $(2^{(\kappa^\omega)})^\nu$ is an upper bound for how many distinct relations there can be in ν worlds of cardinality κ. □

B Limitations of infinitary languages: proofs

Recall that for a predicate F and a formula ϕ, ϕ^F results from restricting quantifiers in ϕ to F. For a world w, let F_w denote the extension of predicate F in

w. Then the following is a version of the Löwenheim Skolem theorem, proved by the method of closing off under Skolem functions:

Theorem 1. *Let ν be any cardinal, and H a predicate. Suppose that the size of H_w is at least 2^ν, and that there are at most ν non-logical terms in $\mathcal{L}_{\nu\nu}$. Then there is $D \subset H_w$, of size 2^ν, such that if $H'_w = D$, then for all sentences ϕ of $\mathcal{L}_{\nu\nu}$, ϕ^H is true in w iff $\phi^{H'}$ is true in w.*

Proof. The aim is to define a monotonically increasing sequence $\langle D_\alpha | \alpha < \nu^+ \rangle$ of subsets of H_w whose union $\bigcup_{\alpha<\nu^+} D_\alpha$ is the desired D that can serve as the extension of H' (where ν^+ is the cardinal successor of ν).

Since the world w is fixed for the whole proof, I shall often suppress relativisations to it.

Pick a subset X of H_w of cardinality 2^ν.

- $D_0 = X$.

Assume that D_ξ has been defined. Let $\mathcal{T}_\xi$ be the set of triples $\langle \phi, Z, f \rangle$, where ϕ is a formula of $\mathcal{L}_{\nu\nu}$, Z a subset of the variables occurring free in ϕ, and f a valuation that maps variables in Z to members of D_ξ. Say that a valuation g is *suitable* for $\langle \phi, Z, f \rangle$ iff (i) $\phi[g]$ is true, (ii) $g(x) \in H_w$ for every x free in ϕ, (iii) $g(x) = f(x)$ for all $x \in Z$, and (iv) g has the same range as its restriction to the variables free in ϕ. Let C be a function that maps a triple in $\mathcal{T}_\xi$ to a suitable g if there is one, and is undefined otherwise. Then:

- For successor ordinals ξ', $D_{\xi'} = D_\xi \cup \bigcup_{t \in \mathcal{T}_\xi} \text{Range}(C(t))$
- For limit ordinals ξ, $D_\xi = \bigcup \{ D_\alpha : \alpha < \xi \}$

Set $D =_{df} \bigcup_{\alpha<\nu^+} D_\alpha$.

Let the extension of H' be D. We need to show that the following holds for all formulas $\phi \in \mathcal{L}_{\nu\nu}$ and all g that map the free variables in ϕ to members of D:

$$\phi^H[g] \text{ is true} \Leftrightarrow \phi^{H'}[g] \text{ is true.}$$

This is proved by induction on the complexity of formulas.

Let ϕ be atomic. Then $\phi^H = \phi^{H'}$, and the claim holds trivially.

The steps for negation and conjunction are straightforward.

Let $\phi = \exists Z \psi$. For the $\Leftarrow$-direction, suppose that $\phi^{H'}[g]$ and hence $\exists Z(\bigwedge_{z \in Z} H'z \wedge \psi^{H'})[g]$ is true. It follows that $\bigwedge_{z \in Z} H'z \wedge \psi^{H'}[g']$ is true for some Z-variant g' of g. Since $\bigwedge_{z \in Z} H'z[g']$ is true, the range of g' is in D, and we can apply the induction hypothesis to obtain that $\psi^H[g']$ holds; and since $D \subseteq H_w$, $\bigwedge_{z \in Z} Hz[g']$ holds. Hence $\exists Z(\bigwedge_{z \in Z} Hz \wedge \psi^H)[g]$ and thus $\phi^H[g]$ is true.

For the $\Rightarrow$-direction, suppose $\phi^H[g]$ and hence $\exists Z(\bigwedge_{z \in Z} Hz \wedge \psi^H)[g]$ is true, where g takes its values in D. Then $\bigwedge_{z \in Z} Hz \wedge \psi^H[g']$ is true in w for some Z-variant g' of g whose range is in H_w. By cardinal arithmetic, we can show that

the range of g is included in D_α, for some α.[41] Consider the triple $\langle \psi^H, Z, g\rangle \in \mathcal{T}_\alpha$. The existence of g' ensures that C is defined for that triple; let it be g^*. By the definition of C, $\psi^H[g^*]$ is true in w. By the definition of $D_{\alpha'}$, the range of g^* is included in it, and hence in D. Hence g^* is a valuation whose range is in D. By the induction hypothesis, $\psi^{H'}[g]$ is true in w. It follows that $\phi^{H'}[g]$ is true in w.

It remains to show that $\mathrm{card}(D) = 2^\nu$. We prove by induction that for every α, D_α has cardinality 2^ν. It then follows that as a union of at most 2^ν sets of cardinality 2^ν, D has cardinality 2^ν.

By hypothesis, $D_0 = X$ has cardinality 2^ν. Assume that D_ξ has cardinality 2^ν. Then the index set $\mathcal{T}_\xi$ has at most 2^ν members. For given that $\mathcal{L}_{\nu\nu}$ has at most ν non-logical terms, there are 2^ν formulas in $\mathcal{L}_{\nu\nu}$, and at most 2^ν subsets Z of the set of the less than ν free variables in any given formula. Since there are 2^ν elements in D_ξ, by the induction hypothesis, and since there are at most ν free variables in Z, there are at most $(2^\nu)^\nu = 2^\nu$ functions from a given Z to D_ξ. Moreover, $\mathrm{Range}(C, t)$ has cardinality at most ν, since there are fewer than ν free variables in a formula. Suppose now that ξ is a limit ordinal, and that for all $\alpha < \xi$, D_ξ has cardinality 2^ν. Then D_ξ is union of at most 2^ν sets of cardinality 2^ν, and thus has cardinality 2^ν itself. □

In preparation for proving Theorem 2, recall the definition of compartmentalization:[42]

Definition 3. *A class of relations B, with $F, G \in B$, is* compartmentalized *by F and G in W iff every member of B is either restricted to F in all $w \in W$, or restricted to G in all $w \in W$, and $\forall x(Fx \leftrightarrow \neg Gx)$ is true in all $w \in W$.*

(A relation R was defined to be *restricted to* ϕ in a class of worlds W iff $\forall x_1 \ldots x_n(Rx_1 \ldots x_n \rightarrow \bigwedge_{1\leq i\leq n} \phi(x_i))$ is true in $w \in W$.)

Let B_F be the subclass of relations of B which are, in some world, instantiated by Fs; and analogously for B_G. If B is compartmentalized by F and G, these classes will be disjoint.

We divide the variables of $\mathcal{L}$ into three disjoint classes V_F, V_G, and $V \setminus (V_F \cup V_G)$, none of which is set-sized. Let $\mathrm{Form}^F_{\kappa\lambda}$ ($\mathrm{Form}^G_{\kappa\lambda}$) be $\{\phi^F : \phi$ a B_F-formula of $\mathcal{L}_{\kappa\lambda}$, with no variables in $V_G\}$ ($\{\phi^G : \phi$ a B_G-formula of $\mathcal{L}_{\kappa\lambda}$, with no variables in $V_F\}$).

Say that a w-valuation g is *sorted* just in case $Fx[g]$ is true in w whenever $x \in V_F$, and $Gy[g]$ is true in w whenever $y \in V_G$.

Definition 5. *Formulas ϕ and ψ are W-equivalent ($\phi \sim_W \psi$) just in case for all worlds $w \in W$, and all sorted w-valuations g, $(\phi \leftrightarrow \psi)[g]$ is true in w.*

[41] To see this, note that the range of g has cardinality less than ν^+. For *reductio*, suppose that it is not included in any D_α. Consider the set $\{\beta$: there is a variable x not in Z and free in ϕ such that $g(x)$ is in $D_{\beta'}$ but not in $D_\beta\}$. This set of ordinals, ordered by inclusion, is cofinal in ν^+. But then the cofinality of ν^+ is smaller than ν^+, which is impossible.

[42] The proof is adapted from those of Corollary 5.2.6 and Theorem 5.2.2 in Dickmann, *op. cit.*

The following features are immediate:

(A) $\sim^W$ is an equivalence relation.

(B) If $\phi \sim_W \phi'$, then $\neg\phi \sim_W \neg\phi'$, and if $\phi_i \sim_W \psi_i$ for all $i \in$, then $\bigwedge_{i\in I} \phi_i \sim_W \bigwedge_{i\in I} \psi_i$ and $\bigvee_{i\in I} \phi_i \sim_W \bigvee_{i\in I} \psi_i$.

(C) If ϕ and ψ are strictly equivalent, then $\phi \sim^W \psi$.

Definition 6. *A B-formula ϕ of $\mathcal{L}_{\kappa\kappa}$ is* $\sim^{\wedge}_{W}$-normalizable *in ν (relative to F and G) iff there is $\phi^{\wedge} = \bigwedge_{l\in L} \bigvee_{j\in J_l} \phi_{lj}$, such that*

(i) $\phi \sim_W \phi^{\wedge}$, and no variable is free in $\phi^{\wedge}$ that is not free in ϕ.

(ii) all ϕ_{lj} are in $Form^F_{\nu\kappa} \cup Form^G_{\nu\kappa}$.

(iii) L and J_l are smaller than ν.

(To reduce clutter, I shall sometimes write $\sim$ instead of $\sim_W$; and I shall omit $\bigwedge$ if L has one member, and $\bigvee$ if every J_l has one member.)

The definition of what it is for a B-formula ϕ of $\mathcal{L}_{\kappa\kappa}$ to be $\sim^{\vee}$*-normalizable* in ν is the same, except that what is required is the existence of a $\phi^{\vee} = \bigvee_{l\in L'} \bigwedge_{j\in J'_l} \phi_{lj}$ satisfying these three conditions. Finally, ϕ is $\sim$*-normalizable* in ν if it is $\sim^{\wedge}$-normalizable in ν as well as $\sim^{\vee}$-normalizable in ν.

Lemma 2. *If ϕ is $\sim^{\wedge}$-normalizable ($\sim^{\vee}$-normalizable) in ν, then it is $\sim^{\vee}$-normalizable ($\sim^{\wedge}$-normalizable) in 2^{2^ν}.*

Proof. Suppose ϕ is $\sim^{\wedge}$-normalizable, with $\phi^{\wedge} = \bigwedge_{l\in L} \bigvee_{j\in J_l} \phi_{lj}$ satisfying conditions (i)–(iii) of Definition 6. Define:

$$\phi^{\vee} = \bigvee_{f\in\prod_{l\in L} J_l} \bigwedge_{l\in L} \phi_{l,f(l)}$$

Clearly, $\phi^{\vee}$ is strictly equivalent to $\phi^{\wedge}$, and hence $\phi^{\vee} \sim \phi^{\wedge}$ by (C). By (A) and the fact that $\phi \sim \phi^{\wedge}$, it follows that $\phi \sim \phi^{\vee}$. Since all ϕ_{lj} are in $\mathrm{Form}^{FX}_{\nu\kappa} \cup \mathrm{Form}^G_{\nu\kappa}$, they are also in $\mathrm{Form}^F_{2^{2^\nu}\kappa} \cup \mathrm{Form}^G_{2^{2^\nu}\kappa}$. Finally, since L and all the J_l are smaller than ν, $\prod_{l\in L} J_l$ is smaller than 2^{2^ν}. Hence ϕ is $\sim^{\vee}$-normalizable in 2^{2^ν}.

The argument in the other direction is the same, *mutatis mutandis*. □

Lemma 3. *If B is compartmentalized by F and G in W, then every B-formula ϕ of $\mathcal{L}$ is $\sim$-normalizable in $c(\phi)$.*[43]

Proof. Fix a cardinal κ, a world $w \in W$, and a sorted w-valuation g. (Reference to w shall hence be omitted.) We show by induction that the result holds for all B-formulas of $\mathcal{L}_{\kappa\kappa}$. Since κ was chosen arbitrarily, and since every formula of $\mathcal{L}$ belongs to $\mathcal{L}_{\kappa\kappa}$, for some κ, the result holds.

[43] Thanks to Bruno Whittle for showing me how to solve a problem with an earlier version of the proof of this Lemma.

(1) ϕ atomic:
Let ϕ be an atomic formula of $\mathcal{L}_{\kappa\kappa}$. Then $d(\phi) = 0$, and $c(\phi) = 2^{2^\kappa}$. Set $\phi = \phi^\wedge$ if ϕ is in $\mathrm{Form}^F_{\kappa\kappa} \cup \mathrm{Form}^G_{\kappa\kappa}$, and $\phi = \bot$ otherwise. Then condition (ii) of Definition 6 is satisfied because atomic formulas and $\bot$ are in $\mathrm{Form}^F_{\kappa\kappa} \cup \mathrm{Form}^G_{\kappa\kappa}$; and condition (iii) is satisfied because L and J have just one member. As for condition (i), let g be a sorted w-valuation. We need to show that $(\phi \leftrightarrow \phi^\wedge)[g]$ is true. If ϕ is in $\mathrm{Form}^F_{\kappa\kappa} \cup \mathrm{Form}^G_{\kappa\kappa}$, we are done. So suppose it is not, and distinguish two cases.

Case (i): ϕ is of the form $v = v'$. Without loss of generality, suppose $v \in V_F$ and $v' \in V_G$. Then since g is sorted, $Fv[g]$ and $Gv'[g]$ are true. Since W is compartmentalized by F and G, $v = v'[g]$ is false, and $(v = v' \leftrightarrow \bot)[g]$ hence true.

Case (ii): ϕ is of the form $Rv_1 \dots v_n$. Without loss of generality, suppose $R \in B_F$ and $v_i \in V_G$ for some i. Since $w \in W$, and B is compartmentalized by F and G in W, $Rv_1 \dots v_n[g]$ is true in w only if $Fv_i[g]$ is true in w for every i. Since g is sorted and $v_i \in V_G$, $Gv_i[g]$ is true in w, and hence $Fv_i[g]$ false. Thus $Rv_1 \dots v_n[g]$ is false in w, and $(Rv_1 \dots v_n \leftrightarrow \bot)[g]$ is true.

Hence in all cases, $(\phi \leftrightarrow \phi^\wedge)[g]$ is true. Moreover, it is clear that no variable is free in $\phi^\wedge$ that is not free in ϕ. (In subsequent steps of the proof, verification of this condition will remain implicit.) It follows that ϕ is $\sim^\wedge$-normalizable in κ.

By the same reasoning, ϕ is $\sim^\vee$-normalizable in κ, and hence $\sim$-normalizable in κ.

(2) $\phi = \neg\psi$.
By the induction hypothesis, there is $\psi^\vee = \bigvee_{l \in L'} \bigwedge_{j \in J'_l} \phi'_{lj}$, with ϕ'_{lj} in $\mathrm{Form}^F_{\nu\kappa} \cup \mathrm{Form}^G_{\nu\kappa}$, where $\nu = c(\psi)$, and L, L', J_l and J'_l smaller than ν. Define:

$$\phi^\wedge = \bigwedge_{l \in L'} \bigvee_{j \in J'_l} \neg\phi'_{lj}$$

By the induction hypothesis, $\psi \sim \psi^\vee$. By (B), $\neg\psi \sim \neg\psi^\vee$, that is, $\phi \sim \neg\psi^\vee$. But $\phi^\wedge$ is strictly equivalent to $\neg\psi^\vee$. Hence $\phi^\wedge \sim \neg\psi^\vee$ by (C), and $\phi \sim \phi^\wedge$ by (A). This shows that ϕ is $\sim^\wedge$-normalizable in $\nu = c(\psi)$. A similar argument shows that ϕ is $\sim^\vee$-normalizable in $\nu = c(\psi)$

(3) $\phi = \bigwedge_{i \in I} \phi_i$ (with $\mathrm{card}(I < \kappa)$
Suppose that the supremum of the depths among the conjuncts is α. By the induction hypothesis, for all $i \in I$, $\phi^\wedge$ is of the form $\bigwedge_{l \in L} \bigvee_{j \in J_l} \phi_{lj}$, with ϕ_{lj} in $\mathrm{Form}^F_{\nu\kappa} \cup \mathrm{Form}^G_{\nu\kappa}$ (where $\nu = c(\alpha)$), and L, J_l smaller than ν. It is enough to show that ϕ is $\sim^\wedge$-normalizable in ν. It then follows by Lemma 2 that ϕ is $\sim^\vee$-normalizable in $c(\phi) = 2^{2^\nu}$.

Set

$$\phi^\wedge = \bigwedge_{i \in I} \phi_i^\wedge = \bigwedge_{k \in \bigcup_{i \in I} L_i} \bigvee_{j \in J_k} \phi_{k,j}$$

By (B), $\phi^\wedge \sim \phi$, such that condition (i) of Definition 6 holds. Clearly, (ii) is satisfied. Since I is smaller than $\kappa < c(\alpha)$ and every L_i smaller than $c(\alpha)$ by the induction hypothesis, $\bigcup_{i\in I} L_i$ is smaller than ν, such that (iii) is satisfied.

The case of disjunction is similar.

(4) $\phi = \forall X\psi$ (with $\mathrm{card}(X) < \kappa$)

Suppose $c(\psi) = \nu$. For every $x \in X$, let x_F (x_G) be a V_F-variable (V_G-variable) that does not occur in ψ, such that x_F and y_F (x_G and y_G) are distinct variables whenever x and y are distinct. Further, X_F is $\{x_F : x \in X\}$, and X_G $\{x_G : x \in X\}$. Given $Y \subseteq X$, $\overline{Y} =_{df} X \setminus Y$. Let $\psi(x/x_Y)$ be the result of substituting, in ψ, every variable $x \in Y$ by x_G, and every variable $x \in \overline{Y}$ by x_F. Set:

$$\phi_1 = \bigwedge_{Y\in\mathcal{P}(X)} [\forall \overline{Y}_F \forall Y_G(\bigwedge_{x\in\overline{Y}_F} Fx \wedge \bigwedge_{y\in Y_G} Gy \rightarrow \psi(x/x_Y))]$$

We need to show that $\phi \sim \phi_1$. Let g be a sorted w-valuation.

Suppose first that $\phi[g]$ is false. Then for some X-variant g' of g, $\psi[g']$ is false. Let g'' be like g except that $g''(x_F) = g'(x)$ for all $x \in X$ such that $Fx[g']$ is true, and $g''(x_G) = g'(x)$ for all $x \in X$ such that $Gx[g']$ is true. It can be verified that for $Y = \{x \in X : Gx[g'] \text{ is true}\}$, $\psi(x/x_Y)[g'']$ is true iff $\psi[g']$ is true. Hence $\psi[g']$ is false. Moreover, g'' differs from g at most on the variables in $\overline{Y}_F$ and Y_G. Hence $\forall\overline{Y}_F\forall Y_G(\bigwedge_{x\in\overline{Y}_F} Fx \wedge \bigwedge_{y\in Y_F} Gy \rightarrow \psi(x/x_Y)[g]$ is false, and it follows that $\phi_1[g]$ is false.

For the other direction, suppose that $\phi_1[g]$ is false. Then for some Y, $\bigwedge_{x\in\overline{Y}_F} Fx \wedge \bigwedge_{y\in Y_F} Gy \rightarrow \psi(x/x_Y)[g']$ is false for some g' that differs from g at most on the variables in $\overline{Y}_F$ and Y_G. Let g'' be like g except that $g''(x) = g'(x_F)$ whenever $x_F \in \overline{Y}_F$, and $g''(x) = g'(x_G)$ whenever x_G in Y_G. Then it can be verified that $\psi(x/x_Y)[g']$ is true iff $\psi[g'']$ is true. Moreover, g'' differs from g at most on the variables in X. Hence $\forall X\psi[g]$, that is, $\phi[g]$, is false.

By the induction hypothesis, $\psi(x/x_Y) \sim \psi(x/x_Y)^\wedge$. Set:

$$\phi_2 = \bigwedge_{Y\in\mathcal{P}(X)} [\forall \overline{Y}_F \forall Y_G(\bigwedge_{x\in\overline{Y}_F} Fx \wedge \bigwedge_{y\in Y_F} Gy \rightarrow \psi(x/x_Y)^\wedge)]$$

To show that $\phi_1 \sim \phi_2$, consider any sorted w-valuation g.

Suppose $\phi_1[g]$ is false. Then for some Y and some g' that differs from g at most on $\overline{Y}_F \cup Y_G$, $\bigwedge_{x\in\overline{Y}_F} Fx \wedge \bigwedge_{y\in Y_G} Gy \rightarrow \psi(x/x_Y)[g']$ is false. Now suppose $x \in V_F$. Then $x \notin Y_G$, since V_F and V_G are disjoint. If x is in $\overline{Y}_F\{x_F : x \in X \setminus Y\}$, then $Fx[g']$ is true because $\bigwedge_{x\in\overline{Y}_F} Fx[g']$ is true. If $x \not\in \overline{Y}_F \cup Y_G$, then $g'(x) = g(x)$, and thus $Fx[g']$ is true because g is sorted. Similarly, we show that if $x \in V_G$, then $Gx[g']$ is true. It follows that g' is sorted.

By the induction hypothesis, $(\psi(x/x_Y) \leftrightarrow \psi(x/x_Y)^\wedge)[g']$. Hence $\bigwedge_{x\in\overline{Y}_F} Fx \wedge \bigwedge_{y\in Y_G} Gy \rightarrow \psi(x/x_Y)[g']$ is false. Since g' that differs from g at most on $\overline{Y}_F \cup Y_G$, $\forall\overline{Y}_F\forall Y_G(\bigwedge_{x\in\overline{Y}_F} Fx \wedge \bigwedge_{y\in Y_F} Gy \rightarrow \psi(x/x_Y)^\wedge[g]$ is false. If follows that $\phi_2[g]$ is false.

In a similar way, we prove that if $\phi_2[g]$ is false, so is $\phi_1[g]$.

The formula $\psi(x/x_Y)^\wedge$ is of the form $\bigwedge_{l\in L_Y}\bigvee_{j\in J_l}\psi_{lj}$, with ψ_{lj} in $\mathrm{Form}^F_{\nu\kappa}\cup \mathrm{Form}^G_{\nu\kappa}$, and L and all J_l smaller than ν. Consider a conjunct of ϕ_2, for some Y:

$$\forall\overline{Y}_F\forall Y_G(\bigwedge_{x\in\overline{Y}_F} Fx\wedge\bigwedge_{y\in Y_F} Gy\rightarrow\bigwedge_{l\in L_Y}\bigvee_{j\in J_l}\psi_{lj})$$

This is logically equivalent to:

$$\bigwedge_{l\in L_Y}[\forall\overline{Y}_F\forall Y_G(\bigwedge_{x\in\overline{Y}_F} Fx\wedge\bigwedge_{y\in Y_F} Gy\rightarrow\bigvee_{j\in J_l}\psi_{lj})]$$

Consider now a conjunct of this conjunction, for some $l\in L$:

$$\forall\overline{Y}_F\forall Y_G(\bigwedge_{x\in\overline{Y}_F} Fx\wedge\bigwedge_{y\in Y_F} Gy\rightarrow\bigvee_{j\in J_l}\psi_{lj})$$

Let $J_{F,l}$ ($J_{G,l}$) be $\{j:\psi_{lj}$ is in $\mathrm{Form}^F_{\nu\kappa}\}$ ($\{j:\psi_{lj}$ is in $\mathrm{Form}^G_{\nu\kappa}\}$). Since for every l and j, ψ_{lj} is in $\mathrm{Form}^F_{\nu\kappa}\cup\mathrm{Form}^G_{\nu\kappa}$, $\bigvee_{j\in J_l}\psi_{lj}$ is equivalent to $\bigvee_{j\in J_{F,l}}\psi_{lj}\vee\bigvee_{j\in J_{G,l}}\psi_{lj}$. Hence the above is equivalent to:

$$\forall\overline{Y}_F\forall Y_G(\bigwedge_{x\in\overline{Y}_F} Fx\wedge\bigwedge_{y\in Y_F} Gy\rightarrow\bigvee_{j\in J_{F,l}}\psi_{lj}\vee\bigvee_{j\in J_{G,l}}\psi_{lj})$$

Since no variables in V_F occur in $\bigvee_{j\in J_{G,l}}\psi_{lj}$, and no variables in V_G occur in $\bigvee_{j\in J_{G,l}}\psi_{lj}$, this is logically equivalent to:

$$(\forall\overline{Y}_F\bigwedge_{x\in\overline{Y}_F} Fx\rightarrow\bigvee_{j\in J_{F,l}}\psi_{lj})\vee(\forall Y_G\bigwedge_{y\in Y_F} Gy\rightarrow\bigvee_{j\in J_{G,l}}\psi_{lj})$$

By repeated use of (B), we can show that $\phi\sim\phi^\wedge$, for $\phi^\wedge$ defined as follows:

$$\phi^\wedge=\bigwedge_{Y\in\mathcal{P}(X)}\bigwedge_{l\in L}[(\forall\overline{Y}_F(\bigwedge_{x\in\overline{Y}_F} Fx\rightarrow\bigvee_{j\in J_{F,l}}\psi_{lj})\vee(\forall Y_G\bigwedge_{y\in Y_F} Gy\rightarrow\bigvee_{j\in J_{G,l}}\psi_{lj})]$$

Since for each l and each $j\in J_{F,l}$, ψ_{lj} is in $\mathrm{Form}^F_{\nu\kappa}$, $J_{F,l}$ is smaller than ν, and X smaller than κ, $\forall\overline{Y}_F(\bigwedge_{x\in\overline{Y}_F}Fx\rightarrow\bigvee_{j\in J_{F,l}}\psi_{lj})$ is in $\mathrm{Form}^F_{\nu\kappa}$; a similar argument shows that $\forall Y_G(\bigwedge_{x\in Y_G}Gx\rightarrow\bigvee_{j\in J_{G,l}}\psi_{lj})$ is in $\mathrm{Form}^G_{\nu\kappa}$. Since $Y\leq 2^{2^\kappa}\leq\nu$, there are fewer than ν conjuncts. Hence ϕ is $\sim^\wedge$-normalizable in ν. By Lemma 2, ϕ is normalizable in $c(\phi)$.

The case of existential quantification is similar.

□

Theorem 2. *Let B be compartmentalized by F and G in W, and ϕ be a B-sentence of $\mathcal{L}_{\kappa\kappa}$. Then if for all $w,w'\in W$ and B-sentences ψ of $\mathcal{L}_{\nu\kappa}$, where $\nu=c(\phi)$, ψ^F and ψ^G have the same truth-value in w as they have in w', so does ϕ.*

Proof. Let ϕ be a B-sentence of $\mathcal{L}_{\kappa\kappa}$, of depth α. By Lemma 3, ϕ is $\sim$-normalizable in $\nu = c(\phi)$. That is, there is a B-sentence $\phi^\wedge$ of $\mathcal{L}_{\nu\kappa}$ such that $\phi^\wedge \sim \phi$. By Definition 5, and since w and w' are in W, the following holds for every sorted variable assignment g:

- $\phi[g]$ is true in w iff $\phi^\wedge[g]$ is true in w.
- $\phi[g]$ is true in w' iff $\phi^\wedge[g]$ is true in w'.

Since ϕ is a sentence, so is $\phi^\wedge$, given that no variable is free in the latter that is not free in the former. As sentences, they are true iff they are true under every sorted assignment. Hence:

- ϕ is true in w iff $\phi^\wedge$ is true in w.
- ϕ is true in w' iff $\phi^\wedge$ is true in w'.

The formula $\phi^\wedge$ is of the form $\phi^\wedge = \bigwedge_{l\in L}\bigvee_{j\in J_l}\phi_{lj}$, where all ϕ_{lj} are in $\mathrm{Form}^F_{\nu\kappa}\cup\mathrm{Form}^G_{\nu\kappa}$. Since for all B-sentences ψ of $\mathcal{L}_{\nu\kappa}$, ψ^F and ψ^G have the same truth-value in w as they have in w', so do ϕ_{lj}. By a trivial induction, we obtain:

- $\phi^\wedge$ is true in w iff $\phi^\wedge$ is true in w'.

Putting the last three bulletted biconditionals together, it follows that ϕ is true in w iff it is true in w'. □