To Be F Is To Be G

Cian Dorr

Forthcoming in *Philosophical Perspectives*
Revised version: 4th November 2016
Comments welcome

1 Identifications

I am interested in a certain way of understanding claims of the form 'To be F is to be G', which I take to have played a central role in philosophy from its inception. Here are some examples where the target reading is natural:

(1) a. To be a vixen is to be a female fox.

b. To be square is to be rectangular and equilateral.

c. To be just is to be such that each part of one's soul does its own proper work.

d. To be a human being is to be a rational animal.

e. To be a hydrogen atom is to be an atom whose nucleus contains exactly one proton.

f. To be a continuous function is to be such that for every open set in one's range, the set of things in one's domain which one maps to members of that set is open.

As (1c) and (1d) illustrate, questions whose answers can be given in the form 'To be F is to be G' have been of central interest to philosophers since the beginning. (1e) illustrates that we cannot always tell whether to be F is to be G using "armchair" methods: sometimes, we need to do experiments. But not always, as witness (1f).

The reading I am interested in is not the only possible reading of 'To be F is to be G'. The readings in play in the following examples seem rather different in character:

(2) a. To be a black athlete in Colombia is to be constantly reminded of your otherness.[1]

[1] `http://www.okayafrica.com/controversial/afro-colombian-gold-medalist/`.

b. To be a teacher is to be forever an optimist.[2]

c. To be red is to be coloured.

d. To be a vixen is to be female.

Here are some diagnostics which may help us isolate the target reading. First: one can encourage the target reading by emphasising (focusing) the word 'is'. One says: 'To be a vixen *is* to be a female fox'. Or: 'To be a vixen *just is* to be a female fox'. Second: on the target reading, 'To be F is to be G' can be rephrased as 'To be F and to be G are the same/are one and the same', or (perhaps more idiomatically) 'Being F and being G are the same'.[3] Third: on the target reading, 'To be F is to be G' is non-contingent: necessary if true, impossible if false. Fourth: on the target reading, 'To be F is to be G' entails 'Necessarily, all and only F things are G'. And fifth: on the target reading, the claim that to be F is to be G constitutes a very satisfying *explanation* of the fact that necessarily, all and only F things are G. One might be puzzled as to why it should be necessary that everything F is G, or that everything G is F. But if to be F *is* to be G, there is nothing more to be puzzled about.

The work 'is' does in 'To be F is to be G' on the target reading seems very like the work it does in paradigm cases where it is said to express identity, like 'Hesperus is Phosphorus' and 'This is she'. Indeed, I think that this parallel is the key to understanding our 'is'. However, there is a way of explaining the parallel that we must be wary of, which is to treat the target sentences simply as ways of talking about the identity of *properties*, so that (1a), for example, would be assimilated to (3):

(3) The property of being a vixen is the property of being a female fox.

This assimilation misses the fact that the expressions flanking 'is' in (1a) and (3) differ in syntactic category:

(4) a. I hope to be an astronaut.

b. *I hope the property of being an astronaut.

The inference from (1a) to (3) seems similar in status to the inference from 'Nelly is a vixen' to 'Nelly has the property of being a vixen'. Nominalists who deny that the latter inference is strictly truth-preserving (on the grounds that strictly speaking there *are* no properties) will presumably say the same about the former inference. It is not incoherent to be a nominalist while fully endorsing claims like (1a); or at least, such

[2] `http://www.presidency.ucsb.edu/ws/?pid=56389`.

[3] Zoltan Szabó (p.c.) tells me that whereas the grammaticality of 'To be F is to be G' is idiosyncratic to English, the forms with 'are the same' are more widespread cross-linguistically.

a combination is incoherent only if nominalism itself is incoherent. Indeed, even those with no nominalistic sympathies have reason to be careful about the inference from 'To be F is to be G' to 'The property of being F is the property of being G'. For by such a move one can get from the unproblematic (5a) to the paradoxical, deeply controversial (5b):

(5) a. To be a non-self-instantiator is to fail to instantiate oneself

b. The property of being a non-self-instantiator is the property of failing to instantiate oneself

For these reasons, explicit property-talk will only play a heuristic role in our investigation of 'To be F is to be G'.

One important parallel between the 'is' in the target sentences and the 'is' in 'Hesperus is Phosphorus' involves the point I made earlier about explaining necessity. The fact that Hesperus is Phosphorus explains in a supremely satisfying way why it is necessary that everyone who lands on Hesperus will land on Phosphorus. Identities are excellent stopping places for explanation; they do not cry out for explanation in their own right. Indeed, there is something odd about questions like 'Why is Hesperus Phosphorus?'. Unless this is understood as a request to be reminded of the reasons for *believing* that Hesperus is Phosphorus, it is hard to know what would count as a satisfying answer. It is tempting to respond by citing some metalinguistic facts, as if one had been asked why 'Hesperus' refers to the same thing as 'Phosphorus'. But of course that is a quite different question. And this also applies to questions like 'Why is it that to be a vixen is to be a female fox'? Once we set aside the "remind me of reasons to believe" reading, and metalinguistic questions about the word 'vixen', it is hard to see what an answer would even look like.[4]

There are other environments besides 'To be F is to be G' where something that looks like the 'is' of identity occurs with arguments that are syntactically different from ordinary determiner phrases (like 'Hesperus' and 'the property of being an astronaut'); despite the title, these other constructions will also be part of the topic of this paper. For one thing, we can put untensed verb phrases involving verbs other than 'be' on either side of 'is':

(6) a. To be a triangle is to have three angles

[4]Rayo 2013 strongly emphasises this point, which he calls 'why closure'. However 'Why is it that to be F is to be G?' is not always odd in this way. 'Why is it that to be a hungry vixen is to be a female fox feeling in need of food?' seems to be quite reasonably answered by 'Because to be a vixen is to be a female fox, and to be hungry is to feel in need of food'. The oddity seems specific to the case were at least one of the expressions flanking 'is' lacks relevant syntactic complexity, like 'to be a vixen'. Thanks to Timothy Williamson for discussion on this point.

b. To die is to cease to live

c. To kill is to cause to die

d. ? To resemble is to be similar to

e. To kill something is to cause it to die

We can also use gerunds:

(7) a. Being square is being rectangular and equilateral

b. Killing is causing to die

c. Being a non-self-instantiator is not instantiating oneself

These seem equivalent to their infinitival counterparts ((1b), (6c), and (5a)).[5] Finally, 'is' can be flanked by untensed clauses headed by 'for':

(8) a. For there to be vixens is for there to be female foxes

b. For Obama to be a bachelor is for Obama to be an unmarried man

c. For something to be square is for it to be rectangular and equilateral

d. For someone to kill someone is for them to cause them to die

e. For a line x to be parallel to another line y is for x and y to be coplanar and non-intersecting.

'To be F is to be G' seems obviously equivalent (on the relevant reading) to 'For something to be F is for it to be G'. But 'for' clauses are more flexible: for example, they seem to provide the only way of getting across what is expressed by (8b) or (8e).

I will introduce the colourless label 'identifications' for sentences in which 'is' is the main verb, flanked by expressions of the kinds just considered, and understood in the way I have tried to make salient. In other work I have used other labels for the same class of sentences: 'metaphysical analyses' (Dorr 2004, 2005) and 'real definitions' (Dorr 2007). But I will not use these labels here, since it is artificial to speak of 'analyses' and 'definitions' as species of *sentences*: a proper account of analyses and definitions should go hand in hand with an account of the activities of analysing

[5] It is somewhat more tempting to assimilate 'Being F is being G' to 'The property of being F is the property of being G', perhaps using the italicised '*Being F* is *being G*' as an intermediary. But so long as one recognises an unproblematic reading for (5a), it seems arbitrary to deny that (7c) has a corresponding reading.

and defining, of which they seem to be the products (in the sense in which assertions and promises are the products of asserting and promising).[6] These activities seem to centrally involve believing, knowing, and/or asserting things expressible by identifications: for example, 'She analysed lying as intentionally misleading' seems roughly equivalent to 'She maintained that to lie is to intentionally mislead'.[7] However, in this paper we will be concerned with the activities only insofar as we are concerned with their cognitive subject matter.

The questions I will be investigating concern what one might call the *logic* of identifications. But all I mean by this is that the questions are extremely general ones. For example, we will not consider whether to be morally right is to maximise happiness, but we will ask whether, for arbitrary F and G, to be F is to be F and either G or not G. There is no suggestion that the answers to 'logical' questions must be in some sense non-substantive, or analytic, or neutral with respect to the more specific disputes. On the contrary, I believe that these questions are among the hardest and deepest questions in metaphysics, and that differences in how we answer them will interact very significantly with differences in how we approach more specific questions.

The order of events will be as follows. Section 2 will articulate some logical principles that should be relatively uncontroversial, insofar as they simply generalise standard principles of the logic of identity. Sections 3 and 4 will introduce some formal tools for regimenting identifications, and re-express the basic principles from §2 in terms of them. The remainder of the paper will then use these tools to formulate certain further principles that I think *should* be controversial, and to explore some of their consequences.

2 Basics

Given the intimate relation we have seen between the 'is' in identifications and the 'is' in ordinary identity sentences, it is clear that the logic of identifications must be importantly parallel to the logic of identity. At a minimum, all instances of the following schemas had better be true:

[6] Also, 'real definition' is opposed to 'nominal definition', so there is pressure to take them to be two species of some genus. If we take real definitions to be declarative sentences, we will therefore be expected to take nominal definitions to be declarative sentences too. But which sentences would they be?

[7] The same goes for 'She reduced lying to intentionally misleading'. But there are several other uses of 'reduction' floating around. In my view, the use of the word 'reduction' is such a mess that we would do better to ban it.

Reflexivity: To be F is to be F

Transitivity: If to be F is to be G and to be G is to be H, then to be F is to be H

Symmetry: If to be F is to be G, then to be G is to be F

I have come across some resistance to *Symmetry*—indeed, I seem to have once rejected it myself. But it now strikes me as manifestly valid.[8] If to be a vixen simply *is* to be a female fox—if being a vixen and being a female fox are one and the same—then of course it is equally true that to be a female fox is to be a vixen. However, there are two factors which may mask the obviousness of this implication. First, *Symmetry* fails for some of the other readings of 'To be F is to be G' which I attempted to set aside in §1—'To be red is to be coloured' has a true reading, whereas 'To be coloured is to be red' seems not to. And second, even on the target reading, there are pragmatic factors that make 'To be a vixen is to be a female fox' a more natural-sounding speech than 'To be a female fox is to be a vixen'. These are plausibly explained in terms of the idea that certain declarative sentences make salient certain questions, to which they present themselves as helpful answers.[9] In particular, 'To be F is to be G' suggests the question 'What is it to be F?' in a way that it does not suggest the question 'What is it to be G?'. And while 'To be a vixen is to be a female fox' is a helpful answer to 'What is it to be a vixen?', 'To be a female fox is to be a vixen' is—although true—not a *helpful* answer to 'What is it to be a female fox?'. Someone who asked this would more likely be hoping for something like 'To be a female fox is to be a fox with two X chromosones'. However, these pragmatic effects are quite context-sensitive: there are settings where 'To be a female fox *is* to be vixen' seems completely fine, for example that of an argument about whether some particular animal is a vixen or not.[10]

Some resist *Symmetry* because they think identifications are intimately connected in some way with some asymmetric notion such as "metaphysical priority" or "grounding". To my mind, the most plausible way to forge such a connection—which we will revisit in §9—is to claim that 'To be a vixen is to be a female fox' entails that being female and being a fox are each metaphysically prior to being a vixen. Perhaps it also entails that being a fox and being female jointly *ground* being a vixen. But we should be careful to distinguish these claims, which are perfectly

[8]Thanks to Kieran Setiya for convincing me of this.

[9]This intuitive thought has been fruitful in explaining many pragmatic phenomena: see, e.g., Simons et al. (2010).

[10]This ordering-bias is especially noticeable in claims of the form 'What it is to be F is to be G'. But while I am not sure what to make of this 'What it is' syntactically or semantically, it seems unlikely that it could make for a difference in truth value.

consistent with *Symmetry*, from the claims that 'To be a vixen is to be a female fox' entails that being a female fox is metaphysically prior to, or a ground of, being a vixen. If one thinks that being a vixen *is* being a female fox, I don't think one should feel any pull towards these latter claims; but they are what one would need to rely on to make a priority- or grounding-theoretic argument against *Symmetry*.

If the 'is' in identifications works like 'is' in ordinary identity sentences, we should also expect its logic to include some analogue of Leibniz's Law or the principle of substitution of identicals. For example, the following seem to be consequences of (1b) ('To be square is to be rectangular and equilateral'):

(9) a. Everything square is rectangular and equilateral.

b. If it is possible for there to be a square planet, it is possible for there to be a rectangular and equilateral planet.

c. To be either round or square is to be either round, or rectangular and equilateral.

d. For it to be necessary that everything square is square is for it to be necessary that everything square is rectangular and equilateral.

Perhaps there are exceptions to this pattern. For example, (10a)–(10c) seem *prima facie* to be false despite the truth of (1b):

(10) a. If someone wants a square garden, someone wants a rectangular and equilateral garden.

b. Whoever believes that something is square believes that it is rectangular and equilateral.

c. To say that everyone believes that everything square is square is to say that everyone believes that everything square is rectangular and equilateral

The case is not beyond dispute; the literature on attitude reports contains a whole battery of techniques which could be used to explain away the apparent invalidity of arguments like these.[11] But assuming we take the appearance at face value, it points towards something quite distinctive about linguistic environments like 'Someone

[11]Examples include Stalnaker's pragmatic techniques for defending the validity of substitution of necessarily equivalent sentences in attitude reports (Stalnaker 1999), and the pragmatic (Salmon 1986a, Soames 1987b), error-theoretic (Braun 1988, Saul 2007), and contextualist (Dorr 2014c, Schiffer 1979) techniques that have been used to defend the validity of substitution of co-referential names in attitude reports.

believes that…'—something that distinguishes them, for example, from the environments created by 'and', 'not', 'all', 'some', and 'it is metaphysically necessary that…'. As we might put it, the former environments are "opaque", sensitive to distinctions in "mode of presentation", whereas the latter are "transparent", only concerned with distinctions "out in the world".[12]

This picture strongly suggests that any operators we might need to appeal to in stating questions that are central to the subject matter of metaphysics should be transparent. For example, if 'because' is supposed to express a "worldly" notion of explanation—something like grounding—then if we think that to be a vixen *is* to be a female fox, and reject the claim that every vixen is a vixen because it is a vixen, we must also reject (11):

(11) Every vixen is a vixen because it is a female fox.

Perhaps (11) has true readings where 'because' is understood in epistemological or psychological or metalinguistic terms. But how could it be true if it is just supposed to be about how things are, out in the world?[13]

Note that one could endorse this argument against (11) while still accepting (12):

(12) Every vixen is a vixen because it is female and it is a fox.

Given that to be vixen is a female fox and that 'because' is transparent, (12) implies

(13) Every female fox is a female fox because it is female and it is a fox.

But (13) is not obviously false, or inconsistent with the irreflexivity of 'because'. In §5 we will consider a certain fine-grained picture which systematically rejects claims like (14):

(14) For something to be a female fox is for it to be the case that it is female and it is a fox

[12]The semantically relevant notion of "mode of presentation" need not be conceived along Fregean lines. One might tie differences in modes of presentation to differences in syntactic structure, holding that the only cases where substitution of 'G' for 'F' changes the truth value of an attitude report despite the truth of 'To be F is to be G' are cases where 'F' and 'G' differ in syntactic structure, so that in particular such substitution is always legitimate when 'F' and 'G' are syntactically simple predicates like 'groundhog' and 'woodchuck' or 'doctor' and 'physician' (see Salmon 1986a, 2010, Soames 1987a).

[13]Most theorists of grounding stress the "worldly" or "metaphysical" character of grounding in introducing the notion (Audi 2012, Fine 2012, Rosen 2010). However, Correia (2010) distinguishes a 'worldly' and 'conceptual' sense of 'ground', and advocates very different logics for the two. But I have little grip on what the conceptual sense of 'ground' is supposed to be, or how it is suppose to relate to other, non-grounding-theoretic uses of 'because'.

And if (14) is false, there is no route from (13) to a violation of the irreflexivity of 'because'.

The thesis that operators that are central to the subject matter of metaphysics should be transparent is relevant to the suggestion, which I have encountered in discussion, that the use of 'To be F is to be G' that is of most importance for metaphysics is one of the readings for which *Symmetry* fails. I imagine that those who made this suggestion were thinking that the metaphysically important reading would be one on which 'To be a vixen is to be a female fox' is true while 'To be a female fox is to be a vixen' is false.[14] I doubt that there is any such reading. But even if there were one, the operator expressing it would be opaque, and thus if the thesis is correct, it cannot have the claimed kind of importance.

The thesis is controversial enough to be worth one last round of argument. Consider how we use words introduced by explicit definitions. For example, let me now define 'schmixen' by issuing the following stipulation:

(15) To be a schmixen is to be a female fox.

When we have introduced a word like this, we can substitute the definiendum ('schmixen') for the definiens ('female fox') in a wide range of contexts. Granted, such substitutions are problematic in speech and attitude reports. Talking to someone who misheard my introduction of the word 'schmixen' and went on to get quite confused, it seems like we could speak the truth by saying 'You have been confusedly thinking that schmixens are *e-mail boxes*, but really they are female foxes', even though there is no true reading of 'You have been confusedly thinking that female foxes are e-mail boxes'. However, whatever is going on here is something quite distinctive about our practice of characterising psychological states, and does not impugn the usual practice of substituting stipulatively defined words for their definienda in non-psychological contexts. But if we are happy substituting 'schmixen' for 'female fox' in such contexts, why would we not substitute 'vixen' for 'female fox'? The differences in how 'schmixen' and 'vixen' got their meanings may be relevant to epistemology, but it is hard to see how it could matter for metaphysics. Speaking a language which didn't provide a single word meaning 'vixen' would not cut one off from any *facts*: at worst, it would diminish one's repertoire of modes of presentation of the facts.

[14]Rosen (2010, p. 123) equates 'p reduces to q' with 'For it to be the case that p just is for it to be the case that q', while taking it for granted that if p reduces to q, q does not reduce to p. Similarly Fine (2015, p. 308) proposes an asymmetric reading for 'IS' on which 'H_2O IS water' must be false if 'Water IS H_2O' is true.

3 Formalisation

To make further progress in debating the logic of identifications, it will be helpful to have a formal language in which identifications can be stated without the idiosyncratic syntactic trappings that English grammar requires. Two different approaches to formalisation suggest themselves: rather than plumping for one of them, I will present both, since they both have their advantages and there is something to be learnt by thinking about how they can be translated into one another.

The first approach is to make the formal-language counterpart of the 'is' in identifications something that combines with two *predicates* to make a sentence. So for example, we could represent 'To be triangular is to be trilateral' simply as 'Triangular ≡ Trilateral', where 'Triangular' and 'Trilateral' are two one-place predicates (so that, e.g., we could write '$\exists x(\text{Triangular}(x))$' for 'Something is triangular'). To formalise sentences like 'To be square is to be rectangular and equilateral', our language will need some mechanism for building complex predicates. English contains several such mechanisms: for example we have complex adjectival phrases like 'rectangular and equilateral', complex noun phrases like 'female fox', and complex verb phrases 'has at least one proper part'. But in a formal language we would hope to get by with something more uniform. The most familiar formal device for forming complex predicates is lambda-abstraction, whereby we combine a list of n distinct variables $v_1 \ldots v_n$ with an open sentence φ to form a one-place predicate, written as $(\lambda v_1 \ldots v_n.\varphi)$, which applies to objects $x_1, \ldots, x_n$ iff φ is true on an assignment that maps v_1 to x_1,…, and v_n to x_n. This gives us such translations as the following:

(16) a. $\text{Square} \equiv \lambda x.\, \text{Rectangular}(x) \wedge \text{Equilateral}(x)$
To be square is to be rectangular and equilateral.

b. $\text{Composite} \equiv \lambda x.\exists y(\text{ProperPart}(y, x))$
To be composite is to have a proper part.

c. $\text{Parallel} \equiv \lambda x y.\text{Line}(x) \wedge \text{Line}(y) \wedge \text{Coplanar}(x, y) \wedge \neg\text{Intersect}(x, y)$
For an object x to be parallel to an object y is for x and y to be coplanar lines that do not intersect.

We can also allow ≡ to be flanked by two *sentences*, allowing for sentences like

(16) d. $\exists x(\text{Vixen}(x)) \equiv \exists x(\text{Female}(x) \wedge \text{Fox}(x))$
For there to be vixens is for there to be female foxes.

This is no great departure, since sentences can be thought of as 0-adic predicates.[15]

[15]It would be ideally perspicuous to mark a symbolic difference between the different syntactic

Philosophers occasionally use λ-terms as formal equivalents of English expressions of the form 'The property of being F'. Given that I do not want to equate 'To be F is to be G' with 'The property of being F is the property of being G', it is important to be clear that that is *not* how I am using λ-terms—they are predicates, just like 'is square' and 'is rectangular and equilateral'. If one does want a systematic translation into English, perhaps the best option is to translate '$\lambda x.(\ldots x\ldots)$' as 'is such that …he/she/it/one …' (with the choice depending on the requirements of syntactic agreement).[16]

The fact that many English sentences may contain unpronounced variables for items like instants or intervals of time, events, or situations makes the task of translating English identifications into this formal language more complicated than it might initially seem. For example, the currently dominant approach to tense assigns additional arguments to ordinary adjectives, verbs and nouns, which are saturated by time variables present in the syntax.[17] Against this background, it is plausible that the most natural reading of 'To be a vixen is to be a female fox' would require a representation like '$(\lambda xt.\,\text{Vixen}(x,t)) \equiv (\lambda xt.\,\text{Female}(x,t) \wedge \text{Fox}(x,t))$'. Likewise, one might use event arguments to formalise 'To kill is to cause to die' as '$(\lambda exy.e$ is a killing by x of $y) \equiv (\lambda exy.e$ is a causing-to-die by x of $y)$'.[18] I will ignore these subtleties in what follows by confining my attention to stative sentences and assuming that no time arguments are needed.

The second approach to formalisation attempts to stay closer to the structure of identifications in English, especially the ones using 'for' clauses. In these sentences, pronouns on the right hand side of 'is' can be syntactically linked to indefinites on the

---

roles in which the '$\equiv$' symbol can appear—for example, between the '$\equiv$' in (16a) (which combines with two monadic predicates to make a sentence) and the one in (16c) (which combines with two dyadic predicates to make a sentence). But there is no practical need for this, since it can always be reconstructed from the types of the arguments.

[16]This departs slightly from the translations given above—the more faithful rendition of (16a) would be 'To be square is to be such that one is rectangular and one is equilateral'. This distinction does not matter if (as I believe) to be rectangular and equilateral *is* to be such that one is rectangular and one is equilateral. However, some proponents of an extreme version of the "structured" picture (see §6) may reject this identification, on the grounds that there is a difference in syntactic structure between the two sides. On this view, the kind of formal language we are currently working with is incapable of expressing the fact we express in English by 'To be square is to be rectangular and equilateral'. Expressing this will require a language that contains a "predicate functor" that can do what 'and' seems to do in this sentence: combine directly with two predicates to form a new predicate, without forming an open sentence as an intermediary.

[17]For a helpful survey, see Kusumoto 1999, ch. 1.

[18]The event-theoretic analysis of 'e is a causing-to-die by x of y' is itself controversial. One possibility is to understand it as equivalent to '$\text{Cause}(e) \wedge \text{Agent}(x,e) \wedge \exists f(\text{Die}(f) \wedge \text{Theme}(y,f) \wedge \text{Theme}(f,e))$'.

left hand side, and the meaning of the whole sentence turns on the pattern of links. Consider 'For someone to kill someone is for them to cause them to die'. This is ambiguous: using subscripts, we can distinguish the two readings as 'For someone$_1$ to kill someone$_2$ is for them$_1$ to cause them$_2$ to die' (which is roughly true) and 'For someone$_1$ to kill someone$_2$ is for them$_2$ to cause them$_1$ to die' (which is obviously false). This structural ambiguity is clearly the same sort of thing as the structural ambiguity in 'Someone told someone she would kill her': it is a matter of different patterns of "coindexing", which formal languages usually resolve by using numerically distinct variables as the counterparts of pronouns in natural language. This suggests a formal treatment where the operator playing the role of 'is' is something that always combines with two (open or closed) *sentences* to make a new sentence, and may bind some of the variables that occur free in those sentences.[19] We can write the list of bound variables as a subscript to $\equiv$.[20] This yields formalisations like the following:

(17) a. $\text{Composite}(x) \equiv_x \exists y(\text{ProperPart}(y,x))$
For something to be composite is for there to be a proper part of it.

b. $\text{Parallel}(x, y) \equiv_{x,y} (\text{Line}(x) \wedge \text{Line}(y) \wedge \text{Coplanar}(x, y) \wedge \neg\text{Intersect}(x, y))$
For an object x to be parallel to an object y is for x to be a line and y to be a line and x to be coplanar with y and x not to intersect y.

c. $\exists y[\text{Country}(y) \wedge (\text{German}(x) \equiv_x \text{From}(x, y))]$
There is a country y such that for something to be German is for it to be from y.

This approach might seem to fit 'To be F as to be G' worse than 'For something to be F is for it to be G'. But the case is not so clear. Orthodox syntax posits an unpronounced pronoun-like constituent 'PRO' in such sentences: really we are dealing with 'PRO to be F is PRO to be G'. Whether 'PRO to be F' and 'PRO to be G' combine with local lambda operators before encountering 'is', or whether the bind-

[19]The question what mechanisms underlie the behaviour of indefinites like 'something' in 'For someone to kill someone is for them to cause them to die' is a controversial one in linguistics. In dynamic semantics, it is standard to think of indefinites as variable-like rather than quantifier-like: the quantificational meanings of sentences like 'Something is in this box' arise through a. An alternative approach would be to think of indefinites as still being quantifier-like, but as having covert restrictors that may include variable: 'For someone *identical to* x_1 to kill someone *identical to* x_2 is for them$_1$ to cause them$_2$ to die' (Elbourne 2005). Further options arise if we want indefinites to be predicate-like, as in Graff 2001. I hope these debates can be bypassed for present purposes.

[20]Rayo (2013) also uses this notation, with what I take to be the same interpretation.

ing occurs at the level of the whole sentence, involves hard questions of syntax and semantics that I will not try to resolve here. [21, 22]

Even if we were convinced that the sentential approach provides a more adequate formalisation, we can still make sense of the predicate formalism by understanding $R \equiv S$ as shorthand for $R(v_1, \ldots, v_n) \equiv_{v_1,\ldots,v_n} S(v_1, \ldots, v_n)$ (where R and S are n-ary predicates, and $v_1, \ldots, v_n$ are arbitrarily chosen, distinct variables not free in R or S). Since both uses of $\equiv$ were explained by reference to identifications in English, this translation should be good insofar as 'To be F is to be G' and 'For something to be F is for it to be G' are interchangeable in English. Note however that many sentential identifications are not the translations of any predicate identifications, since they are not syntactically of the form $R(v_1, \ldots, v_n) \equiv_{v_1,\ldots,v_n} S(v_1, \ldots, v_n)$ for any R, S, v_1,..., and v_n. Because of this, any proposal for a translation in the opposite direction—mapping every sentential identification to a predicate identification—will be more controversial. In particular, the obvious suggestion to translate $\varphi \equiv_{v_1,\ldots,v_n} \psi$ as $(\lambda v_1 \ldots v_n.\varphi) \equiv (\lambda v_1 \ldots v_n.\psi)$ leads, in combination with the translation in the other direction to the conclusion that $\varphi \equiv_{v_1,\ldots,v_n} \psi$ is always equivalent to $(\lambda v_1 \ldots v_n.\varphi)(u_1, \ldots, u_n) \equiv_{u_1 \ldots u_n} (\lambda v_1 \ldots v_n.\psi)(u_1, \ldots, u_n)$. If we take 'equivalent' in the sense of '$\equiv$', this is controversial for reasons we will consider in §5. Indeed, even the claim of material equivalence is controversial for certain φ and ψ. For example, let φ abbreviate some closed sentence, say 'snow is white'. Someone might accept the truth of $(\lambda x.(\lambda x.\varphi)(x)) \equiv (\lambda x.\varphi)$—'to be such that one is such that snow is white is to be snow is white'—while rejecting $(\lambda x.\varphi)(x) \equiv_x \varphi$—'for something to be such that snow is white is for snow to be white'—on the grounds that not the case (for example) that for Obama to be such that snow is white is for snow to be white, given that Obama's being such that snow is white is about Obama in a way that snow's being white is not.

Let us now consider how to formalise the basic logical principles discussed informally in §2. In the predicate approach, we can use analogues of the usual identity

[21] The claim that PRO in in infinitival clauses must (at least in certain environments) be semantically bound by a local lambda abstractor features in a popular explanation of the the 'de se' character of expressions like 'She expects to φ' and 'She wants to φ' (Chierchia 1989).

[22] If there turns out to be a structural difference between the arguments of 'is' in 'For something to be F is for it to be G' and in 'To be F is to be G', this will raise the possibility that there is some subtle semantic difference between the forms, so that 'For it to be the case that (for something to be F is for it to be G) is for it to be the case that (to be F is to be G)' can be false—perhaps there are even cases where the two forms differ in truth value. This would require some revision in the present paper, which treats the forms as interchangeable. If evidence against this assumption emerged, I would need to focus more narrowly on 'To be F is to be G'; however I would insist that this kind of claim can intelligibly be generalised to polyadic predicates whether or not the generalisation is expressible in natural languages. (Thanks to Mark Schroeder for helpful discussion here.)

axioms:

Ref	$F \equiv F$
LL	$(F \equiv G) \rightarrow (\varphi \rightarrow [G/_{?}F]\varphi)$

where $[G/_{?}F]\varphi$ is a sentence like φ except that one or more occurrences of F are replaced by occurrences of G, in such a way that no variables free in $F \equiv G$ are bound in φ or $[G/_{?}F]\varphi$. I will adopt the convention that universal generalisations of instances of schemas count as instances; so for example, $\forall x((\lambda y.Rxy) \equiv (\lambda y.Rxy))$ is an instance of Ref. Symmetry and transitivity follow immediately using the instances $(F \equiv G) \rightarrow ((F \equiv F) \rightarrow (G \equiv F))$ and $(G \equiv F) \rightarrow ((G \equiv H) \rightarrow (F \equiv H))$ of LL; the truth-preservation principle $(F \equiv G) \rightarrow (Fx \leftrightarrow Gx)$ follows from the instance $(F \equiv G) \rightarrow ((Fx \leftrightarrow Fx) \rightarrow (Fx \leftrightarrow Gx))$.

LL will of course have to be restricted if our language contains contexts that are "opaque" in the way that propositional attitude ascriptions seem *prima facie* to be. However, given the general perspective on the exceptional character of these contexts defended in §2, we may reasonably avoid the need to constantly make exceptions to generalisations like LL by simply stipulating that our formal languages will not contain any expressions that generate opaque contexts.

The analogous work in the sentential approach can be done by the following axioms:

Ref_{s}	$\varphi \equiv_{v_1 \ldots v_n} \varphi$
Alphabetic Variation	$(\varphi \equiv_{v_1 \ldots v_n} \psi) \rightarrow ([u_i/v_i]\varphi \equiv_{u_1 \ldots u_n} [u_i/v_i]\psi)$
LL_{s}	$(\varphi \equiv_{v_1 \ldots v_n} \psi) \rightarrow (\chi \rightarrow [\varphi/_{?}\psi]\chi)$

where $[u_i/v_i]\varphi$ is the result of replacing each free occurrence of v_i in φ with a free occurrence of u_i (re-lettering bound variables if necessary), and $[\varphi/_{?}\psi]\chi$ is a sentence that results from χ by replacing one or more instances of the open sentence ψ with an instance of the open sentence φ, in such a way that no variable that is free in $\varphi \equiv_{v_1 \ldots v_n} \psi$ is bound in χ or $[\varphi/_{?}\psi]\chi$. Note that in this approach, *Alphabetic Variation* plays a crucial role in the derivation of many implications, such as $Fx \equiv_x Gx, Hx \equiv_x Ix \vdash Fx \wedge Hy \equiv_{x,y} Gx \wedge Iy$. By contrast, in the predicate approach, while we surely will want $(\lambda v_1 \ldots v_n.\varphi) \equiv (\lambda u_1 \ldots u_n.[u_i/v_i]\varphi)$ to be valid, it is not needed for the closest analogue of the above inference, namely $F \equiv G, H \equiv I \vdash (\lambda xy.Fx \wedge Hy) \equiv (\lambda xy.Gx \wedge Iy)$.

4 Higher-orderese

My refusal to take 'The property of being F is the property of being G' as more than a heuristic gloss on 'To be F is to be G' is reminiscent of a certain attitude towards the formalism of higher-order logic defended by, amongst others, Prior (1971) and (Williamson 2003). This position holds that sentences like '$\exists F(F(\text{Frege}) \wedge F(\text{Church}))$' and '$\exists F(\Box \forall x F(x))$' are intelligible and true, but only at best heuristically or misleadingly glossed by sentences using standard English quantified noun phrases, like 'Some property is instantiated by both Frege and Church', and 'Some property is necessarily instantiated by everything'.[23] Perhaps there just aren't any natural-language sentences strictly synonymous with these formal sentences: if so, we will have to learn the language of higher-order logic by the same "direct method" we use when learning foreign languages by immersion (Williamson 2003, p. 459). I will not be arguing for this kind of embrace of higher-order quantification here. But I do want to discuss several ideas about identification which can be more cleanly articulated using higher-order resources.

The formal higher-order language I have in mind is that of simple relational type theory with lambda abstraction. It works as follows (see Appendix A1 for a more rigorous presentation). Every syntactic unit is a *term* having a particular *type*, or syntactic category. Types are defined as follows: the letter e is a type ("the type of objects"); for any $n \geq 0$ and types $\tau_1, \ldots, \tau_n$, the n-tuple $\langle \tau_1, \ldots, \tau_n \rangle$ is a type; nothing else is a type. Formulae (open or closed sentences) are terms of type $\langle\rangle$ ("propositional type"). n-place predicates of the familiar sort—expressions that that can combine with n first-order variables to form sentences—are terms of type $\langle e, \ldots, e \rangle$. There are variables of all types; we may indicate the type of a variable by means of a superscript on its first occurrence. For any terms $A, B_1, \ldots, B_n$ of types $\langle \tau_1, \ldots, \tau_n \rangle, \tau_1, \ldots, \tau_n$ respectively, $A(B_1, \ldots, B_n)$ is a formula. For any distinct variables $v_1, \ldots, v_n$ of types $\tau_1, \ldots, \tau_n$ and formula φ, $(\lambda v_1 \ldots v_n . \varphi)$ is a term of type $\langle \tau_1, \ldots, \tau_n \rangle$. The logical constants are $\neg$ (of type $\langle\langle\rangle\rangle$); $\vee$ and $\wedge$ (of type $\langle\langle\rangle, \langle\rangle\rangle$); and for every type τ, $\forall_\tau$ and $\exists_\tau$ (of type $\langle\langle\tau\rangle\rangle$), and $\equiv_\tau$ (of type $\langle \tau, \tau \rangle$).

Some abbreviations: we write $\forall x^\tau(\varphi)$ and $\exists x^\tau(\varphi)$ instead of $\forall_\tau((\lambda x^\tau . \varphi))$ and $\exists_\tau((\lambda x^\tau . \varphi))$. We freely use infix notation, e.g. writing $\varphi \wedge \psi$ instead of $\wedge(\varphi, \psi)$. $\rightarrow$ abbreviates $\lambda p^{\langle\rangle} q^{\langle\rangle} . \neg p \vee q$, and $\leftrightarrow$ abbreviates $\lambda p^{\langle\rangle} q^{\langle\rangle} . (p \rightarrow q) \wedge (q \rightarrow p)$. We may omit parentheses and type annotations when they can be reconstructed unambiguously.

[23] The point has nothing special to do with the word 'property': the reasons for not embracing these English sentences as unproblematic translations of the higher-order sentences also applies to corresponding sentences using 'concept', 'condition', etc.

By regimenting identification using a higher-order predicate $\equiv_\tau$, I am opting for the "predicate" formalism from the previous section. The convenience and elegance of having lambda-abstraction do all the work of variable-binding is just too great to pass up. This means that the language will lack the ability to neutrally regiment sentential identifications that are not obviously materially equivalent to any particular predicate identification. Fortunately, there are plenty of deep and controversial questions of identification that *do* have obvious (material) equivalents in our higher-order language, so we need not be too sad to postpone the rest for another occasion.

One question that can naturally be raised once higher-order quantifiers are in the language is the question whether identifications are materially equivalent to claims of higher-order indiscriminability. Schematically:

Indiscriminability $\quad \forall x^\tau \forall y^\tau ((x \equiv_\tau y) \leftrightarrow \forall z^{\langle\tau\rangle}(z(x) \leftrightarrow z(y)))$

where τ may be any type.

Indiscriminability can be derived from Ref and LL given standard classical logic. $(x \equiv_\tau y) \rightarrow ((z^{\langle\tau\rangle}(x) \leftrightarrow z(x)) \rightarrow (z^{\langle\tau\rangle}(x) \rightarrow z(y)))$ is an instance of LL, and classically implies $(x \equiv_\tau y) \rightarrow (z^{\langle\tau\rangle}(x) \rightarrow z(y))$; universal generalisation on this formula gives the right to left direction of *Indiscriminability*. In the other direction, suppose that $\forall z^{\langle\tau\rangle}(z(x) \leftrightarrow z(y))$; then by universal instantiation, $(\lambda u.x \equiv_\tau u)(x) \leftrightarrow (\lambda u.x \equiv_\tau u)(y)$, which implies $x \equiv_\tau x \leftrightarrow x \equiv_\tau y$ (by "extensional β-conversion", see next section), so $x \equiv_\tau y$ by Ref.

Some will reject *Indiscriminability* on the basis of an argument like this:

1. Hank believes all vixens are vixens and does not believe all vixens are female foxes.

2. So $(\lambda p^{\langle\rangle}$.Hank believes $p)$(all vixens are vixens) $\wedge \neg(\lambda p^{\langle\rangle}$.Hank believes $p)$(all vixens are female foxes)].

3. So $\exists z^{\langle\langle\rangle\rangle}(z$(all vixens are female foxes)$\wedge\neg z$(all vixens are vixens)).

4. All vixens are female foxes $\equiv$ all vixens are vixens

5. So, $\exists p^{\langle\rangle}\exists q^{\langle\rangle}(p \equiv q \wedge \exists z^{\langle\langle\rangle\rangle}(z(p) \wedge \neg z(q)))$.

Accepting the conclusion of this argument means holding that higher-order quantifiers themselves create an opaque context, so that we cannot add such quantifiers to the language without disrupting the ban on opacity required for all instances of LL to be true.

The one way of resisting this argument that I want to firmly rule out (at least considered as a general strategy) is that of denying 4: it is crucial to the philosophically important use of 'To be F is to be G' that such claims cannot be refuted so easily. But there are other ways of resisting the argument that are much more promising. One might reject 1, regarding it as literally false, or at least literally false on every uniform interpretation.[24] One might—currently my favourite option—reject the inference from 1 to 2, treating propositional attitude ascriptions as creating a special syntactic environment that makes for exceptions to generally valid rules like extensional beta-conversion (see §5), similar in this respect to "mixed quotation". Or, most radically, one might reject the classical rule of existential generalisation, and with it one or both of the inferences from 2 to 3 or from 3 and 4 to 5 (see Bacon and J. S. Russell MS).[25]

I think that we should accept *Indiscriminability*, and resist the argument against it in one of the above ways. However, *Indiscriminability* will not play a major role in what follows—indeed, most of the claims we will be considering will be intelligible even to those who insist that higher order quantifiers are meaningless. (Identification would give complex higher-order predicates a *raison d'être* even if there weren't any quantifiers around for them to serve as arguments for.) When I do on occasion use higher-order quantifiers in ways that assume *Indiscriminability*, those who reject *Indiscriminability* because of opaque contexts will in most cases be able to make sense of what's going on by interpreting the quantifiers as restricted to the domain of the transparent.[26]

It is worth mentioning that if we do accept *Indiscriminability*, there is a strong

[24]For the ideology of uniform interpretations and its relevance to arguments like this one, see Dorr 2014c. There I argue that when N and M are ordinary proper names, 'If $N = M$, everyone who believes that $\ldots N \ldots$ believes that $\ldots M \ldots$' is true on every uniform interpretation. One could imagine saying the same thing about all identifications that I say about identities between ordinary names. However, this generalisation is harder to defend than the view in the paper. It is fairly easy to imagine situations where 'Lois believes that Clark flies' could be used to assert a truth even though Lois would accept the sentence 'Clark doesn't fly, although Superman does'; it is much harder to imagine situations where 'Hank believes that there are female foxes' could be used to assert a truth even though Hank would accept the sentence 'There are no female foxes, although there are vixens—vixens are sexless Martian spy-robots, not female foxes'. The difference is that ordinary proper names (excluding compound names like 'Oxford University') lack relevant syntactic structure, whereas expressions in other categories can differ structurally in a way that seems to make for systematic, conventionalised differences in the communicative possibilities for speech and attitude reports involving them.

[25]Another possible view is that higher-order quantification is ambiguous, so that *Indiscriminability* has both true and false readings.

[26]This applies in particular to the initial universal quantifiers which, according to our convention concerning schemas, may be added to bind any otherwise free variables that appear in an instance of a schema.

case to be made for strengthening it to an identification:[27]

The Identity Identity $\equiv_\tau \equiv_{\langle\tau,\tau\rangle} \lambda x^\tau y^\tau . \forall z^{\langle\tau\rangle}(z(x) \leftrightarrow z(y))$

In words: identification *is* higher-order indiscernibility. *The Identity Identity* provides a powerful explanation of the truth of *Indiscriminability* and of the instances of Ref and LL; it allows the entire logic of $\equiv$ to be subsumed under that of quantification, which we need in any case.[28] However, once we have *Indiscriminability*, the question whether *The Identity Identity* is true will only be relevant to questions that turn on embedded identifications, which will not be central in what follows.

5 Beta-equivalence

An object x is such that it is rectangular and it is equilateral if and only if x is rectangular and x is equilateral. More generally: x is such that …it… iff …x…. In a language with lambda abstracts, this pattern can be captured by the following schema:

Extensional β-equivalence $(\lambda v_1 \ldots v_n . \varphi)(A_1, \ldots, A_n) \leftrightarrow [A_i/v_i]\varphi$

where $v_1, \ldots, v_n$ are any distinct variables, $A_1, \ldots, A_n$ are any terms (not necessarily all distinct), and $[A_i/v_i]\varphi$ is a sentence that results from replacing each free occurrence of v_1 in φ with an occurrence of A_1, and each occurrence of v_2 with an occurrence of A_2, and so on, replacing bound variables in such a way that no free variables in any A_i become bound. *Extensional β-equivalence* should be uncontroversial given that the language doesn't contain opaque contexts.[29]

A much more controversial question is whether this claim of coextensiveness can be strengthened to an identification:

Immediate β-equivalence $(\lambda v_1 \ldots v_n . \varphi)(A_1, \ldots, A_n) \equiv_{\langle\rangle} [A_i/v_i]\varphi$

[27] I have stolen this name from Bacon and J. S. Russell MS.

[28] *The Identity Identity* is also convenient for metalogical purposes, since it lets us work with a shorter list of logical constants and a simpler definition of a model: for this reason I take it for granted in the Appendix. But this is only a convenience: reintroducing $\equiv$ as a primitive in the model theory would be straightforward.

[29] *Extensional β-equivalence* becomes controversial once we have expressions like 'believes' in the language. For example, it is controversial whether $(\lambda f$.Hank believes $\forall y$(Vixen(y) $\rightarrow$ $f(y)))(\lambda z$. Female(z) $\wedge$ Fox(z)) $\rightarrow$ Hank believes $\forall y$(Vixen(y) $\rightarrow$ $(\lambda z$. Female(z) $\wedge$ Fox(z))(y)): this conditional will be false on views where the syntactic structure of the argument of 'believes' makes a truth-conditional difference.

For example: for Nelly to be such that she is female and she is a fox is for it to be the case that Nelly is female and Nelly is a fox.

Those who accept *Immediate β-equivalence* will presumably also want to accept other identifications that result from performing the same kind of substitution in an embedded context, for example the following:

$$(\lambda y.((\lambda x.\boldsymbol{R}(x, y))(z))) \equiv_{\langle e \rangle} (\lambda y.\boldsymbol{R}(z, y))$$
$$\boldsymbol{R}(x, (\lambda y.\boldsymbol{S}(y, (\lambda z.Fz \wedge Gz)(x)))) \equiv_{\langle\rangle} \boldsymbol{R}(x, (\lambda y.\boldsymbol{S}(y, Fx \wedge Gx)))$$

So, the picture that strengthens *Extensional β-equivalence* to an identification would seem to go along with the following more general schema:[30]

β-conversion: $\varphi \leftrightarrow \varphi^*$, where φ^* is derived from φ by replacing some constituent of the form $(\lambda v_1 \ldots v_n.\psi)(A_1, \ldots, A_n)$ with $[A_i/v_i]\psi$.

When one term can be derived from another by a replacement of this sort, we say that the latter *one-step β-reduces* to the former, and that the two are *one-step β-equivalent*. Two terms are β-equivalent *simpliciter* when they can be connected by a sequence of terms each of which is one-step β-equivalent to its predecessor. Given the symmetry and transitivity of $\equiv$, *β-conversion* obviously implies the more general principle that $\varphi \leftrightarrow \varphi^*$ is true whenever φ and φ^* are β-equivalent. And given Ref, we also get the seemingly stronger principle that $\boldsymbol{A} \equiv_\tau \boldsymbol{B}$ is true whenever $\boldsymbol{A}$ and $\boldsymbol{B}$ are β-equivalent terms of type τ: for in this case, $(\boldsymbol{A} \equiv_\tau \boldsymbol{A}) \leftrightarrow (\boldsymbol{A} \equiv_\tau \boldsymbol{B})$ is an instance of *β-conversion*, and implies $(\boldsymbol{A} \equiv_\tau \boldsymbol{B})$ given Ref.[31] This explains how we get back from *β-conversion* to *Immediate β-equivalence*.

The question whether to endorse *β-conversion* is a crucial choice point for theorising about the logic of identifications. My view is that *β-conversion* is *close* to being valid. To be precise: *β-conversion* holds whenever the substituted term $(\lambda v_1 \ldots v_n.\psi)(A_1, \ldots, A_n)$ is such that each of the variables $v_1, \ldots, v_n$ have at least one free occurrence in ψ—call these instances of *nonvacuous β-conversion*. In this

[30] Interestingly, although there is no obvious route to *β-conversion* from *Immediate β-equivalence*, if we were taking the sentential identification connective as primitive the schema $(\lambda v_1 \ldots v_n.\varphi)(A_1, \ldots, A_n) \equiv_{u_1 \ldots u_m} [A_i/v_i]\varphi$ could do all the work of *β-conversion*.

[31] If we are taking $\equiv$ as primitive, we could also take the schema whose instances are $\boldsymbol{A} \equiv_\tau \boldsymbol{B}$ when $\boldsymbol{A}$ and $\boldsymbol{B}$ are β-equivalent as our basic axiom, and derive *β-conversion* from $\varphi \equiv \varphi^*$ and $\varphi \leftrightarrow \varphi$ using LL. However, if we are using *The Identity Identity* to avoid taking $\equiv$ as a primitive, we will need to take *β-conversion* as the more basic axiom, since we need to rely on it to get from $\varphi \equiv \varphi^*$, i.e. $(\lambda pq.\forall z(z(p) \leftrightarrow z(q)))(\varphi, \varphi^*)$, to $\varphi \leftrightarrow \varphi^*$, via $\forall z(z(\varphi) \leftrightarrow z(\psi))$ (one application of *β-conversion*), then $(\lambda p.p)(\varphi) \leftrightarrow (\lambda p.p)(\varphi^*)$ by universal instantiation, and finally $\varphi \leftrightarrow \varphi^*$ (two more applications of *β-conversion*).

section and the next, I will attempt to make this plausible, though what I have to say will fall far short of being a knock-down argument. In the present section, I will survey four possible strategies for arguing against instances of *β-conversion*, arguing that none of them are compelling. In §6, I will sketch what seems to me to be the most principled and systematic view on which which β-conversion fails, give some arguments against it, and say some positive things in favour of (non-vacuous) β-conversion.

The first strategy for arguing against *β-conversion* is based on familiar arguments for thinking that the *proposition* that $(\lambda v_1 \ldots v_n.\varphi)(t_1, \ldots, t_n)$ is (in many cases) distinct from the proposition that $[t_i/v_i]\varphi$. For example, Salmon (2010) considers a case where someone thinks they are seeing photos of two yachts when in fact they are seeing two photos of a single yacht *a*, and sincerely utters 'This yacht is longer than that one is'. According to Salmon, the speaker in this case believes the proposition that *a* is larger than *a* is, but does not believe the proposition that ($\lambda x.x$ is larger than x is)(*a*).[32] Similarly, the Babylonians believed that Venus was a planet visible in the morning and Venus was a planet visible in the evening, and communicated this belief to one another by uttering the Babylonian equivalent of 'Phosphorus is visible in the morning and Hesperus is visible in the evening', but they did not believe that ($\lambda x.x$ is a planet visible in the morning and x is a planet visible in the evening)(Venus).

Considerations of this kind are very important to the metaphysics of propositions (understood as "objects of the attitudes"); but they are terrible as an argument against instances of *β-conversion*. This argument would be no better than the argument that since it is possible for someone to believe that there are vixens without believing that there are female foxes or to want to be a vixen without wanting to be a female fox, it is not the case that for there to be vixens is for there to be female foxes or that to be a vixen is to be a female fox. Those (including Salmon and Soames) who think that the proposition that there are vixens is distinct from the proposition that there are female foxes on the basis that it is possible to believe the former without believing the latter should regard 'the proposition that…' as an opaque context, rejecting the inference from 'For it to be the case that φ is for it to be the case that ψ' to 'The proposition that φ = the proposition that ψ'. If they insist on understanding the identification in terms of the identity of entities of some sort, they should think of these entities not as *propositions* (the "objects of the attitudes"), but as "states of affairs"—entities which stand to the number 0 as properties stand to 1 and binary relations stand to 2.[33]

[32] See also Salmon 1986b and Soames 1987a.

[33] It is unfortunate that philosophy has no accepted label for such entities. 'Proposition' is hostage

The second argumentative strategy is more promising, since it turns on environments that have nothing obvious to do with propositional attitudes or "intentionality". Gideon Rosen (2010) and Kit Fine (2012) suggest certain general principles about *grounding* which, if true, would provide a widely-applicable strategy for arguing against instances of *β-conversion*. Rosen puts the point in terms of an ontology of facts: he maintains that in general, the fact that $[a/x]\varphi$ grounds the fact that $(\lambda x.\varphi)(a)$. Since no fact grounds itself, this entails that the facts in question are distinct. Fine thinks of grounding claims as involving a sentential operator which (at least in the straightforward case where it connects two sentences) we can pronounce 'because'. So for him, the key claim is that whenever $(\lambda x.\varphi)(a)$, $(\lambda x.\varphi)(a)$ because $[a/x]\varphi$, although it is never true that $(\lambda x.\varphi)(a)$ because $(\lambda x.\varphi)(a)$.[34] These claims certainly *sound* like they should entail that it is not true, in our target sense, that for it to be the case that $(\lambda x.\varphi)(a)$ is for it to be the case that $[a/x]\varphi$. Rosen and Fine are at the forefront of a movement to give questions expressed in grounding-theoretic terms a central role in metaphysics, not merely as tools for investigating some other questions (in the way that, e.g., questions about conceptual analysis might be), but as topics of investigation for their own sake. At least insofar as one is convinced by the picture I presented in §2—according to which the "subject matter of metaphysics" is conceived of as being about the world as opposed to our representations of it, and true identifications license substitution within claims of this sort—one will not want to resist the grounding-theoretic argument against *β-conversion* at its last step.[35]

However, so long as we conceive of grounding as a worldly matter, I see no good reason for accepting the premise that $[a/x]\varphi$ *ever* grounds $(\lambda x.\varphi)(a)$. When we are

to the propositional attitudes; 'fact' is ruled out since there cannot be a fact that φ unless φ; and some treat 'state of affairs' as like 'fact' in this respect. I am tempted by 'factoid'.

[34]Fine suggests just one possible exception to the generalisation that $[a/x]\varphi$ strictly grounds $(\lambda x.\varphi)(a)$, namely when φ is a predication $F(x)$ where x is not free in F, so that $(\lambda x.\varphi)(a) \equiv [a/x]\varphi$ is an instance of η-conversion (see below) as well as *β-conversion*.

[35]Interestingly, however, there are indications that Fine and some other grounding-enthusiasts are not thinking along these lines. Fine is open to a view on which propositions are individuated too coarsely to respect grounding-theoretic distinctions: 'the truth of "A, B < C" might be taken to depend not merely upon the propositions expressed by "A", "B" and "C" but also upon how these propositions are expressed'. This suggests a picture where grounding-theoretic claims are in some important sense about our representations, rather than simply about how things are in the world. Correia (2010) takes seriously the idea that 'grounding' admits a "conceptual" interpretation that works like this, as well as a "worldly" interpretation, and interprets Fine and Rosen as concerned with the "conceptual" notion. If he is right, the present argument against *β-conversion* from putative failures of substitutivity in grounding claims would have no more force than the previously considered argument from failures of substitutivity in attitude reports. However, I have little sense of what the conceptual interpretation of grounding claims is supposed to be, or why anyone would regard such claims as having a distinctive interest for metaphysics.

first being introduced to the language of grounding, we will be tempted to deploy it quite promiscuously. For example, we will be tempted to claim that the fact that Nelly is a vixen is grounded by the fact that Nelly is a female fox. After all, 'Nelly is a vixen because Nelly is a female fox' certainly sounds true, and there are no obvious tests for distinguishing the 'because' here from the 'because' of grounding. But this temptation must certainly be resisted, as discussed in §2. Given that to be a vixen is to be a female fox, it certainly follows that for Nelly to be a vixen is for Nelly to be a female fox, and hence that the fact that Nelly is a female fox does not ground the fact that Nelly is a vixen, since it does not ground itself. Of course Fine and Rosen need not dispute this, since they can claim that this one fact is distinct from the fact that Nelly is female and Nelly is a fox. But once we have realised that we need to be careful in going from intuitive 'because' claims to grounding claims, it is hard to see any principled grounds for resisting the temptation in the case of the fact that Nelly is a female fox while yielding to it in the case of the fact that Nelly is female and Nelly is a fox.

There is a third influential strategy for arguing against instances of *β-conversion* whose application is more limited. It has two premises. The first is contingentism, the view that it is metaphysically possible for there to be something such that it is not metaphysically necessary that it exists (in the sense of being identical to something):

Contingentism $\quad \Diamond\exists x\neg\Box\exists y(y = x)$

The second premise is what Williamson (2013) calls "the being constraint", which can be stated schematically as follows:

(BC) $\quad \Box\forall x\Box(Fx \rightarrow \exists y(y = x))$

Here F stands for any predicate.[36] The combination of contingentism, (BC), and classical modal logic requires the failure of *β-conversion*. For example, we cannot have the following instance of *β-conversion*:

$$\Box\forall x\Box((\lambda z.\neg\exists y(y = z))(x) \rightarrow \exists y(y = x)) \leftrightarrow \Box\forall x\Box(\neg\exists y(y = x) \rightarrow \exists y(y = x))$$

since the formula on the left is an instance of (BC), while the one on the right is equivalent in classical modal logic to $\Box\forall x\Box\exists y(y = x)$ and thus inconsistent with contingentism.

Similarly, contingentists who endorse (BC) will have to reject the following in-

[36]The variable x could occur free in F, but the argument does not depend on instances of this sort.

stance of *β-conversion* for any φ:

$$\Box\forall x\Box((\lambda x.\varphi \vee \neg\varphi)(x)) \leftrightarrow \Box\forall x\Box(\varphi \vee \neg\varphi)$$

since the right hand side is a theorem of classical modal logic, whereas the left hand side implies $\Box\forall x\Box(\exists y(y = x))$ given (BC).[37]

My main complaint about this strategy is that the motivation for the Being Constraint seems weak given contingentism. The central contrast this package draws between subject-predicate sentences and other kinds of sentences just doesn't seem to be borne out when we actually look at natural languages.

A naïve way to make this argument would be as follows. 'Obama doesn't exist' is a subject-predicate sentence: it results from combining the name 'Obama' with the complex predicate 'doesn't exist'. So if the Being Constraint were correct, it would have to be necessary that if Obama doesn't exist, Obama exists, in which case it would be necessary that Obama exists, which is something no contingentist will grant. The reason this is naïve is that surface syntax may be misleading. Sentences where 'not' occurs inside the verb phrase sometimes have readings in which the subject really occurs within the scope of the negation, in the sense of 'scope' that matters for semantics:

(18) a. Everyone hasn't yet had a chance to read the minutes.

b. All that glitters is not gold.

(18a) and (18b) are structurally ambiguous: they have weak readings equivalent to 'It is not the case that everyone has already had a chance to read the minutes' and 'It is not the case that all that glitters is gold' as well as the strong readings equivalent to 'No-one has yet had a chance to read the minutes' and 'Nothing that glitters is gold'. Proponents of the Being Constraint can therefore respond to the naïve argument by saying that 'It is possible that Obama doesn't exist' is similarly ambiguous, and is true only on the reading where the negation takes scope over 'Obama'.[38]

The problem with this response is that when 'doesn't exist' has a quantified subject, the reading where negation takes scope over the subject is often much too weak. If contingentism is true, sentences like the following are surely true on *both* readings:

[37] Stalnaker (1994) develops a quantified modal logic in which (BC) is upheld and β-conversion fails. Many other authors, most influentially Plantinga (1983), have defended a structurally similar package in the context of a theory of properties, namely that property-exemplification entails existence, so that the intersubstitutability of '*a* has the property of being an *x* such that φ' and '$[a/x]\varphi$' must be restricted in modal contexts.

[38] See Plantinga 1983, p. 13.

(19) a. It could have happened that both of us didn't exist.

b. If the second world war had not been fought, everyone who was actually born since then wouldn't have existed.

These sentences are clearly ambiguous in the same way as (18a) and (18b). But given standard contingentist views about the extent of contingent existence, it seems wrongheaded to insist that they are true only on their weak readings, where they are equivalent respectively to (20a) and (20b):

(20) a. It could have happened that it was not the case that both of us existed.

b. If the second world war had not been fought, it would not have been the case that everyone who was actually born since then existed.

The most prominent reading of (19a) is the strong one on which, as uttered by A to B, it is true only if it could have happened that neither A nor B existed; assuming contingentism, it should be true on that reading. Similarly for (19b). But given (BC), a sentence of the form 'DP VP' will be modally equivalent to 'DP exist(s) and VP', so long as nothing in the VP takes scope over the DP. So the readings of (19a) and (19b) where 'not' takes scope below the NP will be equivalent to (21a) and (21b):

(21) a. It could have happened that both of us existed and didn't exist.

b. If the second world war had not been fought, everyone who was actually born since then would and wouldn't have existed.

And this looks bad, since (21a) and (21b) seem clearly false.[39]

(A further argument against the combination of contingentism and the Being Constraint, due to Fritz and J. Goodman (forthcoming, n. 14), turns on higher-order quantification. In a higher-order setting, there is an natural analogue of contingentism involving quantification into sentence position:

$$\Diamond\exists p\Diamond(\neg\exists q(q \equiv_{\langle\rangle} p))$$

There is some pressure on contingentists to endorse such "higher-order contingentism": see Williamson 2013, ch. 6. Similarly, there is some pressure on those who

[39]Contingentist proponents of (BC) might reply at this point that (21a) and (21b) are in fact true, for the same reason that 'All unicorns both are and are not unicorns' is true. There are many problems with this move, but perhaps the worst one is that it does not generalise to examples using other quantifiers. 'If that had happened, most of us wouldn't have existed' has a reading where, assuming contingentism, it is true if 'us' refers to A, B, and C, and if the relevant thing had happened, A and B would never have been born but C still would have. But 'If that happened, most of us both would and wouldn't have existed' isn't true in this circumstance.

endorse (BC) to accept its higher-order analogue:

$(\mathrm{BC}_{\langle\rangle})$ $\qquad\qquad \Box\forall p\Box(Op \rightarrow \exists q(q \equiv_{\langle\rangle} p))$

Here O is schematic for a sentential operator—a term of type $\langle\langle\rangle\rangle$. But since $\neg$ and $\Diamond$ are sentential operators, $(\mathrm{BC}_{\langle\rangle})$ implies propositional necessitism:

1.	$\Box\forall p\Box(\Diamond p \vee \neg p)$	classical modal logic (KT)
2.	$\Box\forall p\Box(\Diamond p \rightarrow \exists q(q \equiv p))$	instance of $\mathrm{BC}_{\langle\rangle}$
3.	$\Box\forall p\Box(\neg p \rightarrow \exists q(q \equiv p))$	instance of $\mathrm{BC}_{\langle\rangle}$
5.	$\Box\forall p\Box(\exists q(q \equiv p))$	1-3, classic

Higher-order contingentists must thus reject $(\mathrm{BC}_{\langle\rangle})$, making (BC) look ill-motivated.)

A fourth strategy for arguing against *β-conversion* targets only the *vacuous* instances. For example, one can appeal to the concept of *aboutness*, arguing against the claim that for Obama to be such that snow is white is for snow to be white on the grounds that snow being white is not about Obama, whereas Obama being F is about Obama (for any F). 'About' is a bit too vague for this argument to carry much weight by itself.[40] But it does help to undermine the positive case for full *β-conversion* based on examples, by drawing our attention to the possibility of a weaker generalisation that fits the examples equally well. Sections 8 and 9 will introduce some other considerations that count against vacuous β-conversion but not against nonvacuous β-conversion.

For formal purposes, if we reject vacuous β-conversion, it is convenient to work with a so-called *λI-language*, where '$\lambda v_1 \ldots v_n.\varphi$' is not well-formed unless all of $v_1, \ldots, v_n$ have free occurrences in φ. (For details see Appendix A1.) This restriction lets us use the usual β-conversion rule rather than constantly having to make exceptions for the vacuous case. The more common form of language ('λK-language'), where vacuous binding is allowed, can be translated into the λI-language by appending trivial conjuncts or disjuncts to abstracts so that every abstracted variable has a free occurrence: for example, when y is not free in F, $\lambda xy.Fx$ could be translated as $\lambda xy.Fx \wedge y = y$. It might be objected that this is arbitrary. Why not instead choose $\lambda xy.Fx \vee y \neq y$, or $\lambda xy.Fx \wedge \exists z(z = y)$, or $\lambda xy.Fx \wedge (x = y \vee x \neq y)$, for example? This is not an issue if these various options are themselves equivalent (i.e. if the identifications between them are true). But even if the options are not equivalent, opponents of vacuous β-conversion can respond to the worry about arbitrariness by

[40] J. Goodman (MS) shows how the way of thinking about aboutness that underlies this strategy can be developed into a systematic theory.

saying that the original λK-term is vague and has several different λI-terms as admissible precisifications.[41] This seems a strong response: if vacuous β-conversion fails, our use of terms involving vacuous binding is less constrained than our use of λI-terms in a crucial respect, in a way that might be expected to make for vagueness. Thus opponents of vacuous β-conversion may legitimately take λI-languages to be metaphysically more perspicuous than λK-languages, as well as formally more convenient.

From now on, when I say that something follows from something else by β-conversion, I will always mean nonvacuous β-conversion. If we need the vacuous case I will say so explicitly.

In the next section I will further support (nonvacuous) β-conversion by pointing out some problems for the most systematic kind of theory in which it fails.

6 Structure

The "structured picture" involves a kind of thinking familiar from the theory of structured propositions (Cresswell 1985, Lewis 1970, Salmon 1986a, Soames 1987a), which holds that propositions have a kind of structure analogous to that of the sentences that express them.[42] One signature commitment of the theory of structured propositions is that the proposition that a is F = the proposition that b is G only if $a = b$ and the property of being F = the property of being G. (This is often expressed by saying that these propositions are, or can harmlessly be identified with, ordered pairs of objects and properties.) The idea of the structured picture is that identifications work analogously. So we have (the universal closure of) the following axiom:

Atomic Structure $\qquad (f(x) \equiv_{\langle\rangle} g(y)) \rightarrow ((f \equiv_{\langle e\rangle} g) \wedge (x = y))$

Atomic Structure requires widespread failures of β-conversion. For example, β-conversion implies that $(\lambda x.R(x,x))(a) \equiv (\lambda x.R(x,a))(a)$. But *Atomic Structure* allows this only if $(\lambda x.R(x,x)) \equiv (\lambda x.R(x,a))$. For most R, this will be obviously

[41] Different classical-logic-friendly theories of vagueness offer different tools to help one face down arbitrariness-based objections to classical theorems like 'Everyone is either bald or not bald'. The suggestion is that opponents of vacuous β-conversion should respond to the present worry in the same way.

[42] See also Bealer 1982, whose theory of 'concepts' provides an especially close analogue of the structured picture in a first-order setting.

false—typically, $\lambda x.(R(x, x))$ and $\lambda x.(R(x, a))$ are not even coextensive.[43]

Atomic Structure is only a partial articulation of the structured picture, which would not really qualify as "systematic" if it only applied to identifications of the form $F(a) \equiv G(b)$. In a higher-order setting, a principled theory endorsing *Atomic Structure* should surely also endorse the analogous schema involving sentential operators:

Propositional Structure $\qquad x(p) \equiv_{\langle\rangle} y(q) \rightarrow ((x \equiv_{\langle\langle\rangle\rangle} y) \wedge (p \equiv_{\langle\rangle} q))$

This extends the analogy with the theory of structured propositions, which involves the idea that, for example, when one proposition is the negation of another proposition, it is not also the result of applying some *other* operator to that proposition or any other, and it is not the result of applying negation to any other proposition.

Note that even those who reject the intelligibility of higher-order quantifiers or higher-order identifications might accept the following schemas, which capture much of the force of *Atomic Structure* and *Propositional Structure*:

Schematic Atomic Structure $\qquad F(x) \equiv G(y) \rightarrow ((F \equiv G) \wedge (x = y))$

Schematic Propositional Structure

$$(X(\varphi) \equiv Y(\psi)) \rightarrow ((X(\theta) \equiv Y(\theta)) \wedge (\varphi \equiv \psi))$$

I could say more to flesh out the structured picture, considering analogues of *Atomic Structure* for other types, as well as principles like $r(x, y) \not\equiv f(x)$ and $f(x^e) \not\equiv z(p^{\langle\rangle})$ corresponding to the idea that each proposition has a unique structure. But the most important objections to the structured picture require only *Propositional Structure*. I will consider five objections. The first three are in my view much weaker than the last two; I discuss them here because if they worked, they would threaten not just the structured picture but many other "fine-grained" theories, including the

[43]The standard theory of structured propositions suggests the following polyadic generalisation of *Atomic Structure*:

$$(r(a_1, \ldots, a_n) \equiv s(b_1, \ldots, b_n)) \rightarrow (r \equiv s \wedge a_1 = b_1 \wedge \cdots \wedge a_n = b_n)$$

(See, e.g., Audi 2012.) However, this is subject to a further objection which does not impugn *Atomic Structure*: it rules out the possibility that any $r^{\langle e,e\rangle}$ is symmetric in the strong sense that $r \equiv \lambda xy.r(y, x)$. While *Atomic Structure* clearly needs to extend *somehow* to the polyadic case to be worth taking seriously, the desire to allow for symmetry motivates weakening the extension somehow, perhaps to

$$(R(a_1, a_2) \equiv S(b_1, b_2)) \rightarrow (R \equiv S \vee R \equiv (\lambda xy.Syx)) \wedge ((a_1 = b_1 \wedge a_2 = b_2) \vee (a_1 = b_2 \wedge a_2 = b_1))$$

and its natural generalisation to the n-adic case.

theory I will be developing in §8.

The first objection involves apparent counterexamples: cases where $A \equiv B$ just seems true despite the fact that A and B differ in structure in a way disallowed by the structured picture. For example, perhaps it just seems obvious that for London to be north of Paris is for Paris to be south of London. But the English sentences 'London is north of Paris' and 'Paris is south of London' have a binary subject-predicate structure: they result from applying the monadic predicates 'is north of Paris' and 'is south of London' to the names 'London' and 'Paris'. The identification is thus ruled out by *Atomic Structure*. Similarly, it might just look obvious that for it not to be necessary that P is for it to be possible that not P; but this is ruled out by *Propositional Structure*, since it is false that for it to be necessary that P is for it not to be the case that P.

The problem with such direct "appeals to intuition" is that it isn't clear that the judgments in question really involve our target interpretation of 'To be F is to be G', understood literally. We are often pretty permissive in our use of 'To be F is to be G', allowing ourselves the freedom to substitute not only logically equivalent expressions but expressions that we do not even regard as metaphysically necessarily equivalent. For example, when we are doing physical geometry, we might at one time say 'To be a line is to be the shortest path between two points' and at another 'To be a line is to be such that, of any three of one's points, one is between the other two', even if we are not convinced that these conditions are necessarily coextensive. Whatever is going on here, it suggests that we should not be too impatient if proponents of the structured picture respond to putative counterexamples by invoking some kind of "loose talk".

The second objection depends not on case-by-case judgments but on a general principle which is inconsistent with the structured picture, and which is encouraged by some natural ways of thinking about higher order logic. Syntactically, the task of an operator is to combine with a sentence to make another sentence. This makes it natural to think of the semantic value of an operator as a function mapping propositions to propositions. A "function" here is simply a binary relation that is functional—every proposition bears it to some proposition, and no proposition bears it to more than one proposition. Those in the grip of this picture may well find it obvious that quantification into operator position is interchangeable with quantification into binary connective position restricted by a functionality requirement. More generally, quantification into type $\langle\tau\rangle$ is interchangeable with quantification into type $\langle\tau, \langle\rangle\rangle$ restricted by functionality. This is made precise by the following

principle:

Plenitude $\forall x^{\langle\tau,\langle\rangle\rangle}(\text{Functional}_{\langle\tau,\langle\rangle\rangle}(x) \rightarrow \exists z^{\langle\tau\rangle}\forall y^{\tau}(x(y, z(y))))$

where

$$\text{Functional}_{\langle\tau,\langle\rangle\rangle}(x) =_{\text{df}} \forall y^{\tau}\exists p(x(y, p) \wedge \forall q(x(y, q) \rightarrow q \equiv_{\langle\rangle} p))$$

Loosely speaking: for any functional relation x between type-τ things and propositions, there is a corresponding property z of type-τ things, such that for each type-τ thing, the proposition that it has z is the very proposition to which it is related by x.

Plenitude is drastically inconsistent with the structured picture. For some distinct objects a and b and property $f^{\langle e\rangle}$, let R be

$$\lambda y^{e} p^{\langle\rangle}.((y = a) \wedge (p \equiv f(b))) \vee ((y \neq a) \wedge (p \equiv \neg f(b)))$$

Since R is functional, *Plenitude* entails that $\exists z^{\langle e\rangle}\forall y^{e} R(y, z(y))$. Choose a witnessing z and set $y = a$. By Extensional Beta, the fact that $R(a, z(a))$ implies that $((a = a) \wedge (z(a) \equiv f(b))) \vee ((a \neq a) \wedge (z(a) \equiv \neg f(b)))$ and hence that $z(a) \equiv f(b)$. *Atomic Structure* entails that this is true only if $a = b$, but by stipulation $a \neq b$.

But it is not clear what is to be said in favour of *Plenitude*, once we learn to be careful about the heuristic way of thinking in terms of functions that might make it seem undeniable. True, it is a strong and simple generalisation; but so is *Atomic Structure* (and so, as we shall see later, are certain other generalisations inconsistent with *Plenitude*). The final comparison between the packages that include *Plenitude* and those inconsistent with it will have to be made on other grounds.

The third objection takes off from the observation that the structured picture is inconsistent with *Plenitude*. Take any $x^{\langle\langle\rangle,\langle\rangle\rangle}$ that is counterexample to *Plenitude*—a functional relation among propositions that does not correspond to any operator $z^{\langle\langle\rangle\rangle}$. Couldn't we introduce into our language a new symbol $\odot$, with the syntax of a sentential operator, just by stipulating that whenever a sentence φ means that p and $x(p, q)$, $\odot\varphi$ will mean that q? If this stipulation is effective, *Schematic Propositional Structure* will fail in our new language, assuming that our new symbol $\odot$ counts as a legitimate substitution instance for the schematic letter X. For example, x might be $\lambda pq.q \equiv \neg\varphi$, for some chosen sentence φ.[44] Then we have $\odot\psi \equiv \neg\varphi$ for all ψ, though obviously it is not true that $\psi \equiv \varphi$ for all ψ. Even more simply, we could take x to be $\lambda pq.p \equiv q$; then we have $\odot(\neg\varphi) \equiv \neg\varphi$ despite the fact that $\varphi \not\equiv \neg\varphi$.[45]

[44] In a λI-language, use $\lambda pq.q \equiv \neg\varphi \wedge p \equiv p$.

[45] As Jeremy Goodman pointed out, these stipulations do not specify any meaning for sentences

These considerations about made-up languages do not of course show that *Schematic Propositional Structure* has any false instances in our actual language. But they do raise a challenge: given that languages where *Schematic Propositional Structure* fails are possible, what reason to we have for thinking that our own language is not one of them? After all, the primary social functions which explain how our languages socially evolved could, it seems, be fulfilled just as well by the imagined extended languages. As Rayo (2013, p. 10) puts the point: 'It is simply not the case that ordinary speakers are interested in conveying information about metaphysical structure.'[46] Their goals are much more down to earth. One can also turn this into a worry about the non-schematic *Propositional Structure*. Since this does not contain the new symbol $\odot$, it will still be true in the new language if it was true in the old language: this means that higher-order existential generalisation will not be valid for $\odot$. The challenge is to say why, if we reject *Plenitude*, we should ever be confident that existential generalisation works for terms in our current language, given that our communicative purposes can be perfectly well served by languages in which it fails. The underlying worry is that while quantification into operator position might initially seem like a readily intelligible generalisation of our ordinary quantificational idioms, its legitimacy becomes much harder to defend if its application requires us to make a metaphysically contentious distinction between the "bona fide" connectives (which admit existential generalisation) and the merely "apparent" connectives (which do not).

One way to respond to this argument is to insist that the relevant stipulation is simply impossible. This is plausible enough for some of the relevant functional relations x. The idea would be that in some cases where a certain sentence φ means that p, we fail to *know which* q is such that $x(p, q)$ in some metasemantically important sense of 'know which', and because of this, are not in a position to *understand* the new sentence $\odot\varphi$ in a way that conforms to the attempted stipulation. For example, if $x(p, q)$ is 'Either Caesar once asserted p and $q \equiv p$, or Caesar never asserted p and $q \equiv$ snow is white', we arguably know too little to really understand '$\odot$Elephants have trunks'. But for other x, it is hard to see how this kind of complaint could be sustained, since we seem to "know the extension of x" perfectly well. The previous

where $\odot$ occurs as an argument of some higher-order term (e.g. of type $\langle\langle\langle\rangle\rangle\rangle$), rather than having a sentence as its argument. If one wants a stipulation that can make $\odot$ behave just like $\neg$ or any other operator in the higher-order language, one will need something vastly more radical, perhaps a complete reinterpretation of the entire language mapping every term of type $\langle\langle\rangle\rangle$ to one of type $\langle\langle\rangle, \langle\rangle\rangle$ and extending this mapping to all types. This does much to undercut the force of the present objection. Thanks to Jeremy Goodman and Peter Fritz for discussion.

[46] Rayo is arguing against a view he calls "metaphysicalism", which seems quite close to the structured picture: see Dorr 2014b.

example, where $x \equiv (\lambda pq.q \equiv \neg\varphi)$ seems to be like this, in the case where φ is a sentence we understand. In this case, it is hard to see what understanding-theoretic obstacle there could be to introducing $\odot$ by stipulating the truth of $\forall p.\chi(p, \odot p)$. Thus, I think that we will have to get used to the idea that there are possible "ill-behaved" languages in which not all of the connectives are bona fide, so we cannot dismiss out of hand the suggestion that some of the connectives in English might turn out to be like this.[47]

The real problem with the argument, I think, is that the challenge from made-up languages is simply too general to have any bite against the structured picture in particular: one can raise essentially the same challenge concerning *any* putatively valid schema. It would not be a compelling argument against, say, the law of non-contradiction (understood as a schema) to point out that we could stipulatively modify our language in such a way that it ceased to be valid, and that the new language would be no worse than the old from the point of view of everyday communication. So what? One might argue: 'The social function of language would be served equally well by a language in which *most* ordinary instances of this schema were true as by a language in which *all* instances of the schema were true; therefore it would be a surprising coincidence if a community were to end up speaking a language in which *all* instances of the schema were true'. But this seems a bit silly, since the truth of different instances of a schema in a community's language are not probabilistically independent events. Rather, the truth about metasemantics—about what it is for one abstractly specified language rather than another to be spoken by a given community—means that it is simply easier for a community to end up speaking a "regular", "well-behaved" language, other things being equal.[48] They don't have to *care* about regularity, or logic, or metaphysics, for this to happen. It just

[47]Someone might object that when we introduce $\odot$ in the imagined way, φ does not really occur as a syntactic, as opposed to merely orthographic, constituent in the sentence $\odot\varphi$, so that the failure of existential generalisation or the substitution into *Schematic Propositional Structure* is completely unsurprising. (The apparatus of schemas needs to be understood in such a way as to rule out merely orthographic "embeddings": 'c=d → (Fido is a dog → Fido is a cog)' is not an instance of Leibniz's Law.) While this may be correct in some cases, I do not think it would be wise for proponents of the structured picture to rely on the the science of syntax to save them from these kinds of objections. Whether something is a constituent in the sense relevant to *syntax* presumably turns either on cognitive-psychological facts about how speakers process the compound formula, or on sociological facts about the systems of linguistic rules prevalent in a community. To be in a position to insist that stipulations of the kind we have envisaged never create new sentences with genuine syntactic constituents, one would have to be thinking of syntax as directly answerable to metaphysics in a way that seems alien to the practice of actual syntacticians.

[48]The notion of easiness here could be cashed out in terms of physical probability, as in Dorr and Hawthorne 2014.

happens by default, unless something special happens to stop it from happening, such as someone issuing some strange stipulation that would only ever occur to a philosopher trying to prove a point about the power of stipulation.

The fourth objection involves the inconsistency of the structured picture with the following principle:

Involution $\qquad p \equiv_{\langle\rangle} \neg\neg p$

The inconsistency with *Propositional Structure* (or *Schematic Propositional Structure*) is straightforward. Consider, say, a possibility claim $\Diamond\psi$. By Involution, we have $\Diamond\psi \equiv \neg\neg\Diamond\psi$; by *Propositional Structure*, this is true only if $\psi \equiv \neg\Diamond\psi$. But this is false for any true ψ, since sentences flanking a true identification cannot differ in truth value.

But why believe Involution? The strongest case I know of is based on the following thought experiment from Ramsey (1927):

> We might, for instance, express negation not by writing a word "not", but by writing what we negate upside-down. Such a symbolism is only inconvenient because we are not trained to perceive complicated symmetry about a horizontal axis, and if we adopted it we should be rid of the redundant "not-not", for the result of negating the sentence "p" twice would simply be the sentence "p" itself. (Ramsey 1927, pp. 42-43)

If we spoke Ramsey's imagined language, we would simply have no pairs of distinct formulae in our language that relate to one another in the same way that φ relates to $\neg\neg\varphi$ in our actual language.[49] If there are truths of the form $\varphi \not\equiv \neg\neg\varphi$, they are inexpressible in such a language. But it is hard to believe that the use of such a language would be any sort of a *handicap* from a metaphysical point of view.[50]

The point doesn't turn on the lack of any symbol for $\neg$ in Ramsey's language. Suppose the speakers of the language are willing to introduce such a symbol by stipulation. One obvious way for them to accomplish this would be to stipulate that all instances of the schema $\neg\varphi \equiv \mathfrak{b}$ should be true. But $\neg\neg\varphi \equiv \varphi$ will certainly be

[49] We had better assume that the basic symbols of the language are chosen so that we never have vertically mirror-symmetric sentences like the English 'HE DOCKED' (see Sorensen 1999, p. 159).

[50] It would, perhaps, make it harder to describe certain possible mental states, such as the mental states of those intuitionistic logicians who rejected the claim that every set of natural numbers has a least element while still accepting that every set of natural numbers does not not have a least element. But this is not the kind of deficiency metaphysicians should care about: *any* form of language will enable people to get confused in certain distinctive ways, and will be better suited than others for the task of characterising those particular forms of confusion.

true in their language if this stipulation is successful. For starting with $\neg\neg\varphi \equiv \neg\neg\varphi$, substitution in accordance with the schema yields $\neg\neg\varphi \equiv \neg\psi$, which we can then turn into $\neg\neg\varphi \equiv \varphi$ using $\neg\psi \equiv \varphi$, which is another instance of the schema. The other way in which we could imagine them introducing the $\neg$ symbol would be stipulate that it is equivalent to $\lambda p.\flat$ (that's an upside-down 'p' after the dot); if *β-conversion* fails, this does not guarantee that instances of the schema $\neg\varphi \equiv \psi$ are true. But insofar as we think that this gives them a way of expressing the facts we express using 'not', we face a problem of expressive limitation in the opposite direction: *our* language seems to lack the resources to express the facts they express using inverted sentences. There seems to be no way for deniers of Involution to do justice to the intuition that the two languages are on a par.

The fifth and last argument against *Propositional Structure* that I want to discuss is well-known (though it has been neglected): it is known as the "Russell-Myhill paradox", and establishes that *Propositional Structure* is actually *inconsistent* when we have classical higher-order quantification. Let me start by stating the argument loosely in terms of propositions and properties. Choose some arbitrary proposition p, say that snow is white. Let a "heteropredicative" proposition be one that that predicates of p some property that it itself lacks. Now consider the proposition that p is heteropredicative, call it q. Is q heteropredicative? If not, then q must have every property that it predicates of p, and in particular the the property of being heteropredicative; contradiction. So q is heteropredicative: it predicates of p some property f that it, q, lacks. This f cannot be the property of being heteropredicative, which as we have just seen, q does *not* lack. So, there must be two distinct—and indeed non-coextensive—properties which this single proposition q predicates of p.

For those who feel like working through it, here is a rigorous statement of the argument.[51] Let O ("is heteropredicative") abbreviate

$$\lambda q^{\langle\rangle}.\neg\forall f^{\langle\langle\rangle\rangle}((q \equiv f\,p) \rightarrow f\,q)$$

[51]My version of the argument is similar to the versions given (and endorsed) by Hodes 2015 and J. Goodman forthcoming.

Then we argue as follows:

1.	$O(Op) \leftrightarrow \neg\forall f^{\langle\langle\rangle\rangle}(Op \equiv f\,p \rightarrow f(Op))$	(*Extensional β-equivalence*)
2.	$\forall f(Op \equiv f\,p \rightarrow f(Op)) \rightarrow (Op \equiv Op \rightarrow O(Op))$	∀-elim
3.	$\neg O(Op) \rightarrow (Op \equiv Op \rightarrow O(Op))$	(1, 2)
4.	$Op \equiv Op$	(Ref)
5.	$\neg O(Op) \rightarrow O(Op)$	(3, 4)
6.	$O(Op)$	(5)
7.	$\neg\forall f(Op \equiv f\,p \rightarrow f(Op))$	(1, 6)
8.	$\forall f(O \equiv f \rightarrow (f(Op) \leftrightarrow O(Op)))$	(LL)
9.	$\forall f(O \equiv f \rightarrow f(Op))$	(6, 8)
10.	$\neg\forall f(Op \equiv f\,p \rightarrow O \equiv f)$	(7, 9)

The conclusion is plainly inconsistent with the substitution instance $\forall f((Op \equiv f\,p) \rightarrow ((O \equiv f) \wedge (p \equiv p)))$ of *Propositional Structure*.[52]

When we think about *why Propositional Structure* fails in this case, we can see that we should expect failures to be quite pervasive. The argument is essentially Cantorian: one can think of the conclusion as saying that the domain of properties of propositions is larger than the domain of propositions, so that there can be no one-one correspondence between the two domains, and in particular the relation of being a property f and a proposition q such that q is $f(p)$ cannot be one-one as required by *Propositional Structure*. So it is wrong to think of the failure of uniqueness in the case of $O(p)$ as an isolated exception—in the absence of some plausible criterion for confining failures of uniqueness to some special propositions, it seems that we should expect just about any proposition that is the result of applying an operator f to also be the result of applying some other operator that is not even coextensive with f.

One way to block this reasoning is to adopt a ramified type theory like that of Whitehead and B. Russell 1910. Even explaining the basic idea behind this move, let alone properly evaluating it, would take me too far afield; so let me just echo the widespread consensus that this would be a major cost.[53]

[52]The argument remains valid if we uniformly replace all constituents of the form '$X(p)$' with $[X/y^{\langle\langle\rangle\rangle}]\varphi$ for any formula φ. The conclusion will be interesting, and arguably inconsistent with the structured picture, when φ contains at least one occurrence of the operator variable $y^{\langle\langle\rangle\rangle}$. We can recover something close to Russell's original argument (B. Russell 1903, Appendix B) by taking φ to be $\forall p^{\langle\rangle}((y^{\langle\langle\rangle\rangle}p) \rightarrow p)$ ('every y proposition is true').

[53]See Bacon, Hawthorne and Uzquiano 2016, sect. 7 for a survey of some of the forms that ramific-

I conclude that the structured picture is false. Since the structured picture looks to be the simplest and most systematic alternative to *β-conversion*, this bolsters the case for *β-conversion*. But of course, intermediate positions that accept neither *β-conversion* nor the structured picture are imaginable. So, let me close this section by tentatively presenting a positive argument for β-conversion. This argument turns on the familiar practice of stipulative definition, in which new terms are introduced by writing down things like (22a)–(22c):

(22) a. x is a schmixen $=_{\text{df}}$ x is female and x is a fox

b. $\text{Collinear}(x, y, z) =_{\text{df}} \text{Between}(x, y, z) \vee \text{Between}(y, z, x) \vee \text{Between}(z, y, x)$

c. x is a transitive set $=_{\text{df}} \forall y \forall z(y \in x \wedge z \in y \rightarrow z \in x)$

As discussed in §2, it is central to the practice of introducing new predicates in this way that having done so, we get to substitute the open sentence on the right of '$=_{\text{df}}$' for the one on the left, *salva veritate* (perhaps with a special exception for attitude and speech reports). In particular, we should be able to substitute in identifications, licensing claims like (23):

(23) $\text{Schmixen}(\text{Nelly}) \equiv \text{Female}(\text{Nelly}) \wedge \text{Fox}(\text{Nelly})$

For Nelly to be a schmixen is for it to be the case that Nelly is female and Nelly is a fox.

But words we have introduced in this way also seem to be perfectly genuine predicates in our expanded language. We should therefore be able to existentially generalise from (23) to (24):

(24) $\exists f^{\langle e \rangle}(f(\text{Nelly}) \equiv \text{Female}(\text{Nelly}) \wedge \text{Fox}(\text{Nelly}))$

But this sits very strangely with the denial of the following instance of *β-conversion*:

(25) $(\lambda x.\, \text{Female}(x) \wedge \text{Fox}(x))(\text{Nelly}) \equiv \text{Female}(\text{Nelly}) \wedge \text{Fox}(\text{Nelly})$

If (24) is witnessed by some $f^{\langle e \rangle}$, why on earth should we not take it to be witnessed by $\lambda x.\, \text{Female}(x) \wedge \text{Fox}(x)$? If existentially quantified claims like (24) were true, surely the best conventions about the use of λ-abstracts would be one on which

ation might take, including an approach that (unlike that of Whitehead and Russell) keeps the syntax of the language intact and merely replaces each of our quantifiers with a hierarchy of "restricted" quantifiers. For considerations against ramification, see Ramsey 1926 and Prior 1971, ch. 3. Hodes (2015) considers an argument for ramification based on "converse-compositional" principles like *Propositional Structure*, and finds it wanting.

$\lambda x.\,\mathrm{Female}(x) \wedge \mathrm{Fox}(x)$ could be used to stand for the witnesses. And of course this mode of argument is extremely general. For any open sentence φ in which the variables $x_1, \ldots, x_n$ all occur free, we can introduce a new n-ary predicate F by stipulating that $F(x_1, \ldots, x_n) =_{\mathrm{df}} \varphi$, and then infer by substitution that $F(A_1, \ldots, A_n) \equiv [A_i/x_i]\varphi$ and by existential generalisation that $\exists f(f(B_1, \ldots, B_n) \equiv [A_i/x_i]\varphi)$, which makes the denial of $(\lambda x_1, \ldots, x_n.\varphi)(B_1, \ldots, B_n) \equiv [A_i/x_i]\varphi$ seem bizarre. (Note that this seems a lot less compelling in the vacuous case where some of $x_1, \ldots, x_n$ do not occur free in φ, since we do not normally introduce new predicates by means of stipulations like that.)

Another way to use our practice of stipulative definition to argue for *β-conversion* relies not on existential generalisation but on the following schema, which is considerably less controversial than *β-conversion*:

(η-conversion) $A \equiv A^*$, where A^* is derived from A by replacing some constituent of the form $(\lambda v_1 \ldots v_n.F(v_1, \ldots, v_n))$, where none of $v_1, \ldots, v_n$ is free in F, with F.[54]

η-conversion is not in tension with the structured picture, and several authors who reject *β-conversion* in general have expressed sympathy for η-conversion—for example, Fine (2012, §9) and Salmon (2010, §2) both suggest that η-convertible sentences may be equivalent in some strong sense in which β-convertible sentences are not. Thus, the following argument using η-conversion to *β-conversion* is of some interest. As before, let φ be some formula with $x_1, \ldots, x_n$ free, and introduce F by $F(x_1, \ldots, x_n) =_{\mathrm{df}} \varphi$. First use two applications of definitional substitution to get the following identifications:

$$F(A_1, \ldots, A_n) \equiv [A_i/x_i]\varphi$$
$$(\lambda x_1 \ldots x_n.F(x_1, \ldots, x_n))(A_1, \ldots, A_n) \equiv (\lambda x_1 \ldots x_n.\varphi)(A_1, \ldots, A_n)$$

But

$$(\lambda x_1 \ldots x_n.F(x_1, \ldots, x_n))(A_1, \ldots, A_n) \equiv F(A_1, \ldots, A_n)$$

is an instance of η-conversion. So by symmetry and transitivity, we can derive

$$(\lambda x_1 \ldots x_n.\varphi)(A_1, \ldots, A_n) \equiv [A_i/x_i]\varphi$$

(Again, this is much less compelling in the vacuous case.)

[54] Note that instances of η-conversion where F is $(\lambda v_1 \ldots v_n.\varphi)$ for some formula φ are also instances of (nonvacuous) *β-conversion*.

It is natural to think of the formation of complex predicates by λ-abstraction as a device for "automating" the procedure of introducing a new predicate by stipulation and then using it—when we write $\lambda x_1, \ldots, x_n.\varphi$ in a formula (where all of $x_1, \ldots, x_n$ occur free in φ and no other variables do), it is just as if we had inserted a simple predicate F which we had earlier defined by issuing the stipulation $F(x_1, \ldots, x_n) =_{\text{df}} \varphi$. The differences between these procedures seem to be matters of convenience rather than principle. Thus we should not be surprised by the idea that disputed questions about the logic of λ-abstracts can be settled by reference to the behaviour of stipulatively defined predicates.

Opponents of β-conversion will probably reply to these arguments by insisting that we have to choose between two different interpretations for stipulative definitions like (22a)–(22c). We could treat them as *mere abbreviations*, in which case the uses of existential generalisation and η-conversion are not licensed; or we could treat them as true predicates interchangeable with the corresponding lambda terms, in which case definitional substitutions will not be licensed in all contexts (and in particular, not in identifications). But the idea that we have to make such a choice looks like an artefact of a bad theory. True, logicians theorising in a metalanguage about a distinct object language sometimes introduce things called "metalinguistic abbreviations" which are not predicates at all, but part of a system for forming complex names for expressions in the object language. But despite superficial similarities, this practice is really quite different from the practice of stipulative definition as engaged in by mathematicians, scientists, and philosophers, which manifestly does lead to extensions of the language.

7 Booleanism

The questions discussed so far concern the 'pure logic' of identifications; they have nothing specific to say about the behaviour of other familiar logical vocabulary—truth-functional operators or quantifiers—within the scope of identifications. One simple theory about the interaction of identifications with truth-functional operators is *Booleanism*, according to which the truth functional operators conform to the axioms of a Boolean algebra, with identification playing the role of identity. This theory can be axiomatised by adding the following schemas to the logic we have

already (classical logic plus Ref, LL, and βη-conversion):

∧-*Commutativity*	$(\lambda v_1 \ldots v_n.\varphi \wedge \psi) \equiv (\lambda v_1 \ldots v_n.\psi \wedge \varphi)$
∨-*Commutativity*	$(\lambda v_1 \ldots v_n.\varphi \vee \psi) \equiv (\lambda v_1 \ldots v_n.\psi \vee \varphi)$
∧∨-*Distributivity*	$(\lambda v_1 \ldots v_n.\varphi \wedge (\psi \vee \theta)) \equiv (\lambda v_1 \ldots v_n.(\varphi \wedge \psi) \vee (\varphi \wedge \theta))$
∨∧-*Distributivity*	$(\lambda v_1 \ldots v_n.\varphi \vee (\psi \wedge \theta)) \equiv (\lambda v_1 \ldots v_n.(\varphi \vee \psi) \wedge (\varphi \vee \theta))$
∧∨-*Dissolution*	$(\lambda v_1 \ldots v_n.\varphi \wedge (\psi \vee \neg\psi)) \equiv (\lambda v_1 \ldots v_n.\varphi)$
∨∧-*Dissolution*	$(\lambda v_1 \ldots v_n.\varphi \vee (\psi \wedge \neg\psi)) \equiv (\lambda v_1 \ldots v_n.\varphi)$

Here $v_1 \ldots v_n$ stands for any list of zero or more distinct variables, and φ, ψ, θ are formulae. Given these axioms, various other familiar-looking schemas follow (see Huntingdon 1904 for proofs), including the following

Involution	$(\lambda v_1 \ldots v_n.\neg(\neg\varphi)) \equiv (\lambda v_1 \ldots v_n.\varphi)$
∧-*Associativity*	$(\lambda v_1 \ldots v_n.(\varphi \wedge \psi) \wedge \theta) \equiv (\lambda v_1 \ldots v_n.\varphi \wedge (\psi \wedge \theta))$
∧-*Idempotence*	$(\lambda v_1 \ldots v_n.\varphi \wedge \varphi) \equiv (\lambda v_1 \ldots v_n.\varphi)$
∧∨-*De Morgan*	$(\lambda v_1 \ldots v_n.\neg(\varphi \wedge \psi)) \equiv (\lambda v_1 \ldots v_n.\neg\varphi \vee \neg\psi)$
∧∨-*Absorption*	$(\lambda v_1 \ldots v_n.\varphi \wedge (\varphi \vee \psi)) \equiv (\lambda v_1 \ldots v_n.\varphi)$
∧∨-*Annihilation*	$(\lambda v_1 \ldots v_n.\varphi \wedge (\psi \wedge \neg\psi)) \equiv (\lambda v_1 \ldots v_n.\psi \wedge \neg\psi)$

We also get the dual versions that interchange ∧ and ∨.[55]

Booleanism could alternatively be axiomatised using a single schema with a more complicated condition on what counts as an instance:

Taut: $(\lambda v_1 \ldots v_n.\varphi) \equiv (\lambda v_1 \ldots v_n.\psi)$ whenever $\varphi \leftrightarrow \psi$ is a tautology (theorem of classical propositional logic).

The equivalence of *Taut* to the axioms listed above follows from the soundness and completeness of the Boolean-valued semantics for classical propositional logic, where the semantic values of sentences in a model are members of an arbitrary

[55] In a higher-order λK-language, there is no need to use schemas in axiomatising Booleanism. For example, we can replace ∧∨-Distributivity with the single axiom $(\lambda pqr.p \wedge (q \vee r)) \equiv (\lambda pqr.(p \wedge q) \vee (p \wedge r))$. In a λI-language we can still do this for Commutativity and Distributivity, but it will not work for Dissolution since $\lambda pq.p$ is ill-formed. However, Booleans will see no advantage to λI-languages. As noted in §5, vacuous lambda abstracts can be translated into a λI-language by adding tautologous conjuncts to turn them into non-vacuous abstracts; for example translating $\lambda p.\varphi$ when p is not free in φ as $\lambda p.\varphi \vee (p \wedge \neg p)$. Because they accept Dissolution, Booleans accept full β-conversion even for the expanded language.

Boolean algebra and theorems are sentences whose semantic value in every model is the top element of that model's Boolean algebra.

Booleanism has more often been taken for granted than argued for. One exception is Ramsey (1927), whose rather compelling argument for Involution we already encountered in §6. After making this argument, Ramsey goes on in short order to generalise the conclusion to all the other Boolean equivalences. But it is doubtful whether the mode of argument from alternative possible forms of language actually extends as far as this. It certainly extends to De Morgan: Involution entails that both $\wedge\vee$*-De Morgan* and $\vee\wedge$*-De Morgan* hold when $\vee$ is interpreted as $\lambda pq.\neg(\neg p \wedge \neg q)$ or when $\wedge$ is interpreted as $\lambda pq.\neg(\neg p \vee \neg q)$, and it is hard to believe that the actual interpretations of $\wedge$ and $\vee$ fail to fit together in the way that these possible interpretations do. (Indeed, Ramsey's community where negation is represented by inversion will not have different expressions corresponding to '$\neg(\varphi \vee \psi)$' and '$\neg\varphi \wedge \neg\psi$', or '$\neg(\varphi\wedge\psi)$' and '$\neg\varphi\vee\neg\psi$', if they wisely use $\wedge$ and $\vee$ for conjunction and disjunction.) Another argument in a similar style, turning on possible languages whose sentences do not always have to consist of linearly ordered strings of symbols, can be used to support Commutativity.[56] The fact that our language forces us to choose an order for conjuncts and disjuncts seems no more relevant to enhancing its expressive capacity than the fact that we have to choose a typeface in writing, or a tone of voice in speech.

However, it is hard to see how this mode of argument could be extended to any of the other Boolean equivalences. One could imagine a written language in which symbols, once arranged to form a formula, automatically rearrange themselves to form a tautologically equivalent formula in some canonical form (say, disjunctive normal form). But assuming the community in question retained the ability to introduce new simple symbols stipulatively equivalent to old complex expressions, they could use such symbols to generate stable equivalents to sentences of ours that are not in the canonical form, and use these to express counterexamples to Booleanism, if there are any: by contrast with the case of Involution, allowing for definitional expansions of the language seems to restore expressive parity. Anyway, this thought experiment seems lame as an argument for Booleanism in particular, given that we can also imagine a script whose symbols providentially rearrange themselves into the shortest expression equivalent to what was written down *modulo* the laws of nature, or indeed *modulo* the truth about some arbitrarily chosen subject matter. While users of such languages are lucky in one way, in that it is harder for them to communicate false beliefs about the relevant subject matter, they are surely subject

[56]Cf. Williamson (1985), who in a somewhat different context imagines languages whose sentences are built up by putting expressions into bags.

to genuine expressive limitations (until they introduce new simple symbols that do not get destroyed by the rearrangement). From an anti-Boolean point of view, the biconditionals corresponding to the Boolean axioms are not importantly different from the laws of nature in this respect.

Define $\Box$ as $\lambda p.p \equiv \top$, where $\top$ is some arbitrarily chosen closed tautology, say $\exists p(p) \vee \neg\exists p(p)$. This operator $\Box$ is of great interest in a Boolean setting, partly because it can take over the role of the propositional identification connective $\equiv_t$, since $\Box(\varphi \leftrightarrow \psi) \leftrightarrow (\varphi \equiv \psi)$ is a theorem of Booleanism.[57] Booleanism implies that all instances of the modal axioms K and T hold for $\Box$.[58] Moreover, it would be natural for Booleans to endorse further principles about embedded occurrences of $\equiv$ which have the effect of making $\Box$ be a normal modal operator obeying the logic S4, or perhaps even S5.[59]

If one thought that $\Box$ had an S5 logic, it would be natural to equate the claim that $\Box\varphi$—i.e. that $\varphi \equiv \top$—with the claim that it is *metaphysically necessary* that φ. Some might see this as a welcome explanation of the unfamiliar (identification) in terms of the familiar (metaphysical necessity).[60] My attitude towards the proposed

[57] *Proof:* Suppose $\varphi \equiv \psi$. $(\varphi \leftrightarrow \psi) \equiv (\varphi \leftrightarrow \varphi)$ is true by Ref and LL; $\top \equiv (\varphi \leftrightarrow \varphi)$ is true by *Taut*, so $(\varphi \leftrightarrow \psi) \equiv \top$ is true by LL. In the other direction, suppose $(\varphi \leftrightarrow \psi) \equiv \top$. $\varphi \equiv (\psi \leftrightarrow (\varphi \leftrightarrow \psi))$ is true by *Taut*, so by LL, $\varphi \equiv (\psi \leftrightarrow \top)$; also $(\psi \leftrightarrow \top) \equiv \psi$ by *Taut*, so $\varphi \equiv \psi$ by LL.

[58] For K, suppose that $\varphi \rightarrow \psi \equiv \top$ and $\varphi \equiv \top$; then $(\top \rightarrow \psi) \equiv \top$ by LL; since $(\top \rightarrow \psi) \equiv \psi$ by *Taut*, a second application of LL yields $\psi \equiv \top$. For T, suppose that $\varphi \equiv \top$; $\top$ is true by propositional logic, so we can infer φ by LL.

[59] The following principles suffice for S4:

Necessitated Ref	$(x \equiv_\tau x) \equiv \top$
Necessitated LL	$((A \equiv B) \rightarrow (\varphi \rightarrow [B/_\gamma A]\varphi)) \equiv \top$

Both of these are natural generalisations of *Taut*, since they assert identifications where the corresponding biconditionals were already provable in the system. Furthermore, if we accept *The Identity Identity*, we can derive these principles from the following natural analogues of Booleanism for the quantifiers:

$$((\lambda v_0 \ldots v_n.A) \equiv (\lambda v_0 \ldots v_n.A \wedge B)) \leftrightarrow ((\lambda v_1 \ldots v_n.\exists v_0(A)) \equiv (\lambda v_1 \ldots v_n.\exists v_0(A) \wedge B))$$
$$((\lambda v_0 \ldots v_n.A) \equiv (\lambda v_0 \ldots v_n.A \vee B)) \leftrightarrow ((\lambda v_1 \ldots v_n.\forall v_0(A)) \equiv (\lambda v_1 \ldots v_n.\forall v_0(A) \vee B))$$

(Here B is a formula in which v_0 does not occur free. See Dorr 2014a and J. Goodman 2016 for more on these "Adjunction" principles.) To get to S5, on the other hand, we would have to add something much less clearly well-motivated, namely the $\Box$-necessity of distinctness:

$$(x \not\equiv_\tau y) \rightarrow ((x \not\equiv_\tau y) \equiv \top)$$

For arguments that Booleans should reject this principle, see Bacon MS and J. Goodman MS.

[60] Such an explanation would be worth little if it only applied to the sentential connective $\equiv_{\langle\rangle}$,

equation, if I believed that S5 was valid for $\Box$, would be the reverse. Philosophers have struggled to say something helpful to single out the "metaphysical" readings of modal operators from among the panoply of other readings they may bear; their efforts have not been conspicuously successful. So an explication of 'It is metaphysically necessary that φ' as 'For it to be the case that φ is for it to be the case that $\top$' would shed some welcome light on the concept of metaphysical necessity and the interest of questions formulated in terms of it.[61] True, as we saw in §1, 'To be F is to be G' also admits several readings; but at least it doesn't have the same array of readings as 'necessarily', and the task of singling out the target reading is perhaps less challenging. Moreover, whereas it is easy to find philosophers who claim to have no idea what metaphysical necessity is or to regard it as in some sense a defective notion, it is hard to see how any philosopher could be so dismissive of 'To be F is to be G' given the central role questions about the truth of such claims play in most branches of the subject.

'Metaphysically necessary' was introduced into the philosophical vernacular partly through general formulas—for example, the equation of metaphysical necessity with "unrestricted" or "absolute" necessity, or 'necessity in the highest degree—whatever that means' (Kripke 1972, p. 99)—but partly also through philosophers exchanging opinions about which truths are, in fact, metaphysically necessary—e.g. that nothing is green and red all over, that Nixon is not an inanimate object, that a certain lectern is not made of ice, etc. If you were convinced of the falsity of claims like 'Nixon is not an inanimate object $\equiv \top$' or 'The lectern is not made of ice $\equiv \top$', etc., you might worry that the proposed interpretation of 'metaphysically necessary' would be unduly uncharitable. However given the character of 'metaphysically necessary' as a term of art, charity to the explicit explanatory remarks made by those who introduced the term should weigh especially heavily with us in deciding how to interpret it, while charity to claims that they made where they thought of themselves as taking a stand on philosophically controversial questions should count for relatively little. If we can make sense of a notion of necessity with a good claim to labels like 'necessity in the highest degree', it would be perverse to interpret 'metaphysically necessary' as expressing something more restricted just for the sake of being more charitable to the case-by-case modal pronouncements of, say, Saul Kripke.

---

but there are natural strategies for extending it to other types. For example, one might maintain that $(A \equiv_{\langle \tau_1,\ldots,\tau_n \rangle} B) \equiv_{\langle\rangle} \Box\forall x_1 \ldots \forall x_n (A(x_1, \ldots, x_n) \leftrightarrow B(x_1, \ldots, x_n))$. Or if one rejected this on the grounds that it is inconsistent with contingentism, one might still accept $(A \equiv_{\langle \tau_1,\ldots,\tau_n \rangle} B) \equiv_{\langle\rangle} \Box\forall x_1 \Box \ldots \forall x_n \Box (A(x_1, \ldots, x_n) \leftrightarrow B(x_1, \ldots, x_n))$.

[61] The idea of defining necessity in terms of an identity connective occurs, for example, in Cresswell 1965 and Suszko 1975.

This would be as misguided as interpreting philosophers like Shoemaker (1998), who maintain a doctrine that they express by saying 'All laws of nature are metaphysically necessary', as merely meaning by 'metaphysically necessary' what other philosophers have meant by 'nomically necessary'. If the logic of $\Box$ is S5, it has a very strong claim to be the "absolute" form of necessity, since whenever $\varphi \equiv \top$ is true, $O\varphi$ must be true for every other transparent operator O such that $O\top$ is true, a category which certainly includes all non-epistemic necessity operators.

If the project of using identifications to say what it is to be metaphysically necessary were feasible only for S5-endorsing Booleans, that would be a weighty consideration in favour of their view. In fact, however, even if we do not question the orthodox view that the logic of metaphysical necessity is S5, this project is also open to non-Booleans, and to Booleans who reject S5 for $\Box$. The question how this should go is beyond the scope of the present paper, but one obvious strategy would be to first define a higher-order predicate *is an S5 operator* (of type $\langle\langle\langle\rangle\rangle\rangle$), and then identify *being metaphysically necessary* with *being mapped to a truth by every S5 operator*, $\lambda q.\forall x^{\langle\langle\rangle\rangle}(\mathrm{S5}(x) \rightarrow x(q))$.[62] It is not obvious that this operator will turn out to itself be a S5 operator. But if this can be shown, the case for identifying it with metaphysical necessity will be parallel to the case for identifying $\lambda p.p \equiv \top$ with metaphysical necessity if its logic is S5. Thus, while it does seem like an advantage of a theory if it allows for a plausible explanation of metaphysical necessity in terms of identifications, this advantage is not distinctive to Booleanism. The feature that does seem to be distinctive to Booleanism is the possibility of explaining identifications in terms of metaphysical necessity; but the ability to do this does not seem to me to be much of an advantage at all.

[62]For Booleans, "S5(x)" might be defined as $\mathrm{Taut}(x) \wedge \mathrm{K}(x) \wedge \mathrm{T}(x) \wedge 4(x) \wedge \mathrm{B}(x)$, where these are in turn defined in terms of a type-$\langle\langle\rangle\rangle, \langle\langle\rangle\rangle\rangle$ "entailment" connective $\leq_{\langle\langle\rangle\rangle}$ as follows:

$$\begin{aligned}
\mathrm{Taut}(x) &=_{\mathrm{df}} (\lambda q.q \vee \neg q) \leq (\lambda q.x(q \vee \neg q)) \\
\mathrm{K}(x) &=_{\mathrm{df}} \lambda pq.(xp \wedge p \leq q) \leq (\lambda pq.xq) \\
\mathrm{T}(x) &=_{\mathrm{df}} x \leq (\lambda q.q) \\
4(x) &=_{\mathrm{df}} x \leq (\lambda q.x(x(q))) \\
\mathrm{B}(x) &=_{\mathrm{df}} (\lambda q.q) \leq (\lambda q.x(\neg x(\neg q)))
\end{aligned}$$

Here, $x \leq_{\langle\langle\rangle\rangle} y$ can in turn be defined as $x \equiv_{\langle\langle\rangle\rangle} \lambda p.(x(p) \wedge y(p))$. Non-Booleans could adopt the same strategy but with a different definition of $\leq$. One possible definition that is well behaved in a wide variety of non-Boolean settings, including the models considered in the Appendix, is $x \leq_{\langle\langle\rangle\rangle} y =_{\mathrm{df}} (\lambda p.x(p) \wedge (x(p) \wedge y(p))) \equiv_{\langle\langle\rangle\rangle} (\lambda p.x(p) \vee (x(p) \wedge y(p)))$ (see Appendix A8).

8 Non-circularity

Even though the arguments for Booleanism considered in the previous section were unconvincing, Booleanism is a simple and powerful metaphysical vision. At times, indeed, its simplicity has won me over. But before we can properly assess the force of such abductive considerations, we will need a much more detailed appreciation of the space of alternatives to Booleanism.

In the remainder of this paper, I want to consider one particular kind of consideration which might lead one to reject Booleanism, and see what kind of alternative it might suggest. Consider the following claims:

GRUE: To be grue is to be either green and observed before t, or blue and not observed before t.

BLEEN: To be bleen is to be either blue and observed before t, or green and not observed before t.

GREEN: To be green is to be either grue and observed before t, or bleen and not observed before t.

BLUE: To be blue is to be either bleen and observed before t, or grue and not observed before t.

GRUE and *BLEEN* are uncontroversial: just look at the passages of N. Goodman 1954 in which the words 'grue' and 'bleen' are introduced. *GREEN* and *BLUE*, on the other hand, are very odd. It is tempting to think—*pace* Goodman himself—that they are simply false. Colour scientists and philosophers consider various views about what it is to be green and what it is to be blue. Maybe to be green is to be disposed to affect perceivers in a certain way, or to be disposed to reflect light in a certain way, or to have a surface with certain intrinsic physical characteristics. I can get into a frame of mind where *GREEN* and *BLUE* seem like obviously misguided competitors to these serious views. Of course, those who think that they are true will probably want to tell some pragmatic story about why they would be unhelpful things to assert in many contexts. Note however that 'To be green is to be green' does *not* seem false although it is as unhelpful as can be, so those who think that *GREEN* and *BLUE* are true should not be too sanguine about the prospects for a pragmatic explanation of their oddity that undercuts the temptation to think them false.[63]

[63]One reason one might have for rejecting *GREEN* and *BLUE* that has to do with the fact that a particular time t is mentioned on the right hand sides: one might object that being green and being blue are not about any particular times (perhaps on the grounds that they are qualitative), whereas being

The following derivation shows that *GREEN* follows from *GRUE* and *BLEEN* given Booleanism. Here 'Gx', 'Bx', '$G'x$', '$B'x$' and 'Ox' respectively abbreviate 'x is green', 'x is blue', 'x is grue', 'x is bleen' and 'x is observed before t':

1. $(\lambda x.Gx) \equiv (\lambda x.((((Gx \wedge Ox) \vee (Bx \wedge \neg Ox)) \wedge Ox) \vee$ (*Taut*)
   $(((Bx \wedge Ox) \vee (Gx \wedge \neg Ox)) \wedge \neg Ox)))$
2. $G \equiv (\lambda x.(((\lambda y.(Gy \wedge Oy) \vee (By \wedge \neg Oy))(x) \wedge Ox) \vee$
   $((\lambda y.(By \wedge Oy) \vee (Gy \wedge \neg Oy))(x) \wedge \neg Ox)))$ (1, βη-conversion)
3. $G' \equiv (\lambda y.(Gy \wedge Oy) \vee (By \wedge \neg Oy))$ (*GRUE*)
4. $B' \equiv (\lambda y.(By \wedge Oy) \vee (Gy \wedge \neg Oy))$ (*BLEEN*)
5. $G \equiv (\lambda x.(G'x \wedge Ox) \vee (B'x \wedge \neg Ox))$ (2,3,4, LL)

So if we want to reject *GREEN* (which is line 5), we have to give up Booleanism.

One project we could engage in at this point is that of going through the list of Boolean axioms and theorems and try to decide which ones to keep and which to give up. There is a large literature we can draw on in this enterprise: for any logic weaker than classical propositional logic, we could consider a weakening of *Taut* where equivalence in that logic replaces equivalence in classical propositional logic. But we should hope to be able to do better than this, not simply weakening Booleanism but articulating some general principles that are actually *inconsistent* with it, so as to put together a competing package with advantages of simplicity and strength of its own that could be set against those of Booleanism. Of course, one might think that there just are no true principles that are both inconsistent with Booleanism and comparable in simplicity and generality to those of Booleanism. But general methodological considerations suggest that such a view should be regarded as a last resort.[64]

The case of *GREEN* and *BLUE* is suggestive in this regard. For in my case at least, the inclination to think them false is not primarily based on any positive thoughts about what it might really be to be green or blue, or even on the nature of colour in general. Rather, I am inclined to think that *GREEN* and *BLUE* can be ruled out simply on the basis of *GRUE* and *BLEEN*. Just looking at the logical form of these identifications, I have an impulse to say that they cannot possibly all be true

---

either grue and observed before t or bleen and observed after t is about the particular time t. But I am interested in a more general reason that would still apply even if we replaced 'observed before t' throughout with 'observed at some time or other'.

[64]For abductive methodology in metaphysics, see Williamson 2013 (especially its 'Methodological Afterword') and Williamson 2016.

together, since that would be *circular*.

The idea that there is something "vicious" (i.e. impossible) about a kind of circularity that would be exhibited by this set of identifications is encouraged if we think of identifications as "real definitions". For as readers of mathematics textbooks know, our standard practice of stipulative definition is certainly governed by a "no circularity" constraint. If I have stipulatively defined a simple expression using a certain complex expression, I am not allowed later to stipulatively define one of the simple constituents of that complex expression using a different complex expression that contains the originally defined simple expression.[65] The label 'real definition' suggests that there is some subject matter in reality which is importantly analogous to this human practice, and a "no circularity" constraint would provide one natural basis for such an analogy.

Another way to support some kind of "no circularity" principle about identifications is to lean on the idea that identifications *explain* claims about necessity. As noted in §1, identifications—at least those with a syntactically simple predicate on one side—seem to provide maximally satisfying explanations of the necessities that follow from them; if to be a vixen *is* to be a female fox, there is no further sense in wondering why it is necessary that all vixens are female foxes. But if someone offered to explain some initially mysterious necessity by citing identifications which run in a circle, we would feel cheated; the sense of there being nothing left to explain seems to disappear. Suppose, for example, that we thought that (26) was true:

(26) It is metaphysically necessary that whenever distinct lines x and y are both perpendicular to some third line z, x and y do not intersect.

We might wonder *why* this is true. As we might put it: what stops intersecting lines from ever being perpendicular to a third line? It would seem a bad joke if someone proposed (27a) and (27b) as an answer:

(27) a. ($\lambda xy.x$ is perpendicular to y) $\equiv$ ($\lambda xy.x$ intersects y and every line that intersects both x and y intersects each of them obliquely).

b. ($\lambda xy.x$ intersects y obliquely) $\equiv$ ($\lambda xy.x$ intersects y and x is not perpendicular to y).[66]

[65] "Implicit definitions" might be offered as a counterexample to this claim. But in ordinary mathematical practice it is taken for granted that implicit definitions are shorthand for more complicated explicit definitions, e.g. of the form 'x is $F =_{\text{df}} x$ belongs to every set closed under such-and-such operations'. In cases where this recipe cannot be applied, the acceptability of a so-called implicit definition is a matter of deep philosophical controversy, in sharp contrast to the acceptability of standard stipulative definitions.

[66] Understand 'intersect' in (27a) and (27b) in such a way that lines do not intersect themselves.

Although (26) follows straightforwardly from (27a) and (27b)—if x and y are both perpendicular to z, then by (27a) y intersects z, if y intersects both x and z it intersects both x and z obliquely, and by (27b) y does not intersect z obliquely, so y cannot intersect x—the suggestion that they are both true does nothing to allay any puzzlement one might have had about (26).

Similarly, in the philosophy of mind, some philosophers (e.g. Setiya 2007, ch. 1) engaged with the question 'What is it to intend to do something?' take one of their central tasks to be that of explaining (28):

(28) Necessarily, everyone who intends to do something believes that they will do it.

The kind of understanding of (28) these philosophers are looking for certainly cannot be attained just by accepting some "circular" identification like (29):

(29)

If combinations like (27a) and (27b) or single identifications like (29) were tenable, the task of explaining puzzling facts about necessity by deriving them from identifications would be much easier than it seems in fact to be. This provides further motivation for the idea that there is some formal "non-circularity" constraint which is violated by pairs like these.

But what does it even *mean* to say that identifications cannot "run in a circle"? We had better be careful. Given *Reflexivity*, 'To be a vixen is to be a vixen' cannot count as "circular" in the objectionable sense; given *Symmetry*, neither can the combination of 'To be a vixen is to be a female fox' with 'To be a female fox is to be a vixen'. A more promising suggestion is that the relevant notion of circularity involves the term on one side of an identification occurring as a *proper constituent* of the term on the other side:

No Circles: $A \not\equiv_\tau B$, where A and B are terms of any type τ such that B properly contains an occurrence of A in which no occurrence of any variable free in A is bound.

(Note that *No Circles* also rules out conjunctions of identifications $(A_1 \equiv B_1) \wedge (A_2 \equiv B_2)$ where A_1 is a proper constituent of B_2 and A_2 is a proper constituent of B_1 (with no variables bound): we can use LL and the second conjunct to substitute B_2 for the occurrence of A_2 in B_1 on the right hand side of the first conjunct, to get an identification $A_1 \equiv B_1^*$ where A_1 is a proper constituent of B_1^*.) Principles like *No Circles* have been taken seriously: for example, Correia (2010, p. 16) suggests that it may be correct given what he calls a "conceptualist view of factual equivalence".

And perhaps Prior (1964, p. 193) is endorsing something like *No Circles* when he writes 'I cannot see how the sense of a sentence can ever be identical with a logical complication of itself'.[67] But we cannot accept *No Circles* as it stands, since we have endorsed β-conversion, which implies $\varphi \equiv (\lambda p.p)(\varphi)$, and also (on independent grounds) tentatively endorsed $\varphi \equiv \neg\neg\varphi$: both of these are identifications in which the term on one side of $\equiv$ is a proper constituent of the term on the other.

We need a principle that lets us discriminate between benign circles and vicious ones. The idea I want to propose is this: the case where the term on one side of a true identification occurs as a proper constituent of the term on the other side can arise only if all of the *other* expressions on the more complex side—all of the other ingredients which combine with the term on the less complex side to form the term of which it is a proper constituent—are or are equivalent to *logical* terms, like $\neg$ and $(\lambda p.p)$ in the above examples. *Non-logical* entities are indissoluble, and always make for a genuine increase in complexity when they combine with something else. To capture this idea in our higher-order language, let us help ourselves to a predicate Logical_τ, of type $\langle\tau\rangle$, for every type τ. $\text{Logical}_\tau(x^\tau)$ should be true only if x^τ is the denotation of by some closed term whose only constants are the logical constants $\neg$, $\wedge$, $\neg$, $\forall_\tau$, $\exists_\tau$, and $\equiv_\tau$. While this gloss may not constitute a satisfactory *definition* of 'Logical_τ', it seems to convey an adequate grip on the intended interpretation. Using this vocabulary, we can state the proposal schematically as follows:

Only Logical Circles: $(A \equiv_\tau B) \rightarrow \text{Logical}_\sigma(C)$, where σ and τ are any types, and A, B, C are any terms, of types τ, τ, and σ respectively, such that B contains an occurrence of A together with an occurrence of C that neither contains, nor is identical to, nor is contained by that occurrence of A, and none of the variables free in A or C are bound in either of these occurrences.

Recall that we count universal closures of instances of schemas as instances in their own right; thus all of the following are instances of *Only Logical Circles*:

$$(\exists q(q) \equiv (\lambda p.p)(\exists q(q))) \rightarrow \text{Logical}_{\langle\langle\rangle\rangle}(\lambda p.p)$$
$$\forall x^{\langle\langle\rangle\rangle}\forall p^{\langle\rangle}((p \equiv x(p)) \rightarrow \text{Logical}_{\langle\langle\rangle\rangle}(x))$$
$$\forall f^{\langle e\rangle}\forall p^{\langle\rangle}((f \equiv (\lambda x.f(x) \wedge p)) \rightarrow \text{Logical}_{\langle\rangle}(p))$$

Note too that *Only Logical Circles* is a non-starter if we accept vacuous β-conversion. Vacuous β-conversion entails that $p \equiv (\lambda q.p)(p)$ is true for every p, which given *Only*

[67] Note however that there are many passages in Prior that suggest a commitment to β-conversion: he extracts predicates from sentences by replacing some constituents with blanks, and takes it for granted that in doing this he is providing one legitimate 'analysis' of the original sentence.

Logical Circles would imply the obviously false conclusion that $\text{Logical}_{\langle\rangle}(p)$ is true for every p.

Only Logical Circles does not look very elegant: the criterion for what counts as an instance of the schema involves some rather fiddly syntactic considerations. Fortunately, using the power of β-conversion, we can achieve the same effect as *Only Logical Circles* with the following simpler schema:

OLC $\qquad (x \equiv_\tau \lambda v_1 \ldots v_n.y(z, x, v_1, \ldots, v_n)) \rightarrow \text{Logical}_\sigma(z)$

This schema has one instance for every pair σ, τ of types where $\tau \neq e$: this fixes the types of the variables, since when τ is $\langle \tau_1, \ldots, \tau_n \rangle$, $v_1, \ldots, v_n$ must be of types $\tau_1, \ldots, \tau_n$, and y must therefore be of type $\langle \sigma, \langle \tau_1, \ldots, \tau_n \rangle, \tau_1, \ldots, \tau_n \rangle$. (When $\tau = \langle\rangle$, the list of variables $v_1, \ldots, v_n$ is empty, so in this case *OLC* is just $(x \equiv_{\langle\rangle} y(z, x)) \rightarrow \text{Logical}_\sigma(z)$, where y is of type $\langle \sigma, \langle\rangle \rangle$.)

As a special case of *OLC*, we have the principle that if a proposition is the result of applying some operator to itself, that operator must be a logical one:

$$(p \equiv f(p)) \rightarrow \text{Logical}_{\langle\langle\rangle\rangle}(f)$$

(To get this from *OLC* take $x = p$, $z = f$, and $y = \lambda g^{\langle\langle\rangle\rangle} q^{\langle\rangle}.(g(q))$.) As we might put it with apologies to Prior: the sense of a sentence may never be identical to a *non-logical* complication of itself.

OLC follows from *Only Logical Circles*, since each of its instances is an instance of *Only Logical Circles*. It also implies *Only Logical Circles*. Assume for conditional proof that $A \equiv_\tau B$, where B is a complex term containing non-overlapping occurrences of A and C in which none of their free variables are bound. Since B is complex, it is either a formula φ or an abstract $\lambda v_1 \ldots v_n.\varphi$. Let u and w be variables of the same types as A and C which do not occur free in B, and let φ^* be the result of substituting u and w for the two given non-overlapping occurrences of A and C in φ. Then φ is β-equivalent to $(\lambda u w v_1 \ldots v_n.\varphi^*)(C, A, v_1, \ldots, v_n)$, so by β-conversion, we can transform $A \equiv B$ into

$$A \equiv (\lambda v_1 \ldots v_n.(\lambda u w v_1 \ldots v_n.\varphi^*)(C, A, v_1, \ldots v_n))$$

So by universal instantiation, we can substitute A, $(\lambda u w z v_1 \ldots v_n.\varphi^*)$, and C respectively for x, y, z in *OLC* and thus derive Logical(C).

As we hoped, *OLC* can be used to rule out combinations of identifications such as *GRUE*, *BLEEN*,*GREEN*, and *BLUE*. In fact we only need to consider *GRUE* and

GREEN. Suppose both were true:

(30) $a.\quad G' \equiv \lambda x^e.(Gx \wedge Ox) \vee (Bx \wedge \neg Ox)$

$b.\quad G \equiv \lambda x^e.(G'x \wedge Ox) \vee (B'x \wedge \neg Ox)$

Combining these using LL and β-conversion, we get

(31) $G \equiv \lambda x^e.(((Gx \wedge Ox) \vee (Bx \wedge \neg Ox)) \wedge Ox) \vee (B'x \wedge \neg Ox)$

in which the closed term on the left of ≡ is a proper constituent of the one the right. Since the term on the right contains non-overlapping occurrences of the constants G and B, (31) implies $\text{Logical}_{\langle e\rangle}(B)$ by *Only Logical Circles*.[68] But given the intended meaning for 'Logical', this conclusion—that being blue is logical—should seem obviously false.[69]

Of course it is not enough to establish this merely to point out that 'blue' is not itself on our official list of logical constants. There is nothing to stop us from introducing new constants equivalent to the given logical constants, or to complex terms built out of them. Nevertheless, when we consider a sample of closed terms of type $\langle e\rangle$, the idea that 'blue' is even *coextensive* with any such term, let alone equivalent to one in the sense of ≡, looks terribly implausible:

$$\lambda x.x = x \qquad \lambda x.\exists f^{\langle e\rangle} fx$$
$$\lambda x.\forall f^{\langle e\rangle} \neg fx \qquad \lambda p.\exists f^{\langle e\rangle}(fx \wedge \exists y(y \neq x \wedge fy))$$
$$\lambda x.\exists f^{\langle e\rangle}(fx \wedge \forall y(y = x \rightarrow \neg fy)) \qquad \lambda x.\exists f^{\langle e,\langle\rangle\rangle}(f(x, f(x)))$$

The idea that we could, even in principle, state necessary and sufficient conditions for something to blue given only this meagre list of ingredients to work with strains credulity far past the breaking point.[70] We can give a parallel argument from *OLC*

[68]To reach the same conclusion using *OLC*, we need to applying another β-conversion to (31) to get to something which has the right form to instantiate the antecedent of *OLC*:

$$G \equiv \lambda x^e((\lambda f^{\langle e\rangle} h^{\langle e\rangle} x.(((hx \wedge Ox) \vee (fx \wedge \neg Ox)) \wedge Ox) \vee (B'x \wedge \neg Ox))(B, G, x))$$

[69]Note that even Carnap 1928 does not endorse this: although he calls everything on his final list of unreduced vocabulary "logical", it includes not just truth-functional operators and quantifiers but a higher-order predicate 'fund', expressing something like naturalness.

[70]One might be tempted to argue that being blue isn't logical from the premise that any two objects have exactly the same logical properties. But this is not obviously true. As we will see in §9, it is plausible in the present setting that *qualitativeness* is definable in logical terms, in which case so is *qualitative indiscernibility* (sharing all the same qualitative properties) and *being qualitatively*

against many other intuitively "circular" combinations of identifications: for example, we can rules out the combination of (27a) and (27b) using the equally plausible auxiliary premise that *perpendicularity* is not logical.

In fact, in many cases, we can get by without having to rely on any such auxiliary premise. There is no way, in our higher-order language, to build a term of type e out of logical constants. (In fact there are no complex terms of type e at all). Thus on the intended interpretation of 'Logical', 'Logical(z)' is always false when z is of type e. So by setting $\sigma = e$, we can extract from *OLC* the following schema, whose instances are themselves purely logical sentences:

Qual $\qquad x \not\equiv_\tau \lambda v_1 \ldots v_n . y(z^e, x, v_1, \ldots, v_n)$

Using *Qual*, we can rule out the conjunction of *GRUE* and *GREEN* without having to rely on any premises about logicality—indeed the only additional premise we need is the claim that there is at least one object. Choose any object a; then by applying both sides of (31) to a and β-converting, we get

(32) $\qquad Ga \equiv (((Ga \wedge Oa) \vee (Ba \wedge \neg Oa)) \wedge Oa) \vee (B'a \wedge \neg Oa)$

A second β-conversion then yields

(33) $\qquad Ga \equiv (\lambda x^e p^{\langle\rangle}.(((p \wedge Ox) \vee (Bx \wedge \neg Ox)) \wedge Ox) \vee (B'x \wedge \neg Ox))(a, Ga)$

whose negation is an instance of *Qual*.

OLC has the striking consequence that *no non-logical proposition is its own self-conjunction*:

(34) $\qquad (p \equiv (p \wedge p)) \rightarrow \text{Logical}_{\langle\rangle}(p)$

For the antecedent is β-equivalent to $p \equiv (\lambda qr.r \wedge q)(p, p)$, which implies $\text{Logical}_{\langle\rangle}(p)$ by *OLC*. Turning to types other than $\langle\rangle$, it turns out we only need *Qual* to derive the

indiscernible from something distinct from oneself. It possible that some but not all objects have this property: for example, this would plausibly be the case if the world consisted of two spatiotemporally disconnected parts of which one but not the other was mirror-symmetric. The suggestion that being blue is logical is thus not *quite* as absurd as we might have supposed. We can imagine that the everyday world consists of swarms of co-located, qualitatively indiscernible objects, and that swarms containing different numbers of such objects tend to reflect different amounts of light at different wavelengths. Insofar as this is an epistemic possibility, perhaps we should allow that it might turn out, say, that being blue *is* being qualitatively indiscernible from exactly seventeen other objects, in which case being blue is logical after all. However, actual colour science does not provide a fertile ground for such far-fetched speculations.

even more sweeping consequence that *no* property is its own self-conjunction. For example, in type $\langle e \rangle$, *Qual* implies the schema

(35) $$f \not\equiv_{\langle e \rangle} \lambda x.(f(x) \wedge f(x))$$

For suppose that $f \equiv \lambda x.(f(x) \wedge f(x))$; then for any object a, we would have $f(a) \equiv f(a) \wedge f(a)$, which is β-equivalent to $f(a) \equiv (\lambda z^e q^{\langle\rangle}.(q \wedge f(z)))(a, f(a))$, which is inconsistent with *Qual*. This generalises to arbitrary complex types $\langle \tau_1, \ldots, \tau_n \rangle$ $(n \geq 1)$:

(36) $$f \not\equiv_{\langle \tau_1, \ldots, \tau_n \rangle} \lambda x_1 \ldots x_n.(f(x_1, \ldots, x_n) \wedge f(x_1, \ldots, x_n))$$

Suppose f were a counterexample to this; let $A_1 \ldots A_n$ be some terms of types $\tau_1 \ldots \tau_n$ whose only free variable is z^e, and let $B_1 \ldots B_n$ be the result of substituting the name a for z^e in these terms.[71] Then we have

(37) $$f(B_1, \ldots, B_n) \equiv (\lambda z^e q^{\langle\rangle}.(q \wedge (\lambda z^e.f(A_1, \ldots, A_n))z))(a, f(B_1, \ldots, B_n))$$

again contradicting *Qual*. Given this derivation of (36) for all types other than $\langle\rangle$, considerations of uniformity arguably favour strengthening (34) to $p \not\equiv (p \wedge p)$ even in type $\langle\rangle$.

The replacement of Idempotence with its negation (at least in complex types) is perhaps the most distinctive hallmark of the particular kind of "fine-grained" theory we are developing. It distinguishes it, for example, from Goodman's theory of aboutness (mentioned in note 40), and from the theories of "worldly factual equivalence" developed in Correia 2010 and Correia 2016, all of which endorse Idempotence. I admit that this feature is quite surprising: especially given that we are endorsing Involution, you might have expected other especially "trifling" equivalences in propositional logic to correspond to true identifications. Correia and Skiles MS suggest the rejection of Idempotence as a hallmark of a "conceptual" or "representational" (as opposed to "worldly") conception of the kind of claim made by sentences of the form '$\varphi \equiv \psi$', where this is taken to involve, for example, denying that '*a* is a water molecule $\equiv$ *a* is a H_2O molecule' is true, on the grounds that its two sides involve distinct "concepts". But I insist that despite the fact that the present theory rejects Idempotence, it is still intended as a theory of a kind of claim just as "worldly" as ordinary identity claims. The hypothesis is that in the very sense in which it is true that to be a water molecule is to be a H_2O molecule, it is just not true that to be red

[71] A proof that there is at least one such λI-term A_i in every type can be extracted from a proof of a related fact given in the Appendix, note 105.

is to be red and red. I insist that these claims are *not* obviously false. They are "edge cases" of the sort one would only ever consider as part of the kind of systematic logical investigation we are currently engaged in, and as such they should be settled on the basis of broader systematic considerations.[72]

Many of the other Boolean theorems and axioms turn out, given *OLC*, to have the same status as $\wedge$-*Idempotence*: universally false in complex types, and also in type $\langle\rangle$ with a possible exception for logical propositions. For example, consider Dissolution, $p \equiv p \wedge (q \vee \neg q)$. This is β-equivalent to $p \equiv (\lambda rs.s \wedge (r \vee \neg r))(q, p)$, which implies Logical($q$) by *OLC*. In type $\langle e \rangle$, $f \equiv \lambda x.f(x) \wedge (g(x) \vee \neg g(x))$ implies $f(a) \equiv (\lambda x^e q^{\langle\rangle}.q \wedge (g(x) \vee \neg g(x))(a, f(a)))$, which is inconsistent with *Qual*. Absorption and Annihilation go the same way, as does the combination of the two Distributivity principles.[73]

This replacement of $\wedge$-*Idempotence* with its opposite (at least in complex types) shows that the present theory fulfils our aspiration to have a theory that is inconsistent with Booleanism and comparable to it in generality. Moreover, the fact that *Qual* is enough by itself to secure so much of the strength of this result shows that the interest of the theory does not depend crucially on the presence in the language of the undefined predicates 'Logical$_\sigma$'. Someone might object that the apparent intelligibility of these predicates is an illusion based on taking model theory too seriously, or that a theory using such predicates suffers from a pernicious kind of ideological complexity. I think that these objections are mistaken, but I will not engage with them here except to note that they are not objections to *Qual*.

[72] An important worry about the failure of idempotence concerns infinite conjunctions. If infinitary conjunction is intelligible at all, we can form the infinite conjunction of φ, $\varphi \wedge \varphi$, $\varphi \wedge \varphi \wedge \varphi$, …: call it φ^ω. If φ is non-logical, so is φ^ω; so by (34), we have $\varphi^\omega \not\equiv (\varphi^\omega \wedge \varphi^\omega)$. But this is puzzling, since we might think that φ^ω and $\varphi^\omega \wedge \varphi^\omega$ are both just the conjunction of φ with itself countably many times. The best resolution to this puzzle, I suspect, will involve rejecting $\wedge$-*Associativity* even in the finite case. If associativity fails, we should not expect that describing a proposition as a 'conjunction of so-and-so many copies of φ' would suffice to pin it down uniquely. For example, we will distinguish $\wedge(\wedge(\varphi, \varphi), \wedge(\varphi, \varphi))$ from $\wedge(\wedge(\varphi, \wedge(\varphi, \wedge(\varphi, \varphi))))$; we may also want to distinguish both of these from $\wedge(\varphi, \varphi, \varphi, \varphi)$ (a simple, quaternary conjunction). However there are several further choice points we need to consider in order to come up with an account of infinitary conjunction and disjunction that fits with *OLC*.

[73] The hardest case is that of Distributivity. Suppose we have a non-logical proposition $f(a)$, call it p for short. Combining $\wedge\vee-$Distributivity and $\vee\wedge-$Distributivity yields $p \wedge (p \vee p) \equiv ((p \wedge p) \vee p) \wedge ((p \wedge p) \vee p)$; by *Commutativity* this yields $p \wedge (p \vee p) \equiv (p \vee (p \wedge p)) \wedge ((p \wedge p) \vee p)$; substituting in the first disjunct another application of $\wedge\vee$-Distributivity then gives us $p \wedge (p \vee p) \equiv ((p \vee p) \wedge (p \vee p)) \wedge ((p \wedge p) \vee p)$, and another application of $\vee\wedge$-Distributivity turns this into $p \wedge (p \vee p) \equiv ((p \wedge (p \vee p)) \vee (p \wedge (p \vee p))) \wedge ((p \wedge p) \vee p)$. This, finally, has the left hand side as a proper constituent of its right hand side, so that we can apply *Only Logical Circles* to get Logical(p), which is impossible if $p \equiv f(a)$. Note too that given De Morgan's and Involution, which we want to accept, the two Distributivity principles become equivalent.

It is worth noting that *OLC*, and indeed *Qual*, requires rejecting the principle of *Plenitude* which we considered back in §6, according to which every functional relation between propositions corresponds to an operator. For any property f and object a, we can consider the relation $R =_{\text{df}} \lambda pq.(q \equiv f(a) \wedge (p \equiv p))$, which every proposition bears uniquely $f(a)$. Since R is functional, *Plenitude* implies that there is a corresponding operator z, such that for any p, $R(p, z(p))$—i.e. for any p, $z(p) \equiv f(a)$. So in particular, $f(a) \equiv z(f(a) \wedge f(a))$. Since this β-converts to $f(a) \equiv (\lambda xp.f(x) \wedge p)(a, f(a))$, it is inconsistent with *Qual*. The present theory thus commits one just as much as the structured picture did to making a distinction in principle between genuine operators (which allow existential generalisation) and mere "quasi-operators" which use the meanings of their input sentences to fix the meanings of their output sentences in a manner captured by some functional $\zeta^{\langle\langle\rangle,\langle\rangle\rangle}$ that does not correspond to any $\chi^{\langle\langle\rangle\rangle}$.

Now that we have seen how to derive lots of interesting, controversial, and anti-Boolean conclusions from *OLC* (together with β-conversion and the background classical logic), an urgent question is whether you can derive everything else as well, including the negations of these conclusions. In other words: is *OLC consistent*? The answer is not at all obvious, but in the Appendix, I show that it is yes, by constructing a nonempty class of set theoretic models for a λI-language in which all instances of *OLC* have value 1, and in which all the rules of the background logic (including βη-conversion) preserve value 1, although not everything has value 1. Since Involution, De Morgan, Commutativity, and Associativity also hold in some of the models in the class, the proof also shows that *OLC* is consistent with these principles. Moreover, in these models, '$\text{Logical}_\sigma(x)$' has value 1 on an assignment only if the denotation of x on that assignment is the same as that of some closed term built out of logical constants; thus, the models are consistent with 'Logical' meaning what I wanted it to mean. In fact, *OLC* also remains true even when we contract the extension of 'Logical' further so that that $\text{Logical}_\sigma(x)$ has value 1 on an assignment only when the denotation of 'x' is the same of that of some closed term in which the only constant is ¬: i.e. a pure term such as $\lambda p^{\langle\rangle}.p$ or $\lambda f^{\langle e,e\rangle} xy.f(y, x)$ or $\lambda f^{\langle\langle\rangle\rangle} p^{\langle\rangle}.f(f(p))$, or the result of inserting some negation symbols into such a term. One result of this strengthening is that we can no longer build any "logical" terms of type $\langle\rangle$, just as we previously could not build any logical terms of type e. This means that '$\text{Logical}_{\langle\rangle}(p)$' is always false, so we derive another schema not involving 'Logical', analogous to *Qual*:

$$x \not\equiv_\tau \lambda v_1 \ldots v_n.y(z^t, x, v_1, \ldots, v_n)$$

Amongst other things, this lets us drop the exception for logical propositions from (34) and the negations of the other Boolean schemas.[74]

I am not sure whether it is a good move to strengthen *OLC* in this way. On the one hand, if we keep Involution, the strengthened view accords to negation a certain status (that of being able to take part in "circles") that it denies to the other logical constants, which might seem invidious. On the other hand, the strengthened view has the nice feature that it makes type $\langle\rangle$ behave more uniformly with other types with respect to the negations of Boolean axioms. I leave the question open for now.

9 Priority

Despite our disagreements about their logic, I think we understand identifications very well. My grip on them feels much firmer than my grip on expressions like the following, which have also been cropping up in philosophy from its inception:

fundamental
more fundamental than
metaphysically primitive
derivative from
metaphysically prior to
prior in the order of being

These expressions are, I think, among the most obscure in all of philosophy. Looking at the wide variety of ways in which they are used, we may reasonably worry that they are hopelessly vague, to such an extent that no sentence expressed in terms of them would be both interesting and definitely true. So it will be a substantial advance if it turns out that certain such sentences can be glossed or reconstructed in terms of identifications.[75]

Against the background of the kind of theory explored in §8, we can define something which can reasonably be thought of as capturing a notion of metaphysical priority. For any x and y, we can say that x is "weakly prior" to y iff y is the result of

[74]This also applies to all the other types σ in which there are no pure terms, for example $\langle e, \ldots, e\rangle$. Moreover, in types with only finitely pure terms up to βη-equivalence, there are also only finitely many terms whose only constant is $\neg$ up to βη-equivalence and double negation elimination; so for these σ we can also restate *OLC* in purely logical terms by replacing 'Logical$_\sigma$' with a finite disjunction containing one representative of each equivalence classes. For example, in type $\langle\langle e, e\rangle, e, e\rangle$, there are just such four equivalence classes, represented by $\lambda rxy.r(x, y)$, $\lambda rxy.r(y, x)$, $\lambda rxy.\neg r(x, y)$, and $\lambda rxy.\neg r(y, x)$.

[75]I feel the same way about 'grounds' and 'in virtue of', but the ideas in the present section are less directly relevant to them. For one natural approach to analysing grounding in terms of identifications, see Correia and Skiles MS.

saturating an argument place of some z with x, and that x is "strictly prior" to y iff x is weakly prior to y and y is not weakly prior to x. x and y here can variables of any type other than e (since it doesn't make sense for an object to be the result of saturating an argument place of a relation). Formally, we express this using connectives $\preccurlyeq_{\sigma,\tau}$ and $\prec_{\rho,\tau}$, both of type $\langle\sigma,\tau\rangle$ where σ and τ are any types other than e:

$$x \preccurlyeq_{\sigma,\tau} y =_{\text{df}} \exists z(y \equiv_{\tau} \lambda v_1 \ldots v_n(z(x, v_1, \ldots, v_n)))$$
$$x \prec_{\sigma,\tau} y =_{\text{df}} (x \preccurlyeq_{\sigma,\tau} y) \wedge \neg(y \preccurlyeq_{\tau,\sigma} x)$$

Here where τ is $\langle\tau_1, \ldots, \tau_n\rangle$, the variables $v_1, \ldots, v_n$ must be of types $\tau_1, \ldots, \tau_n$ respectively, and so z must be of type $\langle\sigma, \tau_1, \ldots, \tau_n\rangle$. When τ is $\langle\rangle$, the variable list $v_1, \ldots, v_n$ is empty, so $x \preccurlyeq_{\sigma,\langle\rangle} p$ just means $\exists z(p \equiv z(x))$. We can also introduce a connective expressing the case of weak priority without priority:

$$x \approx_{\sigma,\tau} y =_{\text{df}} x \preccurlyeq_{\sigma,\tau} y \wedge y \preccurlyeq_{\tau,\sigma} x$$

This could be pronounced as "x and y are coeval".

Using βη-conversion we can show that weak priority is reflexive ($\forall x(x \preccurlyeq_{\tau,\tau} x)$) and transitive ($\forall x \forall y \forall z((x \preccurlyeq_{\sigma,\tau} y \wedge y \preccurlyeq_{\tau,\rho} z) \rightarrow x \preccurlyeq_{\sigma,\rho} z)$).[76] Weak priority is not a partial order, however: even when x and y are of the same type τ, we can have $x \approx_{\tau,\tau} y$ without $x \equiv_{\tau} y$. For example, Involution entails that $p \approx_{\langle\rangle,\langle\rangle} \neg p$, since $p \equiv \neg(\neg p)$ entails $\exists z(p \equiv z(\neg p))$. Similarly, β-conversion entails that $r \approx_{\langle e,e\rangle,\langle e,e\rangle} \lambda xy.r(y, x)$: each relation is coeval with its converse, since $r \equiv (\lambda sxy.s(y, x))(\lambda xy.r(y, x))$. *Strict* priority, on the other hand, is a strict partial order: transitive because $\preccurlyeq$ is, and irre-

[76] *Proof:* For reflexivity, note that when x is of type $\langle\tau_1, \ldots, \tau_n\rangle$, $x \equiv x$ is η-equivalent to $x \equiv \lambda y_1, \ldots, y_n.x(y_1, \ldots, y_n)$, which is β-equivalent to

$$x \equiv \lambda y_1 \ldots y_n.(\lambda z u_1 \ldots u_n.z(u_1, \ldots u_n))(x, y_1, \ldots, y_n),$$

which implies $\exists z(x \equiv \lambda y_1 \ldots y_n.z(x, y_1, \ldots, y_n))$ by existential generalisation. For transitivity, suppose $a \preccurlyeq_{\sigma,\tau} b$ and $b \preccurlyeq_{\tau,\rho} c$, where τ is $\langle\tau_1, \ldots, \tau_n\rangle$ and ρ is $\langle\rho_1, \ldots \rho_m\rangle$. Then for some $d^{\langle\sigma,\tau_1,\ldots,\tau_n\rangle}$ and $e^{\langle\tau,\rho_1,\ldots,\rho_m\rangle}$, $b \equiv \lambda u_1 \ldots u_n.(d(a, u_1, \ldots, u_n))$ and $c \equiv \lambda v_1 \ldots v_m.(e(b, v_1, \ldots, v_m))$. Substituting the first of these into the second yields

$$c \equiv \lambda v_1 \ldots v_m.(e(\lambda u_1 \ldots u_n.(d(a, u_1, \ldots, u_n)), v_1, \ldots v_m))$$

which is β-equivalent to

$$c \equiv \lambda v_1 \ldots v_m.((\lambda w_0 \ldots w_m.e(\lambda u_1 \ldots u_n.d(w_0, u_1, \ldots, u_n), w_1, \ldots, w_m))(a, v_1, \ldots v_m))$$

which implies $\exists x(c \equiv \lambda v_1 \ldots v_m.(x(a, v_1, \ldots v_m)))$ by existential generalisation. All this looks less forbidding if we work in a functional language of the kind explained in Appendix A2.

flexive because $\preccurlyeq$ is reflexive.

OLC gives us an easy way to establish claims of strict priority: whenever $b \equiv \lambda v_1 \ldots v_n.y(z, a, v_1, \ldots v_n)$ and z is not logical, a must be strictly prior to b. For example, green is strictly prior to grue, since grue $\equiv (\lambda x.(\lambda f^{\langle e \rangle} y^e.(f\,y \wedge Oy) \vee (By \wedge \neg Oy))(\text{green}, x)$, and being green is not logical. Whenever we have a true identification $B \equiv C$ where A is a constituent of C and C also has a non-overlapping, non-logical constituent, *OLC* implies that $A \prec B$, since if there were any way to get back to A by plugging B into an argument place of some other term, the result would then be the kind of circle forbidden by *OLC*. This is a good fit for the way in which the notion of priority is used in informal philosophical settings. For example, the theory that to know something is to have a justified true belief in it would be naturally described as one on which belief is prior to knowledge, whereas the theory that to believe something is to be such that one would know it if one were in normal circumstances would be described as one on which knowledge is prior to belief.

By contrast, three of the competing views we have considered entail that the notions of priority we have just defined are far too undiscriminating to be of any interest: for some or all σ, τ, everything is weakly prior to everything else ($\forall x^\sigma \forall y^\tau (x \preccurlyeq_{\sigma,\tau} y)$) and thus nothing is strictly prior to anything ($\neg\exists x^\sigma \exists y^\tau (x \prec_{\sigma,\tau} y)$). Booleanism and vacuous β-conversion both imply that $\preccurlyeq_{\sigma,\tau}$ is universal for any σ and τ (other than e). *Plenitude*, meanwhile, implies that $\preccurlyeq_{\sigma,\langle\rangle}$ is universal (every proposition is weakly posterior to everything), which is nearly as bad.

Proof. Let $\sigma = \langle \sigma_1, \ldots, \sigma_n \rangle$ and $\tau = \langle \tau_1, \ldots, \tau_m \rangle$.

(i) It is easy to see why Booleanism entails that $\preccurlyeq_{\langle\rangle,\langle\rangle}$ is universal: for any p and q, we have $q \equiv q \wedge (p \vee \neg p)$ by Dissolution, hence $q \equiv (\lambda r.q \wedge (r \vee \neg r))(p)$, so $p \preccurlyeq q$. To generalise this reasoning to show that $x \preccurlyeq_{\sigma,\tau} y$, let $z_1, \ldots, z_n$ be of types $\sigma_1, \ldots, \sigma_n$; then Dissolution implies that

$$y \equiv \lambda v_1 \ldots v_m.(y(v_1, \ldots, v_m) \wedge (x(z_1, \ldots, z_n) \vee \neg x(z_1, \ldots, z_n)))$$

which implies $x \preccurlyeq_{\sigma,\tau} y$ for the same reason as before.

(ii) It is easy to see why vacuous *β-conversion* entails that $\preccurlyeq_{\langle\rangle,\langle\rangle}$ is universal: for any p and q, we have $q \equiv (\lambda r.q)(p)$, which implies $p \preccurlyeq q$. To generalise this reasoning to show that $x \preccurlyeq_{\sigma,\tau} y$, consider the following vacuous β-equivalence:

$$y \equiv \lambda v_1 \ldots v_m.((\lambda u_0 \ldots u_m.y(u_1, \ldots, u_m))(x, v_1, \ldots, v_m))$$

(iii) To show that *Plenitude* implies the universality of $\preccurlyeq_{\sigma,\langle\rangle}$, for a given q let R^q be $\lambda u^\sigma s^{\langle\rangle}.u \equiv u \wedge s \equiv q$. By *Plenitude*, there is a corresponding operator O^q (of type

$\langle\sigma\rangle$) such that for all x^σ, $R^q(x, O^q(x))$. By β-conversion this implies that for all x, $O^q(x) \equiv q$, and hence that $x \preccurlyeq_{\sigma,\langle\rangle} q$.[77] □

Indeed, given Booleanism or *Plenitude*, there seems to be no prospect of finding a non-trivial reconstruction of questions about metaphysical priority in terms of identifications.[78]

We have not provided any way of making sense of the question whether an *object* is strictly prior or strictly posterior to something else: $\prec_{\sigma,\tau}$ is well-defined only when both σ and τ are types other than e. By contrast, our definition of $\preccurlyeq_{\sigma,\tau}$ actually makes perfectly good sense if σ is e, so long as τ is not e. Extending the use of $\preccurlyeq$ to this case, we can say for example that Obama $\preccurlyeq_{e,\langle\rangle}$ Obama is tall, and similarly Obama $\preccurlyeq_{e,\langle e\rangle}$ $\lambda x.x$ admires Obama. Saying that an object is weakly prior to a proposition, property, or relation is a way of saying that it is *about* that object, in a demanding sense of "about"—a sense in which it is at least not *obviously* true that the proposition that every man is mortal is about every man. So long as we deny that $\preccurlyeq_{e,\langle\rangle}$ is universal, there is a strong case for identifying *qualitativeness* with not being, in this demanding sense, about anything.[79] That is:

$$\text{Qualitative}_\tau \equiv \lambda x^\tau.\neg\exists y(y \preccurlyeq_{e,\tau} x)$$

Since the good standing of the notion of qualitativeness is relatively uncontroversial, the fact that we can give this simple analysis of it in logical terms is a significant advantage of the present framework, which rejects Booleanism, vacuous *β-conversion*, and *Plenitude*.[80] For the reasons given above, Booleanism and vacuous *β-conversion*

[77] *Plenitude* is a special case of a stronger principle—*Strong Plenitude*, discussed in §A3 of the Appendix—which implies that $\preccurlyeq_{\sigma,\tau}$ is universal for all $\sigma, \tau \neq e$.

[78] Proponents of vacuous β-conversion might, by contrast, hope to restore nontriviality by somehow restricting the existential quantification in the definition of $\preccurlyeq$. However it is far from obvious how this could be done while preserving transitivity.

Even if we reject all three of Booleanism, vacuous β-conversion, and *Plenitude*, there is of course no guarantee that our $\preccurlyeq$ and $\prec$ will behave in a way that would make it reasonable to think of them as expressing notions of metaphysical priority. For example, although J. Goodman (MS) rejects $q \equiv q \wedge (p \vee \neg p)$ in general, he accepts it in the case where p is qualitative, so his view implies that $x \preccurlyeq y$ whenever x is qualitative; more generally, $x \preccurlyeq y$ exactly when x is about every object that y is about.

[79] Khamara (1988) considers a suggestion like this: 'A property, P, is impure [non-qualitative] if and only if there is at least one individual, y, such that, for any individual, x, x's having P consists in x's having a certain relation to y.'

[80] This advantage does not turn on accepting anything like *OLC*, however. The same definition of qualitativeness is available in the setting of Goodman's theory of aboutness (J. Goodman MS), although that theory makes our $\preccurlyeq$ pretty useless as an account of priority talk, since everything qualitative is weakly prior to everything.

imply that $\forall x^e \forall y^\tau . x \preccurlyeq_{e,\tau} y$, and *Plenitude* implies that $\forall x^e \forall p^{\langle\rangle} . x \preccurlyeq_{e,\langle\rangle} p$, which rules out accepting the above account of qualitativeness on pain of having to say that nothing at all is qualitative.

The ease with which we can make sense of the notion of metaphysical priority, or relative fundamentality, naturally raises the question whether we can also make sense of a corresponding notion of absolute fundamentality.[81] Here is an obvious initial suggestion:

(F1) $$\text{Fundamental}_\tau(x) =_{\text{df}} \neg\exists y(y \prec_{\tau,\tau} x)$$

Loosely: a fundamental property or relation is one to which no other property or relation of the same type is strictly prior.[82] However, this threatens to be in one respect too demanding, and in another respect not demanding enough. To see how it might be too demanding, consider the thesis that conjoining any property with self-identity yields the same property back:

(38) $$\forall f^{\langle e\rangle}(f \equiv (\lambda x. fx \wedge x = x))$$

Since conjunction and identity are logical, this is not ruled out by *OLC*. (38) implies that self-identity is weakly prior to all properties—$\forall f((\lambda x.x = x) \prec_{\langle e\rangle,\langle e\rangle} f)$—and hence, that the only properties that are fundamental in the sense of (F1) are those that are also weakly prior to self-identity. But given *OLC*, (38) also implies that the only properties that are weakly prior to self-identity are logical properties. (Suppose f is weakly prior to self-identity: then there is some z such that $(\lambda x.x = x) \equiv (\lambda x.z(f, x))$, hence $f \equiv (\lambda x. fx \wedge x = x) \equiv (\lambda x.(fx \wedge x = x) \wedge x = x) \equiv (\lambda x.(fx \wedge z(f, x)) \wedge z(f, x))$, which implies $\text{Logical}_{\langle e\rangle}(f)$ since it contains two non-overlapping occurrences of f.) We could respond to this just by giving up (38)—something we will already be committed to if we adopt the strengthening of

[80]The above definition will also look problematic when combined with certain views that postulate special objects whose logical behaviour is very different from that of ordinary individuals. For example, if one believed in *propositions* (understood as special objects) and held that $\forall p^{\langle\rangle}\exists x^e(p \equiv \text{True}(x))$, one would object that the definition wrongly implies that $\forall p^{\langle\rangle}\neg\text{Qualitative}(p)$. This is not the right place for an argument against such views.

[81]For the idea that we should want to be able to talk about naturalness or fundamentality for quantifiers and connectives as well as ordinary predicates, see Sider 2011; for a higher-order implementation of this thought, see Dorr and Hawthorne 2013, sect. 2.

[82]Note that any notion of fundamentality defined in these terms will be one on which anything coeval with something fundamental is itself fundamental; in particular, given Involution, the negations of fundamental properties are themselves fundamental. See Plate 2016 for an analysis of fundamentality (or "logical simplicity") in the setting of a first-order theory of properties that also embraces this consequence.

OLC we were contemplating at end of §8, where '$\text{Logical}_\sigma(x)$' is interpreted so that it is true only when x is the denotation of a closed term whose only constant is $\neg$. But an analogous worry will still arise in other types. For example, we definitely have to admit that the identity operator $\lambda p.p$ is weakly prior to every other operator, since $x^{\langle\langle\rangle\rangle} \equiv (\lambda q.(\lambda yr.y(x(r))))(\lambda p.p, q)$ by *β-conversion*. Thus the only fundamental operators are those (like negation) which count as 'Logical' in the sense relevant to *OLC*, since *OLC* implies that only these could be weakly prior to the identity operator. This is not completely indefensible—no positive claims about fundamentality are uncontroversial—but it is especially odd in the context of Involution, since Involution implies that negation is fundamental in the sense of (F1), and if we are saying this there is pressure to say that at least one of conjunction and disjunction is fundamental too. It seems better to modify (F1) by simply leaving logical entities (in whatever turns out to be the sense relevant to *OLC*) out of consideration altogether:

(F2) $$\text{Fundamental}_\tau(x) =_{\text{df}} \forall y^\tau(y \prec_{\tau,\tau} x \rightarrow \text{Logical}_\tau(y))$$

This addresses the first worry, but still leaves us with the second worry, about not being demanding enough. The problem is that it allows x^τ to count as fundamental even when it is strictly posterior to some y^σ, where σ is some type other than τ. For example, given a (non-logical) binary relation $r^{\langle e,e\rangle}$, we can build up properties like $\lambda x^e.r(x, x)$, $\lambda r^e.r(x, a)$, and $\lambda x^e.\exists y^e(r(x, y))$. These should not count as fundamental in the target sense, since r is strictly prior to all of them. However there is no special reason to expect any non-logical *property* $f^{\langle e\rangle}$ to be strictly prior to any of them.

To address this problem, it looks like (F2) needs to be strengthened to something like this:

(F3) $$\text{Fundamental}_\tau(x) =_{\text{df}} \bigwedge_\sigma \forall y^\sigma(y \prec_{\sigma,\tau} x \rightarrow \text{Logical}_\sigma(y))$$

where the right hand side is an infinite conjunction with one conjunct for every type σ other than e. However, such an infinite conjunction is not expressible in the kind of higher-order language we have been working in, and its intelligibility raises some difficult issues. It is not enough to be tolerant of infinite conjunctions and disjunctions, since such tolerance is naturally combined with a tolerance of types of transfinite adicity (i.e. predicates that make a sentence only when combined with infinitely many arguments), and one might hold that there are too many such types for it to be possible to subsume all of them even in an infinite conjunction or disjunction. A proper treatment of this issue is beyond the scope of the present work.

On an optimistic note, however, it's not clear that we really need infinitely many

conjuncts. If a property $f^{\langle e \rangle}$ is derived from a three-place relation $r^{\langle e,e,e \rangle}$ by reflexivisation, or by quantification, or by saturating its arguments with objects, then it is also derived in one of these ways from some binary relation $s^{\langle e,e \rangle}$. So at least if these operations are representative of the ways in which something of one type can be strictly posterior something of a more complex type, it looks defensible to omit the conjunct corresponding to $\langle e, e, e \rangle$ in the definition of Fundamental$_{\langle e \rangle}$. And this suggests that we might modify (F3) by restricting the conjunction to a finite collection of types whose complexity (on some measure) is not too much greater than that of the given type τ.[83]

An account of absolute fundamentality in logical terms would be a very nice thing to have, since it would help to precisify, and perhaps also to resolve, a wide range of interesting but elusive metaphysical questions.[84] But even setting this aside, a purely logical conception of priority (relative fundamentality) is nothing to sniff at. Claims of priority turn up all over philosophy, especially when we want to engage in a kind of "big picture" thinking that abstracts away from the nitty-gritty details of particular controversial identifications. The fact that the non-circularity picture provides an explanation of priority that justifies our standard practice of reasoning from identifications to priority claims is a significant additional consideration in its favour.

10 Conclusion

Given how central identifications have always been in philosophy, it is surprising how little has been done to explore their logic. I think people have been held back by the assimilation of identifications to questions about the identity of properties, together with the assumption that the latter questions would turn out to be merely verbal. This leads to the idea that the right response to the kinds of general questions we have been concerned with—questions like 'Is it the case that to be red is to be

[83]The resulting analysis of fundamentality will not be quite as neutral as we might have hoped. For example, someone who believed in a fundamental *contemplating* relation of type $\langle e, \langle e, e, e \rangle \rangle$ (a relation between objects and three-place relations) as well as a fundamental *betweenness* relation of type $\langle e, e, e \rangle$ might object that dropping the conjunct for type $\langle e, e, e \rangle$ in the definition of Fundamental$_{\langle e \rangle}$ leads to *contemplating betweenness* being incorrectly classified as fundamental (since we have to ascend to type $\langle e, e, e \rangle$ before we find something nonlogical that is strictly prior to it, namely betweenness). But complete neutrality on all potentially controversial questions is too much to demand. Unlike, for example, a definition of fundamentality that took the form of a mere list, the finitised version of (F3) leaves open a very wide array of views about what is fundamental, and thus provides a reasonable way of reconstructing what might be at stake in many debates about fundamentality.

[84]For example, we could consider to what extent fundamentality as defined plays the various roles for *perfect naturalness* discussed in Dorr and Hawthorne 2013.

red and either square or not square?' and 'Is it the case that to be red is to be red and red?'—will always be something like this: 'If you are talking about "properties" in sense A, then obviously yes; if you are talking about "properties" in sense B, then obviously no'.[85] But as a reason for scepticism, this is premature. Indeed it is quite obscure what conception of the range of available readings for identifications could justify the assumption that our very general questions will turn out to have boringly obvious answers on all their readings, although more specific questions of identification—whether to be morally right is to maximise happiness, whether to be water is to be H_2O, and so forth—still have readings on which their answers are interesting and non-obvious. I hope that the present investigation can illustrate the progress that can be made when one, instead, takes the general questions as seriously as we are used to taking the specific ones. As we have seen, it is not just a matter of settling some intrinsically uninteresting edge cases. Rather, different systematic approaches to such questions reflect deeply and fascinatingly different views about the nature of reality at an extremely general level. Moreover, we have seen how investigating the logic of identifications might illuminate many other questions traditionally of interest to metaphysicians, including questions about metaphysical necessity, priority, and fundamentality. And the exploration has barely begun: there is a whole continent of views waiting to be mapped out, and at this point we can only guess which of them will look most believable in the long run. Onwards![86]

[85]Influential here has been the view of Lewis, who writes, concerning the question whether triangularity and trilaterality are the same property, that 'I don't see it as a matter of dispute. Here there is a rift in our talk of properties, and we simply have two different conceptions' (Lewis 1986, p. 55).

[86]This paper has been in progress in some form or another for a very long time, and there no way I could thank all those who deserve thanks. But I would like to mention Lucas Champollion, Kit Fine, John Hawthorne, Thomas Hofweber, Jessica Moss, Jim Pryor, Mark Schroeder, Kieran Setiya, Ted Sider, Zoltan Szabó, Peter van Inwagen, and Timothy Williamson. Special thinks to Andrew Bacon, Peter Fritz, and Jeff Russell, who provided helpful guidance; and especially special thanks to Jeremy Goodman, whose influence on the final result has been pervasive.

A Appendix

The central goal of the appendix is to prove the consistency of *OLC* in a classical deductive system with nonvacuous βη-conversion. However, I also want to give a more rigorous presentation of the syntax of the higher-languages I have been working with than I gave in the main text, and to present a general model theory for these languages that will also be useful to those investigating other views of logic of identifications.

The plan is as follows. A1 will give a more precise characterisation of the syntax of the higher-order languages introduced in §4 (relationally typed higher order languages). A2 will characterise a different family of languages (functionally typed higher order languages), and show how to translate back and forth between them and relationally typed languages. A3 will present the basic definition of a model, and A4 will put this definition in context by introducing some significant properties of models. A5 and A6 will develop some definitions and results that will be useful in the proof of the main model existence theorem: this proof will then be given in §A7. Finally A8 will consider some extensions of the result.

There is much here that is well-known to those who know it. In particular I will be drawing extensively on the textbook Hindley and Seldin 2008 (henceforth HS) and on Benzmüller, Brown and Kohlhase 2004 (henceforth BBK).

Many of the objects we will be interested in are 'collections' indexed by types, in one of two senses of 'type'. In discussing such entities, the following general definitions will be useful.

- When $\mathscr{T}$ is any set, a ***$\mathscr{T}$-typed collection*** $\mathscr{C}$ is a set of ordered pairs such that the second co-ordinate of each pair is a member of $\mathscr{T}$. When $\mathscr{C}$ is an $\mathscr{T}$-typed collection and $\tau \in \mathscr{T}$, we write $\mathscr{C}_\tau$ for $\{x : \langle x, \tau\rangle \in \mathscr{C}\}$.

- When $\mathscr{C}$ is a $\mathscr{T}$-typed collection, $\bigcup\mathscr{C}$ is the set of all first co-ordinates of members of $\mathscr{C}$. When X is any set, $\mathscr{C} - X$ is the typed collection such that $(\mathscr{C} - X)_\tau = \mathscr{C}_\tau \setminus X$ for every $\tau \in \mathscr{T}$.

- An $\mathscr{T}$-collection $\mathscr{C}$ is ***nonoverlapping*** if $\mathscr{C}_\sigma \cap \mathscr{C}_\tau = \varnothing$ whenever $\sigma \neq \tau$; ***completely overlapping*** if $\mathscr{C}_\sigma = \mathscr{C}_\tau$ for every σ, τ; and ***populated*** if $\mathscr{C}_\sigma \neq \varnothing$ for every σ.

- When $\mathscr{C}$ and $\mathscr{D}$ are $\mathscr{T}$-typed collections, a ***typed function*** f from $\mathscr{C}$ to $\mathscr{D}$ is a function f from $\mathscr{C}$ to $\mathscr{D}$ such that $f(x)$ always has the same second co-ordinate as x. For any $\tau \in \mathscr{T}$, we write f_τ for the function $\mathscr{C}_\tau$ to $\mathscr{D}_\tau$ such that

$f(\langle x, \tau\rangle) = \langle f_\tau(x), \tau\rangle$ for all $x \in \mathscr{C}_\tau$. We denote the set of all typed functions from $\mathscr{C}$ to $\mathscr{D}$ by $\mathscr{D}^{[\mathscr{C}]}$.

- We write '$x_1 \underset{\tau_1}{\mapsto} y_1, \ldots, x_n \underset{\tau_n}{\mapsto} y_n$' for the minimal typed function f such that $f_{\tau_i}(x_i) = y(i)$, i.e. $\{\langle\langle x_1, \tau_1\rangle, \langle y_1, \tau_1\rangle\rangle, \ldots, \langle\langle x_n, \tau_n\rangle, \langle y_n, \tau_n\rangle\rangle\}$.

Often it will be convenient to work with non-overlapping typed collections, since this enables an abuse of notation where, given $f \in \mathscr{D}^{[\mathscr{C}]}$, we can just write $f(x)$ to mean 'the unique y such that for some $\tau \in \mathscr{R}$, $f(\langle x, \tau\rangle) = \langle y, \tau\rangle$'.

We also define a $\mathscr{T}$***-family*** is any function whose domain is $\mathscr{T}$. When F is a $\mathscr{T}$-family and $\tau \in \mathscr{T}$, we write F_τ instead of $F(\tau)$. When we use this subscript notation it should always be clear in context whether we have in mind a typed collection, a typed function, or a typed family. We could instead have defined a typed collection as a typed family of sets, and a typed function as a typed family of functions. The advantage of the above definitions is that standard set-theoretic and function-theoretic concepts can be applied directly to typed collections and functions, for example we can take the union of two $\mathscr{T}$-typed collections or the composition of two $\mathscr{T}$-typed functions.

A1 Relational types and relationally typed languages

In this section I give a more precise definition of the "relational" higher order languages introduced in §4. We begin by defining the set of types or syntactic categories.

Definition 1.1. $\mathscr{R}$, the set of ***relational types***, is the smallest set containing the letter 'e' such that, for any $n \geq 0$, if $\tau_1, \ldots, \tau_n$ belong to the set, the n-tuple $\langle \tau_1, \ldots, \tau_n\rangle$ does too.

We call $\mathscr{R}$-typed collections ***relationally typed collections***. To define our languages we suppose we are given once and for all a certain $\mathscr{R}$-typed collection $\mathrm{Var}_{\mathscr{R}}$ of variables, with infinitely many members in each type. It doesn't matter what we take variables to be, so long as no variable is a string with multiple other variables as constituents. Let a ***relational signature*** be any non-overlapping $\mathscr{R}$-typed collection Σ such that no element of $\bigcup \Sigma$ is a variable or a string containing multiple variables or elements of $\bigcup \Sigma$.

For each relational signature Σ we define two languages, $\mathscr{K}^\Sigma$, the λK-language of Σ, and $\mathscr{I}^\Sigma$, the λI-language of Σ.

Definition 1.2. For any relational signature Σ, $\mathscr{I}^\Sigma$, the ***λI-language of*** Σ, is a function that maps each finite, non-overlapping $V \subseteq \mathrm{Var}_{\mathscr{R}}$ to a typed collection $\mathscr{I}^\Sigma(V)$—the 'terms of $\mathscr{I}^\Sigma$ whose free variables are exactly V'—minimally satisfying the following conditions:[87]

(i) $c \in \mathscr{I}^\Sigma(\varnothing)_\tau$ whenever $c \in \Sigma_\tau$.

(ii) $v \in \mathscr{I}^\Sigma(\{\langle v, \tau\rangle\})_\tau$ whenever $v \in \mathrm{Var}_\tau$.

(iii) $A(B_1, \ldots, B_n) \in \mathscr{I}^\Sigma(V^0 \cup V^1 \cup \cdots \cup V^n)_{\langle\rangle}$ whenever $A \in \mathscr{I}^\Sigma(V^0)_{\langle\tau_1,\ldots,\tau_n\rangle}$, $B_1 \in \mathscr{I}^\Sigma(V^1)_{\tau_1}$, …, $B_n \in \mathscr{I}^\Sigma(V^n)_{\tau_n}$, and $V^0 \cup V^1 \cup \cdots \cup V^n$ is non-overlapping.[88]

(iv) $(\lambda v_1 \ldots v_n.\varphi) \in \mathscr{I}^\Sigma(V - \{v_1, \ldots, v_n\})_{\langle\tau_1,\ldots,\tau_n\rangle}$ whenever $v_1 \in V_{\tau_1}$, …, and $v_n \in V_{\tau_n}$ $(n > 0)$ are any distinct variables, and $\varphi \in \mathscr{I}^\Sigma(V)_{\langle\rangle}$.

The definition of $\mathscr{K}^\Sigma$, the ***λK-language*** of Σ, is the same except for clause (iv), which is modified as follows to allow for vacuous λ-abstracts:

(iv′) $(\lambda v_1 \ldots v_n.\varphi) \in \mathscr{K}^\Sigma(V - \{v_1, \ldots, v_n\})_{\langle\tau_1,\ldots,\tau_n\rangle}$ whenever $v_1 \in \mathrm{Var}_{\tau_1}$, …, and $v_n \in \mathrm{Var}_{\tau_n}$ $(n > 0)$ are any distinct variables, and $\varphi \in \mathscr{K}^\Sigma(V)_{\langle\rangle}$.

A trivial induction shows that if $A \in \mathscr{K}^\Sigma(V)_\sigma$ and $A \in \mathscr{K}^\Sigma(V')_\tau$, $\bigcup V = \bigcup V'$: we call this the set of ***free variables*** of A, $FV(A)$. If we take Var to be non-overlapping, it follows that $\mathscr{K}^\Sigma(V)_\sigma$ is disjoint from $\mathscr{K}^\Sigma(V')_\tau$ for any $V' \neq V$; moreover, another straightforward induction shows that $\mathscr{K}^\Sigma(V)_\sigma$ is disjoint from $\mathscr{K}^\Sigma(V)_\tau$ whenever $\sigma \neq \tau$.[89]

When $\mathscr{L}$ is $\mathscr{I}^\Sigma$ or $\mathscr{K}^\Sigma$, a ***term*** of $\mathscr{L}$ is a member of $\mathscr{L}(V)_\tau$ for some V and τ; $\mathrm{wff}(\mathscr{L})$ is the typed collection such that $A \in \mathrm{wff}(\mathscr{L})_\tau$ iff $A \in \mathscr{L}(V)_\tau$ for some V. A ***pure*** term of $\mathscr{L}$ is a member of $\mathscr{I}^\varnothing(V)_\tau$ or $\mathscr{K}^\varnothing(V)_\tau$ for some V and τ; a ***closed*** term of $\mathscr{L}$ is a member of $\mathscr{L}(\varnothing)_\tau$ for some τ; a ***formula*** of $\mathscr{L}$ is a member of $\mathscr{L}(V)_{\langle\rangle}$ for some V.

[87] 'Minimally' means that for any f satisfying the conditions, $\mathscr{I}^\Sigma(V) \subseteq f(V)$ for every finite, non-overlapping $V \subseteq \mathrm{Var}$.

[88] The requirement that $V^0 \cup V^1 \cup \cdots \cup V^n$ is non-overlapping guarantees that $\mathscr{I}^\Sigma(V)$ will always be empty unless V is non-overlapping, which ensures that strings like $x(x)$ do not belong to $\mathscr{I}^\Sigma(V)_\tau$ for any V and τ.

[89] Both conditions can fail if Var is overlapping. For example, if $y \in \mathrm{Var}_\sigma \cap \mathrm{Var}_\tau$ and $x \in \mathrm{Var}_{\langle\sigma\rangle} \cap \mathrm{Var}_{\langle\tau\rangle}$, $\lambda xy.x(y)$ is in both $\mathscr{I}^\Sigma(\varnothing)_{\langle\langle\sigma\rangle,\sigma\rangle}$ and $\mathscr{I}^\Sigma(\varnothing)_{\langle\langle\tau\rangle,\tau\rangle}$, while $x(y)$ is in both $\mathscr{I}^\Sigma(\{\langle x, \langle\sigma\rangle\rangle, \langle y, \sigma\rangle\})_{\langle\rangle}$ and $\mathscr{I}^\Sigma(\{\langle x, \langle\tau\rangle\rangle, \langle y, \tau\rangle\})_{\langle\rangle}$. The approach where Var is completely overlapping is called "Curry-style" typing, while the approach where Var is non-overlapping is "Church-style": see HS, chapters 10 and 12.

A2 Functional types and functionally typed languages

Most work in higher-order logic uses *functional* languages, in which each complex term has exactly *two* terms as immediate constituents, rather than relational languages, in which a complex term can have any number of immediate constituents. Relational higher-order languages are arguably more metaphysically perspicuous: by treating all the arguments of a polyadic predication as syntactically on a par, they avoid introducing a kind of asymmetry in the notation that does not intuitively correspond to any asymmetry in the metaphysics. Moreover, they have the major advantage that they can be straightforwardly extended to allow for predicates of transfinite adicity.[90] But functional languages are more readable and more convenient for proving things about. In this section I will introduce a certain family of functional languages, and explain the sense in which they are equivalent to the relational languages from §A1.

Definition 2.1. $\mathscr{F}$—the set of ***functional types***—is the smallest set containing the letters 'e' (the type of objects) and 't' (the propositional type—think 'truth evaluable', not 'truth value'), such that whenever σ and τ belong to it and $\tau \neq e$, the ordered pair $\langle \sigma, \tau \rangle$—which we write as $(\sigma \to \tau)$—belongs to it.

A *terminal type* is a functional type other than e; a *complex type* is a functional type other than e or t; a *basic type* is e or t.

Following a standard convention, when talking about functional types we will omit parentheses—they are to be restored from the right, so for example $\rho \to \sigma \to \tau$ abbreviates $(\rho \to (\sigma \to \tau))$. I will use σ, ρ to range over all functional types, while τ always stands for terminal types.

We inductively define functions $\cdot^{\mathscr{R}}$ from $\mathscr{R}$ to $\mathscr{F}$, and $\cdot^{\mathscr{F}}$ from $\mathscr{F}$ to $\mathscr{R}$, as follows:

$$e^{\mathscr{R}} = e \qquad\qquad e^{\mathscr{F}} = e$$

$$\langle\rangle^{\mathscr{R}} = t \qquad\qquad t^{\mathscr{F}} = \langle\rangle$$

$$\langle \tau_0, \ldots, \tau_n \rangle^{\mathscr{R}} = ({\tau_0}^{\mathscr{R}} \to \langle \tau_1, \ldots, \tau_n \rangle^{\mathscr{R}}) \quad (n \geq 0) \qquad (\sigma \to \tau)^{\mathscr{F}} = \langle \sigma^{\mathscr{F}} \rangle^\frown \tau^{\mathscr{F}}$$

where $^\frown$ is concatenation of tuples. It is easy to show that $\cdot^{\mathscr{F}}$ and $\cdot^{\mathscr{R}}$ are bijections and mutual inverses.

Given these functions, we can turn any relationally typed collection $\mathscr{C}$ into a functionally typed collection $\mathscr{C}^{\mathscr{F}} = \{\langle x, \tau^{\mathscr{F}} \rangle : \langle x, \tau \rangle \in \mathscr{C}\}$, and similarly we can turn a functionally typed collection $\mathscr{D}$ into a relationally typed collection $\mathscr{D}^{\mathscr{R}}$.

[90] Thanks to Peter Fritz for pointing this out.

Note that it is crucial to this result that $\mathscr{F}$ does not contain any types of the form $\tau \to e$: if we had allowed for these, we would have something richer than $\mathscr{R}$. We can of course introduce function symbols into the language using Russell-style contextual definition; this will turn quantification into function-symbol position into a notational variant of quantification into dyadic predicate position restricted by functionality. But unless we favour a very coarse-grained logic in the original language, the logical behaviour of the stipulatively extended language will likely be disunified in a way that makes it worse for the purposes of stating general schemas than the original. This is especially clear if we reject *Plenitude* (see §6), since in that case we will think that, for example, quantification into type $t \to t$ works quite differently from quantification into type $t \to t \to t$ restricted by functionality.

To define functional languages, we take $\mathrm{Var}_{\mathscr{F}}$ to be $\mathrm{Var}_{\mathscr{R}^{\mathscr{F}}}$. A ***functionally typed signature*** is any nonoverlapping functionally typed collection Σ that does not contain any variables, or strings consisting of multiple variables or elements of Σ.

Definition 2.2 (Functional languages)**.** For any functionally typed signature Σ, $\mathscr{I}^{\Sigma}$, the ***λI-language of Σ***, is a function that maps each finite nonoverlapping $V \subseteq \mathrm{Var}_{\mathscr{F}}$ to a functionally typed collection $\mathscr{I}^{\Sigma}(V)$ (the 'terms of $\mathscr{I}^{\Sigma}$ whose free variables are exactly V'), minimally satisfying the following conditions:

(i) $c \in \mathscr{I}^{\Sigma}(\varnothing)_{\sigma}$ whenever $c \in \Sigma_{\sigma}$.

(ii) $v \in \mathscr{I}^{\Sigma}(\{\langle v, \sigma\rangle\})_{\sigma}$ whenever $v \in \mathrm{Var}_{\sigma}$.

(iii) $(AB) \in \mathscr{I}^{\Sigma}(V \cup V')_{\tau}$ whenever $A \in \mathscr{I}^{\Sigma}(V)_{\sigma\to\tau}$, $B \in \mathscr{I}^{\Sigma}(V')_{\sigma}$, and $V \cup V'$ is nonoverlapping.

(iv) $(\lambda v.A) \in \mathscr{I}^{\Sigma}(V - \{v\})_{\sigma\to\tau}$ whenever $v \in V_{\sigma}$ and $A \in \mathscr{I}^{\Sigma}(V)_{\tau}$.

The definition of $\mathscr{K}^{\Sigma}$, the ***λK-language of*** Σ, is the same except that clause (iv) is less restrictive:

(iv′) $(\lambda v.A) \in \mathscr{K}^{\Sigma}(V - \{v\})_{\sigma\to\tau}$ whenever $v \in \mathrm{Var}_{\sigma}$ and $A \in \mathscr{K}^{\Sigma}(V)_{\tau}$.

It follows that $\mathscr{I}^{\Sigma}(V)_{\tau} \subseteq \mathscr{K}^{\Sigma}(V)_{\tau}$ for every V and τ.

The concepts of a term, a closed term, a pure term, a formula, and of a formula's set of free variables apply to functional languages just as they did to relational languages. And it is still true that if $\mathrm{Var}_{\mathscr{F}}$ is non-overlapping, $\mathscr{K}^{\Sigma}(V)_{\sigma}$ is disjoint from $\mathscr{K}^{\Sigma}(V')_{\rho}$ whenever $V' \neq V$ or $\sigma \neq \rho$.[91]

[91]The present notation has the following annoying ambiguity: the empty set is officially both a functional and a relational signature, so the expressions '$\mathscr{I}^{\varnothing}$' and '$\mathscr{K}^{\varnothing}$' are ambiguous between the relational and functional languages with no constants. It won't be worth our while to try to resolve this.

We may omit parentheses in writing terms of functional languages; by contrast with the convention for types, they are to be restored from the left, so that ABC abbreviates $((AB)C)$. Also, when it is understood that terms A, B, and C are respectively of types $\rho \to \sigma \to \tau$, ρ, and σ, I will sometimes use infix notation, writing "$B\ A\ C$" for $((AB)C)$: for example, if A is a constant $\wedge$ of type $t \to t \to t$.

Having defined the four classes of languages, we now show how to translate between them. Here we assume that $\mathrm{Var}_{\mathscr{R}}$ is nonoverlapping so that every term has a unique type. First we define a function $\cdot^{\mathscr{F}}$ mapping the set of $\mathscr{K}^{\Sigma}$-terms for each relational signature Σ to the set of $\mathscr{K}^{(\Sigma^{\mathscr{F}})}$-terms, as follows:

Definition 2.3. When Σ is a relational signature and A is a term of $\mathscr{K}^{\Sigma}$, $A^{\mathscr{F}}$ is given by the following inductive definition:

(i) $a^{\mathscr{F}} = a$ when a is a constant or variable.

(ii) $A(B_1, B_2, \ldots, B_n)^{\mathscr{F}} = (A^{\mathscr{F}} {B_1}^{\mathscr{F}} {B_2}^{\mathscr{F}} \ldots {B_n}^{\mathscr{F}})$

(iii) $(\lambda v_1 v_2 \ldots v_n . A)^{\mathscr{F}} = (\lambda v_1 . \lambda v_2 \ldots \lambda v_n . A^{\mathscr{F}})$

It is straightforward to verify that $\cdot^{\mathscr{F}}$ maps $\mathscr{K}^{\Sigma}(V)_{\sigma}$ to $\mathscr{K}^{\Sigma^{\mathscr{F}}}(V^{\mathscr{F}})_{\sigma^{\mathscr{F}}}$ and $\mathscr{I}^{\Sigma}(V)_{\tau}$ to $\mathscr{I}^{\Sigma^{\mathscr{F}}}(V)_{\tau^{\mathscr{F}}}$.

The reverse translation function, from a functional language $\mathscr{K}^{\Sigma}$ to a relational language $\mathscr{K}^{\Sigma^{\mathscr{R}}}$, is slightly more involved:

Definition 2.4. When Σ is a functional signature and A is a term of $\mathscr{K}^{\Sigma}$, $A^{\mathscr{R}}$ is given by the following inductive definition:

(i) $a^{\mathscr{R}} = a$ when a is a constant or variable.

(ii) When $A \in \mathscr{K}^{\Sigma}_{\sigma_0 \to \cdots \to \sigma_n \to t}$ and $B \in \mathscr{K}^{\Sigma}_{\sigma_0}$ (for $n \geq 0$),

$$(AB)^{\mathscr{R}} = (\lambda v_1 \ldots v_n . A^{\mathscr{R}}(B^{\mathscr{R}}, v_1, \ldots, v_n))$$

where $v_1, \ldots, v_n$ are distinct variables of types ${\sigma_1}^{\mathscr{R}}, \ldots, {\sigma_n}^{\mathscr{R}}$ respectively, chosen according to some arbitrary order from among the variables not free in $A^{\mathscr{R}}$ or $B^{\mathscr{R}}$.

(iii) When $v_0 \in \mathrm{Var}$ and $A \in \mathscr{K}^{\Sigma}_{\sigma_1 \to \cdots \to \sigma_n \to t}$ (for $n \geq 0$),

$$(\lambda v_0 . A)^{\mathscr{R}} = (\lambda v_0 v_1 \ldots v_n . A^{\mathscr{R}}(v_1, \ldots, v_n))$$

where $v_1, \ldots, v_n$ are distinct variables of types ${\sigma_1}^{\mathscr{R}}, \ldots, {\sigma_n}^{\mathscr{R}}$ respectively, chosen according to some arbitrary order from among the variables not free in $A^{\mathscr{R}}$ and not identical to v_0.

(If we didn't want take $\mathrm{Var}_{\mathscr{R}}$ to be non-overlapping, the notion of translation would need to be relativised to a type σ and a non-overlapping typed collection of variables V, mapping $\mathscr{K}^{\Sigma}(V)_{\sigma}$ to $\mathscr{K}^{\Sigma^{\mathscr{F}}}(V^{\mathscr{F}})_{\sigma^{\mathscr{F}}}$ and vice-versa.)

Unlike the mapping from relational to functional types, the two translation functions on formulae are not mutual inverses. However, the result of translating a term back and forth always stands in an intimate syntactic relation to that term, namely that of βη-equivalence. The next order of business is to explain this.

Definition 2.5 (Capture-free simultaneous substitution)**.** Suppose that for a relational or functional signature Σ, π is a function mapping a set of variables and constants to a set of terms, such that whenever $\pi(a)$ is defined and $a \in \mathrm{Var}_{\sigma} \cup \Sigma_{\sigma}$, $\pi(a)$ is in $\mathrm{wff}(\mathscr{K}^{\Sigma})_{\sigma}$. We define the ***capture-free simultaneous substitution*** by π of a term A, $[\pi]A$, as follows.

$$
\begin{aligned}
[\pi]a &= \pi(a) \text{ when } \pi(a) \text{ is defined} \\
[\pi]a &= a \text{ when } \pi(a) \text{ is undefined} \\
[\pi](BC) &= ([\pi]B[\pi]C) \\
[\pi]B(C_1, \ldots, C_n) &= [\pi]B([\pi]C_1, \ldots, [\pi]C_n) \\
[\pi](\lambda v_1 \ldots v_n.B) &= (\lambda u_1 \ldots u_n.[\pi[u_i/v_i]]B)
\end{aligned}
$$

where for each i, $u_i = v_i$ if v_i is not free in $\pi(a)$ for any constant or free variable a in $\lambda v_1 \ldots v_n.B$, and otherwise u_i is a variable of the same type as v_i, not identical to any of $v_1 \ldots v_n$ or $u_1 \ldots u_{i-1}$, free in B, or free in $\pi(a)$ for any constant or free variable a in $\lambda v_1 \ldots v_n.B$, chosen according to some arbitrary ordering; and $\pi[u_i/v_i]$ is the substitution function like π except that it maps each v_i to u_i.

We write $[B/v]A$ for $[\{\langle v, B\rangle\}]A$, and $[B_i/v_i]A$ for $[\{\langle v_1, B_1\rangle, \ldots, \langle v_n, B_n\rangle\}]A$.

This gives us what we need to define the notions of $*$-reduction, $*$-equivalence, and $*$-normal forms, where '$*$' may be 'α', 'β', 'η', 'αβ', 'αη', 'βη', or 'αβη'. I will first give the definitions for functional languages, followed by those for relational languages in square brackets when they are different.

Definition 2.6. *A **immediately α-reduces*** to B iff A is an abstract $\lambda v.C$ and for some u not free in C, B is $\lambda u.[u/v]C$. [A is an abstract $\lambda v_1 \ldots v_n.\varphi$, and for some variables

$u_1 \ldots u_n$ each of which is either identical to v_i or not free in φ, B is $\lambda u_1 \ldots u_n.[u_i/v_i]\varphi.$][92]

*A **immediately β-reduces*** to B iff A is a "β-redex"—a term of the form $((\lambda v.C)D)$—and B is $[D/v]C$. [A is a term of the form $(\lambda v_1 \ldots v_n.\varphi)(D_1, \ldots, D_n)$ and B is $[D_i/v_i]\varphi.$]

*A **immediately η-reduces*** to B iff A is an "η-redex"—a term of the form $(\lambda v.(Cv))$ where v is not free in C—and B is C. [A is a term of the form $\lambda v_1 \ldots v_n.C(v_1, \ldots, v_n)$ where none of $v_1, \ldots, v_n$ is free in C, and B is C.]

*A **immediately *-reduces*** to B iff either α is in $*$ and A immediately α-reduces to B, or β is in $*$ and A immediately β-reduces to B, or η is in $*$ and A immediately η-reduces to B.

*A **one-step *-reduces*** to B iff B results from A by replacing one constituent with something to which it immediately $*$-reduces. That is: iff there are two finite sequences $\langle A_1, \ldots A_n\rangle, \langle B_1, \ldots, B_n\rangle$ such that A_1 immediately $*$-reduces to B_1, $A_n = A$, $B_n = B$, and whenever $0 < i < n$, either for some C, A_{i+1} is A_iC and B_{i+1} is B_iC or A_{i+1} is CA_i and B_{i+1} is CB_i, or else for some v, A_{i+1} is $\lambda v.A_i$ and B_{i+1} is $\lambda v.B_i$. [...either for some $C_0 \ldots C_m$, A_{i+1} is $A_i(C_0, \ldots, C_m)$ and B_{i+1} is $B_i(C_0, \ldots, C_m)$ or for some $0 < k \leq m$ A_{i+1} is $C_0(C_1, \ldots, C_k, A_i, C_{k+1}, \ldots, C_m)$ and B_{i+1} is $C_0(C_1, \ldots, C_k, A_i, C_{k+1}, \ldots, C_m)$, or else for some $v_1, \ldots, v_m$, A_{i+1} is $\lambda v_1 \ldots v_m.A_i$ and B_{i+1} is $\lambda v_1 \ldots v_m.B_i$.]

A is in ****-normal form*** iff it does not one-step $*$-reduce to anything.

A is ***one-step *-equivalent*** to B iff either A one-step $*$-reduces to B or B one-step $*$-reduces to A.

A ****-reduces*** to B iff there is a finite sequence of terms $\langle C_1, \ldots, C_n\rangle$ such that $A = C_1$ and $B = C_n$ and whenever $0 < i < n$, C_i one-step $*$-reduces to C_{i+1}.

A is ****-equivalent*** to B—for short, $A \approx_* B$—iff there is a finite sequence of terms $\langle C_1, \ldots, C_n\rangle$ such that $A = C_1$ and $B = C_n$ and whenever $0 < i < n$, C_i is one-step $*$-equivalent to C_{i+1}.

The following consequences of these definitions will be significant for us. We state them for functional languages; the extensions to relational languages are straightforward.

Proposition 2.7. If $A \approx_{\alpha\beta\eta} B$, then $A \approx_{\beta\eta} B$.

Proof. If A immediately α-reduces to B, there is a term C and variables u, v, with u not free in A, such that A is $\lambda v.C$ and B is $\lambda u.[u/v]C$; but then $\lambda u.(\lambda v.C)u$ immediately η-reduces to A and one-step β-reduces to B, hence $A \approx_{\beta\eta} B$. It follows that $A \approx_{\beta\eta} B$ whenever A one-step α-reduces to B, and hence whenever $A \approx_{\alpha\beta\eta} B$. □

[92] When A immediately α-reduces to B, B also immediately α-reduces to A, since if u is not free in C, v is not free in $[u/v]C$ and $C = [v/u][u/v]C$.

Proposition 2.8. If A $*$-reduces to B and $A \in \mathscr{K}^{\Sigma}(V)_{\sigma}$, then $B \in \mathscr{K}^{\Sigma}(V')_{\sigma}$ for some $V' \subseteq V$. Moreover, if $A \in \mathscr{I}^{\Sigma}(V)_{\sigma}$, $B \in \mathscr{I}^{\Sigma}(V)_{\sigma}$.

Proof. By an induction using the definition of substitution, we see that if $C \in \mathscr{K}^{\Sigma}(V)_{\sigma}$ and $D \in \mathscr{K}^{\Sigma}(V')_{\rho}$, $[C/v]D$ belongs to $\mathscr{K}^{\Sigma}(V \cup (V' - \{v\}))_{\rho}$ when $v \in V'$, and $\mathscr{K}^{\Sigma}(V' \setminus \{v\})_{\rho}$ otherwise. Since $(\lambda v.D)C$ also belongs to $\mathscr{K}^{\Sigma}(V \cup (V' - \{v\}))_{\rho}$, it follows that whenever $A \in \mathscr{K}^{\Sigma}(V)_{\sigma}$ immediately β-reduces to $B \in \mathscr{K}^{\Sigma}(V')_{\sigma}$, $V' \subseteq V$, with $V = V'$ in the case where A is of the form $(\lambda v.D)C$ where v has a free occurrence in D: this must be the case if $A \in \mathscr{I}^{\Sigma}(V)_{\sigma}$. It is also easy to show that if A immediately αη-reduces to B and $A \in \mathscr{K}^{\Sigma}(V)_{\sigma}$, $B \in \mathscr{K}^{\Sigma}(V)_{\sigma}$. Two straightforward inductions then establish the result, first for $*$-reduction in one step, and then in general. □

Corollary 2.9. If $A \approx_{*} B$ and $A \in \mathscr{I}^{\Sigma}(V)_{\sigma}$, $B \in \mathscr{I}^{\Sigma}(V)_{\sigma}$.

Proposition 2.10. $AB \approx_{*} CD$ whenever $A \approx_{*} B$ and $C \approx_{*} D$. and $\lambda v.A \approx_{*} \lambda v.B$ whenever $A \approx_{*} B$.

Proof. If $A \approx_{*} B$ and $C \approx_{*} D$, then there exist sequences $\langle A_1, \ldots, A_n\rangle, \langle C_1, \ldots, C_m\rangle$ such that $A = A_1$, $B = A_n$, $C = C_1$, and $D = C_n$, and each $A_i \approx_{*} A_{i+1}$ and $C_i \approx_{*} C_{i+1}$. Then $\langle A_1B_1, A_2B_1, \ldots, A_nB_1, A_nB_2, \ldots, A_nB_m\rangle$ witnesses the $*$-equivalence of AB and CD. Similarly, if $\langle A_1, \ldots, A_n\rangle$ witnesses the $*$-equivalence of A and B, $\langle \lambda v.A_1, \ldots, \lambda v.A_n\rangle$ witnesses the $*$-equivalence of $\lambda v.A$ and $\lambda v.B$: by Corollary 2.9 these are all well-formed so long as $\lambda v.A$ is. □

For proofs of the following theorems, which will be important later, see HS, appendices A–C.

Proposition 2.11 (Substitution)**.** If $A \approx_{\alpha *} B$ and for each i, $C_i \approx_{\alpha *} D_i$, then $[C_i/v_i]A \approx_{\alpha *} [D_i/v_i]B$.[93]

Proposition 2.12 (Strong normalisation)**.** Every term of $\mathscr{K}^{\Sigma}$ $*$-reduces to at least one term in $*$-normal form.

Proposition 2.13 (Church-Rosser)**.** If A $*$-reduces to both B and C, then for some B', C', B $*$-reduces to B', and C $*$-reduces to C', and B' is α-equivalent to C'.

Corollary 2.14. If A $*$-reduces to B, and B and C are both in $*$-normal form, then B is α-equivalent to C.

We can now precisely characterise the way in which our two translation functions are in harmony.

[93] This holds for α$*$-reduction as well as α$*$-equivalence.

Proposition 2.15. When Σ is a functional signature, $A^{\mathscr{R}\mathscr{F}} \approx_\eta A$ for any term A of the functional language $\mathscr{K}^\Sigma$.

Proof. By induction on the complexity of A. If A is a variable or constant, $A^{\mathscr{R}\mathscr{F}}$ is identical to A and thus trivially η-equivalent to A. If A is BC, then for some zero or more distinct variables $v_1, \ldots, v_n$ not free in B or C, $A^{\mathscr{R}\mathscr{F}} = (\lambda v_1 \ldots v_n . B^{\mathscr{R}}(C^{\mathscr{R}}, v_1, \ldots, v_n))^{\mathscr{F}} = \lambda v_1 \ldots \lambda v_n . B^{\mathscr{R}\mathscr{F}} C^{\mathscr{R}\mathscr{F}} v_1 \ldots v_n$. This η-reduces in n steps to $B^{\mathscr{R}\mathscr{F}} C^{\mathscr{R}\mathscr{F}}$, which is η-equivalent to BC by Proposition 2.10 and the induction hypothesis. Finally if A is $\lambda v_0 . B$, then for some zero or more distinct variables $v_1, \ldots, v_n$ not free in B, $A^{\mathscr{R}\mathscr{F}} = (\lambda v_0 v_1 \ldots v_n . B^{\mathscr{R}}(v_1, \ldots, v_n))^{\mathscr{F}} = \lambda v_0 . \lambda v_1 \ldots \lambda v_n . B^{\mathscr{R}\mathscr{F}} v_1 \ldots v_n$. This η-reduces in n steps to $\lambda v_0 . B^{\mathscr{R}\mathscr{F}}$, which is η-equivalent to $\lambda v_0 . B$ by Proposition 2.10 and the induction hypothesis. □

For the analogous result in the other direction, we first need a couple of lemmas.

Lemma 2.16. Whenever A_0 is a functional term of type $\sigma_1 \to \cdots \to \sigma_m \to t$ and for some positive $n \leq m$ $A_1, \ldots, A_n$ are terms of types $\sigma_1, \ldots, \sigma_n$,

$$(A_0 \ldots A_n)^{\mathscr{R}} \approx_\beta \lambda v_{n+1} \ldots v_m . A_0{}^{\mathscr{R}}(A_1{}^{\mathscr{R}}, \ldots, A_n{}^{\mathscr{R}}, v_{n+1}, \ldots, v_m)$$

where $v_{n+1} \ldots v_m$ are zero or more distinct variables, of types $\sigma_{n+1} \ldots \sigma_m$ respectively, not free in $A_0 \ldots A_n$.

Proof. By induction on n. Base case: $n = 1$. $(A_0 A_1)^{\mathscr{R}}$ is $\lambda v_2 \ldots v_m . A_0{}^{\mathscr{R}}(A_1{}^{\mathscr{R}}, v_2, \ldots, v_m)$ for some appropriate $v_2 \ldots v_m$ by definition. Induction step: by definition, for some appropriate $v_{n+2} \ldots v_m$, $(A_0 \ldots A_{n+1})^{\mathscr{R}} = \lambda v_{n+2} \ldots v_m . (A_0 \ldots A_n)^{\mathscr{R}}(A_{n+1}{}^{\mathscr{R}}, v_{n+2}, \ldots, v_m)$, which by the induction hypothesis is βη-equivalent to

$$\lambda v_{n+2} \ldots v_m . (\lambda u_{n+1} \ldots u_m . A_0{}^{\mathscr{R}}(A_1{}^{\mathscr{R}}, \ldots, A_n{}^{\mathscr{R}}, u_{n+1}, \ldots, u_m))(A_{n+1}{}^{\mathscr{R}}, v_{n+2}, \ldots, v_m)$$

for some appropriate $u_{n+1} \ldots u_m$. This β-reduces in one step to $\lambda v_{n+2} \ldots v_m . A_0{}^{\mathscr{R}}(A_1{}^{\mathscr{R}}, \ldots, A_{n+1}{}^{\mathscr{R}}, v_{n+2}$

Lemma 2.17. Whenever $v_1 \ldots v_n$ (n>0) are distinct variables of types $\sigma_1 \ldots \sigma_n$ and A is a functional term of type $\sigma_{n+1} \to \cdots \to \sigma_m \to t$ (where $m \geq n$), $(\lambda v_1 \ldots \lambda v_n . A)^{\mathscr{R}} \approx_{\alpha\beta} \lambda v_1 \ldots v_m . A^{\mathscr{R}}(v_{n+1}, \ldots, v_m)$ where $v_{n+1} \ldots v_m$ are any (zero or more) variables of types $\sigma_{n+1} \ldots \sigma_m$, not free in A.

Proof. By induction on n. Base case: $n = 1$; then by definition, $(\lambda v_1 . A)^{\mathscr{R}}$ is $\lambda v_1 \ldots v_m . A^{\mathscr{R}}(v_1, \ldots, v_m$ for appropriate $v_2, \ldots, v_m$. Induction step: by definition, $(\lambda v_1 \ldots \lambda v_{n+1} . A)^{\mathscr{R}}$ is

$$\lambda v_1 u_2 \ldots u_m . (\lambda v_2 \ldots \lambda v_{n+1} . A)^{\mathscr{R}}(u_2, \ldots, u_m)$$

for some appropriate $u_2 \ldots u_m$, which by the induction hypothesis is β-equivalent to

$$\lambda v_1 u_2 \ldots u_m . (\lambda v_2 \ldots v_m . A^{\mathscr{R}}(v_{n+2} \ldots v_m))(u_2, \ldots, u_m)$$

This β-reduces in one step to $\lambda v_1 u_2 \ldots u_m.([u_i/v_i]A^{\mathscr{R}}(u_{n+2}\ldots u_m))$, which is α-equivalent to $\lambda v_1 \ldots v_m.(A^{\mathscr{R}}(v_{n+2}\ldots v_m))$. □

Proposition 2.18. When Σ is a relational signature, $A^{\mathscr{F}\mathscr{R}} \approx_{\alpha\beta} A$ for any term A of the relational language $\mathscr{K}^{\Sigma}$.

Proof. By induction on the complexity of A. When A is a variable or constant, $A^{\mathscr{F}\mathscr{R}} = A$. When A is $B(C_1, \ldots, C_n)$, $A^{\mathscr{F}}$ is $B^{\mathscr{F}} C_1{}^{\mathscr{F}} \ldots C_n{}^{\mathscr{F}}$, so $A^{\mathscr{F}\mathscr{R}}$ is $(B^{\mathscr{F}} C_1{}^{\mathscr{F}} \ldots C_n{}^{\mathscr{F}})^{\mathscr{R}}$, which is β-equivalent to $B^{\mathscr{F}\mathscr{R}}(C_1{}^{\mathscr{F}\mathscr{R}}, \ldots, C_n{}^{\mathscr{F}\mathscr{R}})$ by the $n = m$ case of Lemma 2.16. When A is an abstract $\lambda v_1 \ldots v_n.B$, $A^{\mathscr{F}}$ is $\lambda v_1 \ldots \lambda v_n.B^{\mathscr{F}}$, so $A^{\mathscr{F}\mathscr{R}}$ is $\lambda v_1 \ldots \lambda v_n.B^{\mathscr{F}\mathscr{R}}$, which is αβ-equivalent to $\lambda v_1 \ldots v_n.B^{\mathscr{F}\mathscr{R}}$ by the $n = m$ case of Lemma 2.17. □

The upshot of Propositions 2.15 and 2.18 is that so long as our logic licenses treating βη-equivalent terms as interchangeable, the choice whether to theorise in a relational or a functional language is just a matter of taste. Moreover, when we are studying properties of languages that do not distinguish βη-equivalent terms, the choice whether to theorise *about* a relational or a functional language is also a matter of taste, since every definition and result about the one will have an analogue for the other. In the rest of the appendix, we will be theorising about functional languages.

A3 Structures and models

We now turn from syntax to semantics. Our models for a given language $\mathscr{L} = \mathscr{I}^{\Sigma}$ or $\mathscr{K}^{\Sigma}$ will consist of an $\mathscr{L}$-structure—a domain for each type together with an interpretation of the language on those domains—together with a *valuation* which assigns truth values to elements of the propositional domain. (The material in this section is largely drawn from BBK: my "$\mathscr{K}^{\Sigma}$-structures" are, roughly, their "η-functional Σ-evaluations".)

Definition 3.1. When $\mathscr{D}$ is a typed collection, an ***assignment*** for $\mathscr{D}$ is any typed function from some nonoverlapping typed collection of variables V to $\mathscr{D}$. Two assignments for $\mathscr{D}$ are ***compatible*** iff they agree on the intersection of their domains and the union of their domains is nonoverlapping.

Definition 3.2. Given a functional language $\mathscr{L} = \mathscr{I}^{\Sigma}$ or $\mathscr{K}^{\Sigma}$, an ***$\mathscr{L}$-structure*** is a pair $\mathfrak{S} = \langle \mathscr{D}, [\![\cdot]\!] \rangle$ where $\mathscr{D}$ is a typed collection and $[\![\cdot]\!]$ is a function which maps every assignment function g for $\mathscr{D}$ to a typed function $[\![\cdot]\!]^g$ from $\mathscr{L}(\mathrm{dom}\, g)$ to $\mathscr{D}$, subject to the following conditions:

(i) $[\![v]\!]^{v \mapsto \mathbf{x}}_{\sigma} = \mathbf{x}$. (Equivalently: $[\![v]\!]^g_{\sigma} = g_{\sigma}(v)$ whenever defined.)

(ii) If $[\![A]\!]^{g_1}_{\sigma\to\tau} = [\![C]\!]^{g_2}_{\sigma\to\tau}$ and $[\![B]\!]^{g_3}_{\sigma} = [\![D]\!]^{g_4}_{\sigma}$, then $[\![AB]\!]^{g_1\cup g_2}_{\tau} = [\![CD]\!]^{g_3\cup g_4}_{\tau}$ if g_1 and g_2 are compatible and g_3 and g_4 are compatible.

(iii) $[\![A]\!]^{g}_{\sigma} = [\![B]\!]^{h}_{\sigma}$ whenever $A \approx_{\beta\eta} B$, g and h are compatible, and both sides are defined.[94]

When A is closed, we write $[\![A]\!]_{\sigma}$ instead of $[\![A]\!]^{\varnothing}_{\sigma}$. For convenience, I will generally assume when discussing $\mathscr{L}$-structures that $\mathscr{D}$ and Var are both nonoverlapping, and that $\bigcup\mathscr{D}$ is disjoint from $\bigcup \mathrm{wff}(\mathscr{L})$. I use boldface variables $\mathbf{x}, \mathbf{y}, \mathbf{z}\ldots$ when talking about the elements of $\bigcup\mathscr{D}$. This enables a useful convention where, for example, given $\mathbf{x} \in \mathscr{D}_{\rho\to\sigma\to\tau}$, $\mathbf{y} \in \mathscr{D}_{\rho}$ and $\mathbf{z} \in \mathscr{D}_{\sigma}$, we can write $[\![\mathbf{xyz}]\!]$ instead of $[\![uvw]\!]^{[u\mapsto\mathbf{x},v\mapsto\mathbf{y},w\mapsto\mathbf{z}]}_{\tau}$, or $[\![\mathbf{x}yz]\!]^{g}$ instead of $[\![uyw]\!]^{g\cup[u\mapsto\mathbf{x},w\mapsto\mathbf{z}]}_{\tau}$. In general, occurrences of boldfaced symbols in expressions inside $[\![\cdot]\!]$ function as abbreviations of variables not otherwise in use, assigned to the relevant element of the domain.[95]

(Note that the definition of an $\mathscr{L}$-structure did not require the domain $\mathscr{D}$ to be populated. When $\mathscr{D}_{\sigma}$ is empty and V_{σ} is nonempty, there are no assignment functions from V to $\mathscr{D}$. In that case, when $A \in \mathscr{L}(V)_{\rho}$, there will be no g for which $[\![A]\!]^{g}_{\rho}$ is defined. But this does not stop $[\![\cdot]\!]$ from being non-trivial for terms with *bound* variables of type σ. If we had taken the more customary approach of assigning denotations to terms relative to variable assignments defined on the entirety of Var, we could not have allowed for nontrivial $\mathscr{L}$-structures without populated domains.)

We can have $\mathscr{I}^{\Sigma}$- and $\mathscr{K}^{\Sigma}$-structures for arbitrary signatures Σ, including $\varnothing$. However, to make a *model*, we will need a an $\mathscr{I}^{\Sigma}$-structure or $\mathscr{K}^{\Sigma}$-structure where Σ contains the logical constants. We will call such signatures "logical".

Definition 3.3. Log is the functional signature containing just the following constants: $\neg \in \mathrm{Log}_{t\to t}$, $\wedge \in \mathrm{Log}_{t\to t\to t}$, $\vee \in \mathrm{Log}_{t\to t\to t}$, and for every type σ, $\forall_{\sigma} \in \mathrm{Log}_{(\sigma\to t)\to t}$ and $\exists_{\sigma} \in \mathrm{Log}_{(\sigma\to t)\to t}$. A functional signature Σ is ***logical*** iff $\mathrm{Log} \subseteq \Sigma$. A functional language $\mathscr{L}$ is logical if it is $\mathscr{I}^{\Sigma}$ or $\mathscr{K}^{\Sigma}$ for some logical signature Σ.

[94] In an $\mathscr{I}$-language, βη-equivalent terms must have the same free variables (Proposition 2.8), so if $[\![A]\!]^{g}_{\sigma}$ and $[\![B]\!]^{h}_{\sigma}$ are both defined, g and h must have the same domain, and so are identical if compatible.

[95] Another way to look at this convention is that, so long as $\mathscr{D}$ is nonoverlapping and disjoint from Σ and Var, any $\mathscr{I}^{\Sigma}$-structure or $\mathscr{K}^{\Sigma}$-structure $\langle\mathscr{D}, [\![\cdot]\!]\rangle$ can be naturally extended to an $\mathscr{I}^{\Sigma\cup\mathscr{D}}$-structure or $\mathscr{K}^{\Sigma\cup\mathscr{D}}$-structure $\langle\mathscr{D}, [\cdot]\rangle$, by the stipulation that that $[A]^{g}_{\sigma} = [\![A]\!]^{g}_{\sigma}$ whenever $[\![A]\!]^{g}_{\sigma}$ is defined, and $[\![\mathbf{x}]\!]_{\sigma} = \mathbf{x}$ for any $\mathbf{x}$ in $\mathscr{D}_{\sigma}$. (Proposition 5.4 below shows that this uniquely determines $[\cdot]$.) Going further in this direction, we could modify the definition of an $\mathscr{L}$-structure so that $[\![\cdot]\!]$ only interprets closed terms of $\mathscr{I}^{\Sigma\cup\mathscr{D}}$ or $\mathscr{K}^{\Sigma\cup\mathscr{D}}$, and define the assignment-relative notion of denotation for open terms by treating the assignment function as specifying a substitution function that replaces free variables with constants from $\mathscr{D}$.

In writing terms in logical languages we use the following metalinguistic abbreviations:

$$\to =_{\text{df}} \lambda p^t.\lambda q^t.\neg p \vee q \qquad \leftrightarrow =_{\text{df}} \lambda p^t.\lambda q^t.(p \to q) \wedge (q \to p)$$
$$\equiv_\tau =_{\text{df}} \lambda x^\tau.\lambda y^\tau.\forall_\tau(\lambda z^{\tau\to t}.zx \leftrightarrow zy) \qquad \not\equiv_\tau =_{\text{df}} \lambda x^\tau.\lambda y^\tau.\neg(x \equiv_\tau y)$$

We also write $\forall v^\sigma(\varphi)$ for $\forall_\sigma(\lambda v^\sigma.\varphi)$ and $\exists v^\sigma(\varphi)$ for $\exists_\sigma(\lambda v^\sigma.\varphi)$.

A model, then, will be the result of supplementing a logical $\mathscr{L}$-structure with a *valuation*, which assigns truth values to elements of the propositional domain in accordance with certain constraints.

Definition 3.4. When $\mathscr{L}$ is logical, an $\mathscr{L}$***-model*** is a triple $\langle \mathscr{D}, [\![\cdot]\!], |\cdot| \rangle$ where $\langle \mathscr{D}, [\![\cdot]\!] \rangle$ is an $\mathscr{L}$-structure and $|\cdot|$ is a function $\mathscr{D}_t \to \{0, 1\}$ such that

(i) For any $\mathbf{p} \in \mathscr{D}_t$, $|[\![\neg\mathbf{p}]\!]| = 1 - |\mathbf{p}|$.

(ii) For any $\mathbf{p}, \mathbf{q} \in \mathscr{D}_t$, $|[\![\mathbf{p} \vee \mathbf{q}]\!]| = \max\{|\mathbf{p}|, |\mathbf{q}|\}$.

(iii) For any $\mathbf{p}, \mathbf{q} \in \mathscr{D}_t$, $|[\![\mathbf{p} \wedge \mathbf{q}]\!]| = \min\{|\mathbf{p}|, |\mathbf{q}|\}$.

(iv) For any $\mathbf{x} \in \mathscr{D}_{\sigma\to t}$, $|[\![\exists_\sigma \mathbf{x}]\!]| = \max\{|[\![\mathbf{xy}]\!]| : \mathbf{y} \in \mathscr{D}_\sigma\}$.

(v) For any $\mathbf{x} \in \mathscr{D}_{\sigma\to t}$, $|[\![\forall_\sigma \mathbf{x}]\!]| = \min\{|[\![\mathbf{xy}]\!]| : \mathbf{y} \in \mathscr{D}_\sigma\}$.

We can generalise $|\cdot|$ to elements of $\mathscr{D}_\tau$ for all complex types τ by defining $|\mathbf{x}|$, for any $\mathbf{x} \in \mathscr{D}_{\sigma\to\tau}$, to be the function with domain $\mathscr{D}_\sigma$ such that, for any $\mathbf{y} \in \mathscr{D}_\sigma$, $|\mathbf{x}|(\mathbf{y}) = |[\![\mathbf{xy}]\!]|$. We call $|\mathbf{x}|$ the ***extension*** of $\mathbf{x}$.

As usual, validity is defined in terms of truth (value 1) in a model:

Definition 3.5. A formula $\varphi \in \mathscr{L}(V)_t$ is ***valid*** on a class of models C iff for every model $\langle \mathscr{D}, [\![\cdot]\!], |\cdot| \rangle \in C$ and $g \in \mathscr{D}^{[V]}$, $|[\![\varphi]\!]^g| = 1$.

Where Γ and Δ are any sets of $\mathscr{L}$-formulae, the sequent $\Gamma \Rightarrow \Delta$ is valid on a class of models C iff there is no model $\langle \mathscr{D}, [\![\cdot]\!], |\cdot| \rangle \in C$ and function f mapping every $\varphi \in \Gamma \cup \Delta$ to an assignment function $f(\varphi)$ such that $\varphi \in \mathscr{L}(\mathrm{dom}\, f(\varphi))_t$, such that any two of these assignment functions are compatible, $|[\![\varphi]\!]^{f(\varphi)}| = 1$ for every $\varphi \in \Gamma$, and $|[\![\varphi]\!]^{f(\varphi)}| = 0$ for every $\varphi \in \Delta$.

The central reason to care about models in the sense of Definition 3.4 comes from the soundness and completeness theorems proved by BBK, which imply that the standard classical rules of (multi-sorted) quantification theory, supplemented by βη-conversion, are sound and complete for the class of all populated models (i.e. models with populated domains). As usual with such results, there is flexibility as

regards exactly what proof system we have in mind; we will have no need to be more specific here. BBK focus on a natural deduction calculus $\mathfrak{NK}_{\beta\eta}$, whose rules are the standard classical natural deduction rules for $\neg$, $\vee$, and $\forall$ (they treat $\wedge$ and $\exists$ as defined), together with rules allowing one to derive the sequent $\Gamma \Rightarrow \{\psi\}$ from $\Gamma \Rightarrow \{\psi\}$ whenever ψ is β-equivalent or η-equivalent to Φ. In this setting, the result is that a sequent of the form $\Gamma \Rightarrow \{\varphi\}$ is valid in the class of all models iff there is a derivation of it in $\mathfrak{NK}_{\beta\eta}$.[96] Although their result concerns only $\mathscr{K}$-models, the proof goes through with essentially no modification for $\mathscr{I}$-models.[97] Moreover, some of the proof procedures that are sound and complete for the class of all populated models need only slight amendment to be sound and complete for the class of all models.[98]

For those of us whose interest in the model theory is driven by metaphysics, the primary importance of the soundness theorem comes from the fact that, by finding a mathematical proof that there is a model of a certain proposed theory in a higher-order language, we can assure ourselves that we will not end up with a contradiction if we endorse that theory and reason in accordance with the classical inference rules and $\beta\eta$-conversion. The importance of the completeness theorem, meanwhile, comes from the guarantee it provides that, in searching for models (in the sense of Definition 3.4) of a theory we like, we are not unnecessarily limiting our search space in a way that would deprive us of the assurance that the discovery of a model would provide. Of course, this kind of assurance is only worth so much, since there are many ways of objecting to theories that don't require deriving contradictions from them using the classical rules and $\beta\eta$-conversion. But investigating the class of models in which a theory is true can also be helpful in other ways when assessing its credibility. For example, since most of us are better at informal reasoning about mathematics than at producing valid arguments in formal languages, exploring models can help us to find deductive consequences of the theory (which may turn out to

[96]Because derivations in $\mathfrak{NK}_{\beta\eta}$ consist of sequents of closed sentences, BBK's result requires a restriction to require the signature Σ contains a sufficiently large infinite supply of constants in each type that do not occur in Γ. We can get around this by allowing free variables to occur in derivations, although we will still need a restriction to rule out the case where Γ is so enormous that (in some type) the set of variables free in Γ is larger than the set of variables not free in Γ.

[97]Note that when $\mathscr{K}$-formulae are allowed to occur in derivations, we can prove $\mathscr{I}$-formulae which are not provable when derivations are required to contain only $\mathscr{I}$-formulae. For example, in the $\mathscr{K}$-language, we can prove $p \equiv (\lambda q.p)p$ from the theorem $p \equiv p$ using β-conversion, and then apply the introduction rules for the quantifiers to derive $\forall p^t \exists x^{t \to t} \forall q^t (p \equiv xq)$, an $\mathscr{I}$-formula which cannot be proved by a derivation consisting entirely of $\mathscr{I}$-formulae.

[98]For example, in a sequent calculus, we could achieve this by changing the rules $\exists$R and $\forall$L to disallow steps which take us from a sequent in which some variable occurs free to a sequent in which that variable does not occur free.

be attractive or objectionable); and thanks to the completeness theorem, it can also help us identify questions that the theory does not deductively resolve (which may be praiseworthy open-mindedness or problematic weakness).[99]

A4 Varieties of structures and models

In this section we will distinguish some interesting subclasses of structures and models. This will help to illuminate why the definitions in A3 look the way they do, and to further refine our sense of what we should be hoping for when looking for models of a theory. Most of these definitions are standard; here again my discussion draws heavily on BBK.

As we go on, it will be helpful to be able to apply some standard algebraic terminology to $\mathscr{L}$-structures and $\mathscr{L}$-models. For future reference, the definitions are as follows:

Definition 4.1. A ***homomorphism*** from an $\mathscr{L}$-structure $\langle \mathscr{D}, [\![\cdot]\!] \rangle$ to an $\mathscr{L}$-structure $\langle \mathscr{D}', [\cdot] \rangle$ is a typed function f from $\mathscr{D}$ to $\mathscr{D}'$ such that for any $A \in \mathscr{L}(V)_\sigma$ and $g \in \mathscr{D}^{[V]}$, $[A]_\sigma^{f \circ g} = f([\![A]\!]_\sigma^g)$. A homomorphism from an $\mathscr{L}$-model $\langle \mathscr{D}, [\![\cdot]\!], |\cdot| \rangle$ to an $\mathscr{L}$-model $\langle \mathscr{D}', [\cdot], \|\cdot\| \rangle$ is a homomorphism f from $\langle \mathscr{D}, [\![\cdot]\!] \rangle$ to $\langle \mathscr{D}', [\cdot] \rangle$ such that for any $\mathbf{p} \in \mathscr{D}_t$, $\|f(\mathbf{p})\| = |\mathbf{p}|$.

An ***isomorphism*** is a homomorphism that is bijective.[100]

An $\mathscr{L}$-structure $\langle \mathscr{D}, [\![\cdot]\!] \rangle$ is a ***substructure*** of an $\mathscr{L}$-structure $\langle \mathscr{D}', [\cdot] \rangle$ if $\mathscr{D} \subseteq \mathscr{D}'$ and $[\![A]\!]_\sigma^g = [A]_\sigma^g$ whenever $A \in \mathscr{L}(V)_\sigma$ and $g \in \mathscr{D}^{[V]}$. $\langle \mathscr{D}, [\![\cdot]\!], |\cdot| \rangle$ is a ***submodel*** of $\langle \mathscr{D}', [\cdot], \|\cdot\| \rangle$ iff $\langle \mathscr{D}, [\![\cdot]\!] \rangle$ is a substructure of $\langle \mathscr{D}', [\cdot] \rangle$ and $|\mathbf{p}| = \|\mathbf{p}\|$ whenever $\mathbf{p} \in \mathscr{D}_t$.

A ***congruence*** on an $\mathscr{L}$-structure $\langle \mathscr{D}, [\![\cdot]\!] \rangle$ is a typed family $\sim$ where $\sim_\sigma$ is an equivalence relation on $\mathscr{D}_\sigma$ for each σ, and for any assignment functions $g, h \in \mathscr{D}^{[V]}$

[99] Those who reject βη-conversion or the classical logic of truth-functional connectives and quantifiers will want to find a broader definition of "model" if they want to pursue similar investigations. If you like the classical rules but not βη-conversion, see Muskens (2007) for a conception of "model" so broad that it does not even require α-equivalent formulae to agree in truth value; it is equivalent to the result of imposing a certain weakening on clause (iii) in our definition of an $\mathscr{L}$-structure while adding a new clause to the definition of a model to require $|\cdot|$ to respect extensional β-conversion. If you like βη-conversion and the classical rules for truth-functional connectives, but are tempted to reject the classical rules for $\forall$ and $\exists$ (and $\equiv$), Bacon and J. S. Russell MS is a good starting point. If you want to weaken the classical rules for truth-functional connectives, there is a vast array of model-theoretic techniques used in the study of non-classical propositional logics which could be adapted to the present setting by taking over the role of the valuation $|\cdot|$ in the definition of a model.

[100] If f is an isomorphism from $\langle \mathscr{D}, [\![\cdot]\!] \rangle$ to $\langle \mathscr{D}', [\cdot] \rangle$, f^{-1} is an isomorphism from $\langle \mathscr{D}', [\cdot] \rangle$ to $\langle \mathscr{D}, [\![\cdot]\!] \rangle$, since $[\![A]\!]_\sigma^{f^{-1} \circ g} = f^{-1}(f([\![A]\!]_\sigma^{f^{-1} \circ g})) = f^{-1}([A]_\sigma^{f \circ f^{-1} \circ g}) = f^{-1}([A]_\sigma^g)$.

such that $g(v) \sim_\rho h(v)$ whenever $v \in V_\rho$, $[\![A]\!]^g_\sigma \sim_\sigma [\![A]\!]^h_\sigma$ for any $A \in \mathscr{L}(V)_\sigma$. A congruence on an $\mathscr{L}$-model $\langle \mathscr{D}, [\![\cdot]\!], |\cdot| \rangle$ is a congruence on $\langle \mathscr{D}, [\![\cdot]\!] \rangle$ with the further property that $|\mathbf{p}| = |\mathbf{q}|$ whenever $\mathbf{p} \sim_t \mathbf{q}$.

When $\sim$ is a congruence on an $\mathscr{L}$-structure $\langle \mathscr{D}, [\![\cdot]\!] \rangle$, the ***quotient of*** $\langle \mathscr{D}, [\![\cdot]\!] \rangle$ ***under*** $\sim$ is the $\mathscr{L}$-structure $\langle \mathscr{D}^\sim, [\![\cdot]\!]^\sim \rangle$, where $\mathscr{D}^\sim_\sigma$ is the set of equivalence classes of $\sim_\sigma$, and for any assignment g for $\mathscr{D}^\sim$ with domain V and $A \in \mathscr{L}(V)_\sigma$, $[\![A]\!]^{\sim g}_\sigma$ is the unique equivalence class that contains $[\![A]\!]^h_\sigma$ for some assignment $h \in \mathscr{D}^{[V]}$ such that $h_\rho(v) \in g_\rho(v)$ whenever $v \in V_\rho$. When $\sim$ is a congruence on an $\mathscr{L}$-model $\langle \mathscr{D}, [\![\cdot]\!], |\cdot| \rangle$, the quotient of the model under $\sim$ is the model $\langle \mathscr{D}^\sim, [\![\cdot]\!]^\sim, |\cdot|^\sim \rangle$, where for each $\mathbf{p} \in \mathscr{D}^\sim_t$, $|\mathbf{p}|^\sim$ is the unique $x \in \{0, 1\}$ such that $|\mathbf{q}| = x$ for some $\mathbf{q} \in \mathbf{p}$.

With these preliminaries out of the way, we can turn to two significant properties of $\mathscr{L}$-structures:

Definition 4.2. A $\mathscr{L}$-structure $\langle \mathscr{D}, [\![\cdot]\!] \rangle$ is ***functional*** if for any $\mathbf{x}, \mathbf{y} \in \mathscr{D}_{\sigma\to\tau}$ such that $\mathbf{x} \neq \mathbf{y}$, there is some $\mathbf{z} \in \mathscr{D}_\sigma$ such that $[\![\mathbf{xz}]\!] \neq [\![\mathbf{yz}]\!]$

Definition 4.3. A $\mathscr{L}$-structure $\langle \mathscr{D}, [\![\cdot]\!] \rangle$ is ***full*** if for any function f from $\mathscr{D}_\sigma$ to $\mathscr{D}_\tau$, there is at least one $\mathbf{x} \in \mathscr{D}_{\sigma\to\tau}$ such that for every $\mathbf{y} \in \mathscr{D}_\sigma$, $[\![\mathbf{xy}]\!] = f([\![\mathbf{y}]\!])$.

These definitions may be helpfully motivated using the concept of *applicative structure*. An ***applicative structure*** on a typed collection $\mathscr{D}$ is a collection of functions $@^{\sigma,\tau}$ where $@^{\sigma,\tau}$ maps $\mathscr{D}_{\sigma\to\tau}$ to $\mathscr{D}_\tau^{\mathscr{D}_\sigma}$. Any $\mathscr{L}$-structure naturally induces an applicative structure on its domain, setting $@^{\sigma,\tau}(\mathbf{x})(\mathbf{y}) = [\![\mathbf{xy}]\!]$. When $\mathbf{x} \in \mathscr{D}_{\sigma\to\tau}$ and $\mathbf{y} \in \mathscr{D}_\sigma$, we may write $\mathbf{x@y}$ as an equivalent to $@^{\sigma,t}(\mathbf{x})(\mathbf{y})$ or $[\![\mathbf{xy}]\!]$. In these terms, functionality and fullness can be explained as follows: an $\mathscr{L}$-structure is functional if $@^{\sigma,\tau}$ is one-one for each σ, τ, and it is full if $@^{\sigma,\tau}$ is onto for each σ, τ.[101]

The unfamiliar form of Definition 3.2 is largely motivated by the desire not to restrict our attention only to functional $\mathscr{L}$-structures. Every functional $\mathscr{L}$-structure is isomorphic to a ***frame***, i.e. an $\mathscr{L}$-structure in which for every complex type $\sigma \to \tau$, $\mathscr{D}_{\sigma\to\tau}$ is a subset of $\mathscr{D}_\tau^{\mathscr{D}_\tau}$, and $\mathbf{x} \in \mathscr{D}_{\sigma\to\tau}$ and $\mathbf{y} \in \mathscr{D}_\sigma$, $|\mathbf{xy}| = \mathbf{x}(\mathbf{y})$—or in other words, an $\mathscr{L}$-structure in which $@^{\sigma,\tau}$ is just the identity operation for every σ, τ.[102] Moreover, in any functional $\mathscr{L}$-structure, the full denotation function $[\![\cdot]\!]$ can be recovered from the application operation $@$ together with the restriction of $[\![\cdot]\!]$ to constants. For example, in any $\mathscr{L}$-structure, $[\![\lambda p.p]\!]_{t\to t}$ must be an $\mathbf{x} \in \mathscr{D}_{t\to t}$ such that $\mathbf{x@q} = \mathbf{q}$ for all $\mathbf{q} \in \mathscr{D}_t$ (since $[\![\lambda p.p]\!]_{t\to t}@\mathbf{q} = [\![(\lambda p.p)\mathbf{q}]\!]_t = [\![\mathbf{q}]\!]_t = \mathbf{q}$); but given functionality there can be at most one $\mathbf{x} \in \mathscr{D}_{t\to t}$ with this property, so

[101] Thanks here to Andrew Bacon.

[102] See BBK, Theorem 3.68.

the identity of $[\![\lambda p.p]\!]_{t\to t}$ is determined by the application operation. So if we were only interested in functional $\mathscr{L}$-structures, it would have been natural to define the structures of interest as frames meeting certain further conditions (i.e. those required for $[\![A]\!]$ to exist for every pure closed term A of $\mathscr{L}$), together with a typed function from Σ to the domain.

We call a model functional or full if its constituent $\mathscr{L}$-structure is functional or full. Here are some further significant properties of models:

Definition 4.4. A model is ***extensional*** if $\mathscr{D}_t$ contains exactly two elements.[103]

Definition 4.5. A model is ***extensionally full*** if, for every set $Z \subseteq \mathscr{D}_{\sigma_1} \times \cdots \times \mathscr{D}_{\sigma_n}$, there is some $\mathbf{x} \in \mathscr{D}_{\sigma_1\to\cdots\to\sigma_n\to t}$ such that, for any $\mathbf{y}_1 \in \mathscr{D}_{\sigma_1}$ … and $\mathbf{y}_n \in \mathscr{D}_{\sigma_n}$, $|[\![\mathbf{x}\mathbf{y}_1\ldots\mathbf{y}_n]\!]| = 1$ iff $\langle \mathbf{y}_1, \ldots, \mathbf{y}_n\rangle \in Z$.

Definition 4.6. A model is ***internally full*** if, for every function f from $\mathscr{D}_\sigma$ to $\mathscr{D}_\tau$, if there is some $\mathbf{x} \in \mathscr{D}_{\sigma\to\tau\to t}$ such that $|[\![\mathbf{x}\mathbf{y}\mathbf{y}']\!]| = 1$ exactly when $\mathbf{y}' = f(\mathbf{y})$, then there is some $\mathbf{z} \in \mathscr{D}_{\sigma\to\tau}$ such that for any $\mathbf{y} \in \mathscr{D}_\sigma$, $[\![\mathbf{z}\mathbf{y}]\!] = f(\mathbf{y})$.

Definition 4.7. A model is ***Leibnizean*** if, for every $\mathbf{z} \in \mathscr{D}_\sigma$, there is some $\mathbf{x} \in \mathscr{D}_{\sigma\to t}$ such that, for any $\mathbf{y} \in \mathscr{D}_\sigma$, $|\mathbf{x}@\mathbf{y}| = 1$ iff $\mathbf{y} = \mathbf{z}$.

A ***Henkin*** model is one that is populated, functional, extensional, and Leibnizean.[104] A ***standard*** model is a full Henkin model.

We can note the following logical relations among these properties.

Proposition 4.8. Every extensionally full model is Leibnizean.

Proof. Take $Z = \{\mathbf{z}\}$. □

Proposition 4.9. A model is full just in case it is both extensionally full and internally full.

Proof. If a model is extensionally full, then for every function f from $\mathscr{D}_\sigma$ to $\mathscr{D}_\tau$ there is some $\mathbf{x} \in \mathscr{D}_{\sigma\to\tau\to t}$ such that $|[\![\mathbf{x}\mathbf{y}\mathbf{y}']\!]| = 1$ iff $\mathbf{y}' = f(\mathbf{y})$, and if it is also internally full, there is a corresponding $\mathbf{z}$ in $\mathscr{D}_{\sigma\to\tau}$ such that for any $\mathbf{y} \in \mathscr{D}_\sigma$, $|[\![\mathbf{x}\mathbf{y}(\mathbf{z}\mathbf{y})]\!]| = 1$, hence $|\mathbf{z}\mathbf{y}| = f(\mathbf{y})$. In the other direction, the implication from fullness to internal fullness is immediate. To show that every full model is extensionally full, we use induction on n. For the base case where $n = 0$ and Z is either $\{\langle\rangle\}$ or $\varnothing$, we need only observe that $|[\![\exists p(p)]\!]| = 1$ and

[103] Note that $\mathscr{D}_t$ could not contain less than two elements, since it follows from the definition of a model that $|[\![\exists p(p)]\!]| = 1$ and $|[\![\forall p(p)]\!]| = 0$.

[104] Henkin (1950) actually did not require Leibnizeanness, leading to a mistake in central theorem, which Andrews (1972) points out, and remedies by imposing a condition equivalent to Leibnizeanness.

$|[\![\forall p(p)]\!]| = 0$. For the induction step, given $Z \subseteq \mathscr{D}_{\sigma_1} \times \cdots \times \mathscr{D}_{\sigma_{n+1}}$, we define a function f_Z from $\mathscr{D}_{\sigma_1}$ to $\mathscr{D}_{\sigma_2 \to \cdots \to \sigma_{n+1} \to t}$ by choosing $f_Z(y_1)$, for each $\mathbf{y}_1 \in \mathscr{D}_{\sigma_1}$, to be some $\mathbf{w}$ such that $|[\![\mathbf{w}\mathbf{y}_2 \ldots \mathbf{y}_{n+1}]\!]| = 1$ iff $\langle \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n+1} \rangle \in Z$: such a $\mathbf{w}$ is guaranteed to exist in every case by the induction hypothesis. Since the model is full, there is some $\mathbf{x} \in \mathscr{D}_{\sigma_1 \to \cdots \to \sigma_{n+1} \to t}$ such that $|\mathbf{x}\mathbf{y}_1| = f_Z(\mathbf{y}_1)$ for all $\mathbf{y}_1 \in \mathscr{D}_{\sigma_1}$. It follows that $|[\![\mathbf{x}\mathbf{y}_1 \ldots \mathbf{y}_{n+1}]\!]| = 1$ iff $\langle \mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n+1} \rangle \in Z$. □

Proposition 4.10. Any functional, extensional, extensionally full model is full.

Proof. First prove (by a straightforward induction on n) that in a functional, extensional model, when $\mathbf{x}, \mathbf{y} \in \mathscr{D}_{\sigma_1 \to \cdots \to \sigma_n \to t}$, $\mathbf{x} = \mathbf{y}$ iff $|\mathbf{x}| = |\mathbf{y}|$: i.e. for all $\langle \mathbf{z}_1, \ldots, \mathbf{z}_n \rangle \in \mathscr{D}_{\sigma_1} \times \cdots \times \mathscr{D}_{\sigma_n}$, $|[\![\mathbf{x}\mathbf{z}_1 \ldots \mathbf{z}_n]\!]| = |[\![\mathbf{y}\mathbf{z}_1 \ldots \mathbf{z}_n]\!]|$. Then suppose f is a function from $\mathscr{D}_\sigma$ to $\mathscr{D}_\tau$, where $\tau = \sigma_1 \to \cdots \to \sigma_n \to t$. By extensional fullness, there exists $\mathbf{x} \in \mathscr{D}_{\sigma \to \tau}$ such that for any $\mathbf{y}_0, \ldots, \mathbf{y}_n$, $|[\![\mathbf{x}\mathbf{y}_0 \ldots \mathbf{y}_n]\!]| = 1$ iff $|f(\mathbf{y}_0)@\mathbf{y}_1@ \cdots @\mathbf{y}_n| = 1$; but then by the foregoing fact, $[\![\mathbf{x}\mathbf{y}_0]\!] = f(\mathbf{y}_0)$ for all $\mathbf{y}_0 \in \mathscr{D}_\sigma$. □

Corollary 4.11. Any extensionally full Henkin model is standard.

The following definition helps to clarify the interest of Leibnizean models:

Definition 4.12. When $\mathfrak{M} = \langle \mathscr{D}, [\![\cdot]\!], |\cdot| \rangle$ is a model, ***Leibniz-equivalence*** in $\mathfrak{M}$ is the typed family $\dot{\sim}$ where $\dot{\sim}_\sigma$ is the equivalence relation on $\mathscr{D}_\sigma$ such that $\mathbf{x} \dot{\sim}_\sigma \mathbf{y}$ iff $|[\![\mathbf{z}\mathbf{x}]\!]| = |[\![\mathbf{z}\mathbf{y}]\!]|$ for every $\mathbf{z} \in \mathscr{D}_{\sigma \to t}$—or equivalently, iff $|[\![\mathbf{x} \equiv_\sigma \mathbf{y}]\!]| = 1$.

A model is Leibnizean, then, if no distinct elements of any $\mathscr{D}_\sigma$ are ever Leibniz equivalent. It is not hard to show that Leibniz equivalence is a congruence (see BBK, Theorem 3.62), and that the quotient of any model by Leibniz equivalence is Leibnizean.

One nice thing about the class of Leibnizean models is that within it, some of our other properties can be characterised by object language axioms.

Proposition 4.13. If a model is Leibnizean, it is is extensional iff *The Fregean Axiom* is true in it; functional iff *Functionality* is true in it for every type σ and terminal type τ; and internally full iff *Strong Plenitude* is true in it for every type σ and terminal type τ.

The Fregean Axiom $\quad \forall p^t \forall q^t ((p \leftrightarrow q) \to (p \equiv_t q))$

Functionality $\quad \forall x^{\sigma \to \tau} \forall y^{\sigma \to \tau} (\forall z^\sigma (xz \equiv_\tau yz) \to x \equiv_{\sigma \to \tau} y)$

Strong Plenitude

$$\forall x^{\sigma \to \tau \to t} (\forall y^\sigma \exists u^\tau (xyu \wedge \forall v^\tau (xyv \to u \equiv_\tau v)) \to \exists z^{\sigma \to \tau} \forall y^\sigma (xy(zy)))$$

(Note that *Plenitude* from §6 is equivalent to the restriction of *Strong Plenitude* to the case where τ is t.)

Proof. These follow straightforwardly from the fact that in a Leibnizean model, $|[\![\mathbf{x} \equiv_\sigma \mathbf{y}]\!]| = 1$ iff $\mathbf{x} = \mathbf{y}$. □

We may also note that any model (Leibnizean or not) is populated iff $\exists x^e(\exists f^{e\to t}(fx))$ is true in it.[105] However, not all of our classes of models can be characterised in this way. One example is the class of Leibnizean models. Since Leibniz-equivalence is a congruence, we can take the quotient of any model under Leibniz-equivalence, and the result is a model in which exactly the same closed sentences are true, and exactly the same sets of open formulae can be made true by providing compatible assignments. This means we would neither gain nor lose anything if we restricted our attention to Leibnizean models: the formulae and sequents valid in any class of models $\boldsymbol{C}$ are exactly those that are valid in the class of Leibniz-quotients of models in $\boldsymbol{C}$. This means that same proof procedures that are sound and complete for the class of all models are valid for the class of all Lebnizean models.

Extensional fullness is another property not characterised by any object-language axiom-schema, but unlike the class of Leibnizean models, no recursive proof procedure is sound and complete for the class of all extensionally full models. Gödel's first incompleteness theorem shows that no such proof procedure is sound and complete for the class of all *standard* models, or indeed for any class of standard models that contains at least one member whose domain in some type is infinite. And this result extends to the class of all extensionally full models, and to any class of extensionally full models that contains at least one member whose domain in some type is infinite.[106] The discovery that a certain theory was not true in any extensionally full

[105]Clearly the truth of $\exists x^e(\exists f^{e\to t}(fx))$ is sufficient for $\mathscr{D}_e$ to be nonempty. This suffices for the domain to be populated because of the following fact: whenever $\langle \mathscr{D}, [\![\cdot]\!] \rangle$ is an $\mathscr{L}$-structure and $\mathscr{L}$ is logical, $\mathscr{D}_\tau$ is nonempty for every terminal type τ. The proof of this is straightforward for $\mathscr{K}$-models: $\mathscr{D}_t$ must be nonempty since it contains $[\![\exists p(p)]\!]$, and so every complex type $\sigma \to \tau$ must be nonempty since $\mathscr{D}_{\sigma\to\tau}$ contains $[\![\lambda u^\sigma.\mathbf{x}]\!]$ for every $\mathbf{x} \in \mathscr{D}_\tau$. For an $\mathscr{I}$-model, it is still true but less obvious. The required lemmas are as follows:

(i) $\mathscr{D}_t$, $\mathscr{D}_{e\to t}$, $\mathscr{D}_{(\sigma\to t)\to t}$, and $\mathscr{D}_{t\to t\to t}$ are nonempty, since they contain $[\![\exists p(p)]\!]_t$, $[\![\lambda x.\exists f(fx)]\!]_{e\to t}$, $[\![\exists_\sigma]\!]$, and $[\![\wedge]\!]$.

(ii) If $\mathbf{x} \in \mathscr{D}_{\tau\to\tau\to\tau}$, $[\![\lambda y.\lambda z.\lambda w.\mathbf{x}(yw)(zw)]\!] \in \mathscr{D}_{(\rho\to\tau)\to(\rho\to\tau)\to\rho\to\tau}$.

(iii) If $\mathbf{x} \in \mathscr{D}_{(\sigma\to\tau)\to\tau}$, $[\![\lambda y.\lambda z.\mathbf{x}(\lambda w.ywz)]\!] \in \mathscr{D}_{(\sigma\to\rho\to\tau)\to\rho\to\tau}$.

(iv) If $\mathbf{x} \in \mathscr{D}_{\pi\to\tau}$ and $\mathbf{y} \in \mathscr{D}_{(\sigma\to\pi)\to\pi}$, $[\![\lambda z.\mathbf{x}(\mathbf{y}z)]\!] \in \mathscr{D}_{(\sigma\to\pi)\to\tau}$.

(v) If $\mathbf{x} \in \mathscr{D}_{\sigma\to\tau}$, $\mathbf{y} \in \mathscr{D}_{\rho\to\tau}$, and $\mathbf{z} \in \mathscr{D}_{\tau\to\tau\to\tau}$, $[\![\lambda v.\lambda w.\mathbf{z}(\mathbf{x}v)(\mathbf{y}w)]\!] \in \mathscr{D}_{\sigma\to\rho\to\tau}$.

Using these we can establish the nonemptiness of $\mathscr{D}_\tau$ for every terminal τ: first for types $\tau \to \tau \to \tau$ and $(\sigma \to \tau) \to \tau$ using (i)–(iii), and then for the remaining types using (iv) and (v).

[106]The following sketch shows that this claim is true in the special case where the infinite type is e and the models are populated; the general result is analogous. For each σ, define a "hereditary coex-

models would be a worrying development for proponents of that theory. Explaining exactly what the reasons for worrying would be is harder than explaining why it would be bad if the theory turned out not to have any models at all, and I won't try to get to the bottom of the matter. One danger is that when we inspected the proof, we could see how to transform it into a proof of the inconsistency of the theory either with the mathematical axioms we used in the proof, or with some analogues of these axioms formulated in higher-order logic (for example, a higher-order analogue of the axiom of choice). Insofar as the axioms were well supported, this would be a blow to the theory's credibility. Another possibility is that the proof would reveal some kind of clash between the way we are treating infinity within the theory and the way we are reasoning about infinity in the syntactic metatheory we use when specifying what the theory is—this would be the case, for example, if the theory contained the higher-order regimentations of 'Fred is a farmer', 'Fred's parents are farmers', 'Fred's parents' parents' are farmers', and so on, while also containing 'Not all Fred's ancestors are farmers', i.e. 'Some non-farmer has every property that Fred has and that is had by the parents of everything has it'. (Such a theory cannot have extensionally full models, since the extension of 'farmer' in any model must include all objects that can be connected to the denotation of 'Fred' by finite chains in which neighbouring objects belong to the extension of 'parent'.) There is clearly something deeply objectionable here, even though it is hard to pin down what it is.

---

tensiveness" predicate $\overset{E}{\sim}_\sigma$ of type $\sigma \to \sigma \to t$ as follows: (i) $x \overset{E}{\sim}_e y =_{\text{df}} x \equiv_e y$; (ii) $x \overset{E}{\sim}_t y =_{\text{df}} x \leftrightarrow y$; (iii) $x \overset{E}{\sim}_{\sigma\to\tau} y =_{\text{df}} \forall z \forall w (z \overset{E}{\sim}_\sigma w \to xz \overset{E}{\sim}_\tau yw)$. Also define $\forall^E_\sigma$ as $\lambda x^{\sigma\to t}.\forall y^\sigma((y \overset{E}{\sim}_\sigma y) \to xy)$ and $\exists^E_\sigma$ as $\lambda x^{\sigma\to t}.\exists y^\sigma((y \overset{E}{\sim}_\sigma y) \wedge xy)$, and note that $\forall^E_\sigma \overset{E}{\sim}_{(\sigma\to t)\to t} \forall^E_\sigma$ and $\exists^E_\sigma \overset{E}{\sim}_{(\sigma\to t)\to t} \exists^E_\sigma$ are valid for every σ. Say $\mathbf{x} \in \mathscr{D}_\sigma$ is *hereditarily extensional* if $|\,[\![\mathbf{x} \overset{E}{\sim}_\sigma \mathbf{x}]\!]\,| = 1$. Given any Log-model $\mathfrak{M} = \langle \mathscr{D}, [\![\cdot]\!], |\cdot| \rangle$, we can make a new model $\mathfrak{M}'$ by first throwing away all elements of the domains that are not hereditarily extensional, and then modifying the denotation function (see Definition 5.3) so that the new denotations of $\forall_\sigma$ and $\exists_\sigma$ are the old denotations of $\forall^E_\sigma$ and $\exists^E_\sigma$. (This makes sense since the old denotations of $\forall^E_\sigma$ and $\exists^E_\sigma$ are hereditarily extensional: see Proposition 5.9 below for a more careful description of this general procedure for modifying models.) Hereditary coextensiveness can be shown to be a congruence on $\mathfrak{M}'$, so we can take the quotient of $\mathfrak{M}'$ by it: call the resulting model $\mathfrak{M}^{\text{ext}}$ the ***extensional core*** of $\mathfrak{M}$. $\mathfrak{M}^{\text{ext}}$ will be extensional, since there are only two equivalence classes under hereditary coextensiveness in $\mathscr{D}_t$. Given the form of the definition of hereditary coextensiveness it is immediate that $\mathfrak{M}^{\text{ext}}$ will be functional; moreover, if $\mathfrak{M}$ was extensionally full, $\mathfrak{M}^{\text{ext}}$ will be extensionally full too. So by Proposition 4.10, if $\mathfrak{M}$ was extensionally full with an infinite domain of objects, $\mathfrak{M}^{\text{ext}}$ will be standard with an infinite domain of objects. Moreover, a sentence is true in $\mathfrak{M}^{\text{ext}}$ iff iff the result of simultaneously substituting $\forall^E_\sigma$ for $\forall_\sigma$ and $\exists^E_\sigma$ for $\exists_\sigma$ in it is true in $\mathfrak{M}$. Thus, any sound and complete proof procedure for a class C of models can be turned into a sound and complete proof procedure for the class C^{ext} of extensional cores of models in C. But if every model in C is populated and extensionally full and at least one is infinite in type e, then every model in C^{ext} is standard and at least one is infinite in type e, so Gödel's theorem rules out the existence of a sound and complete proof procedure for C^{ext}.

But however the objection is best conceived, we can assure ourselves that our theory will not face it if we manage to prove from standard mathematical axioms that the theory has extensionally full models. So, the strategic considerations that motivate the search for models in the first place also motivate the search for extensionally full models.

On the other hand, I see nothing worrying about the discovery that a theory lacks *full* models. As we have seen, any theory inconsistent with *Strong Plenitude* will lack Leibnizean, internally full models, and hence will lack full models (since the Leibniz-quotient of any full model is full); this observation has not suggested any objections to such theories that seem at all comparable in force to those that might follow on the discovery that a theory lacks extensionally full models.[107]

A5 Transformations of structures and models

This section will define a few operations for turning one structure or model into another which we will need to rely on later.

Most obviously, we can get a new structure (or model) just by shrinking the language on which the denotation function is defined.

Definition 5.1. When $\Sigma' \subseteq \Sigma$ and $\mathfrak{S} = \langle \mathscr{D}, [\![\cdot]\!] \rangle$ is a $\mathscr{I}^{\Sigma}$-structure [$\mathscr{K}^{\Sigma}$-structure], $\mathfrak{S}^{\Sigma'}$, the ***restriction of*** $\mathfrak{S}$ ***to*** Σ', is $\langle \mathscr{D}, [\cdot] \rangle$, where for any $g \in \mathscr{D}^{[V]}$, $[A]^g_\sigma = [\![A]\!]^g_\sigma$ when A is in $\mathscr{I}^{\Sigma'}(V)_\sigma$ [$\mathscr{K}^{\Sigma'}(V)_\sigma$] and undefined otherwise. $\mathfrak{M}^{\Sigma'} = \langle \mathscr{D}, [\cdot], |\cdot| \rangle$ is a model if $\mathfrak{M} = \langle \mathscr{D}, [\![\cdot]\!], |\cdot| \rangle$ is: we call this the restriction of $\mathfrak{M}$ to Σ'.

Similarly, we can turn a $\mathscr{K}^{\Sigma}$-structure or $\mathscr{K}^{\Sigma}$-model into an $\mathscr{I}^{\Sigma}$-structure or $\mathscr{I}^{\Sigma}$-model simply by excluding all terms of $\mathscr{K}^{\Sigma}$ that are not terms of $\mathscr{I}^{\Sigma}$ from the domain of each $[\![\cdot]\!]^g$.

Slightly less trivially, when Σ' is any signature, we can transform an $\mathscr{I}^{\Sigma}$-structure or $\mathscr{K}^{\Sigma}$-structure into an $\mathscr{I}^{\Sigma'}$-structure or $\mathscr{K}^{\Sigma'}$-structure by choosing some typed function I from Σ' to $\mathscr{D}$ to provide the interpretations of the constants in Σ'. To prepare the ground for this, we need a lemma to the effect that the denotation on an assignment of the result of performing a substitution operation on a term is the same as the denotation of the original term on an appropriately modified assignment. Recall that $[\pi]A$ is the result of simultaneously substituting $\pi(v)$ for each free occurrence of v in A (see Definition 2.5).

[107] Sometimes, schemas that imply *Strong Plenitude*, like $\forall x^\sigma \exists y^\rho (rxy) \rightarrow \exists f^{\sigma\rightarrow\rho}(\forall x^\sigma (rx(fx)))$, are taken as appropriate higher-order analogues of the axiom of choice. But in effect this simply bundles *Strong Plenitude* together with the weaker schema $\forall x^\sigma \exists y^\rho (rxy) \rightarrow \exists s^{\sigma\rightarrow\rho\rightarrow t}(\forall x^\sigma \exists y^\rho (rxy \wedge sxy \wedge \forall z^\rho (sxz \rightarrow z \equiv_\rho y)))$, and the intuitive and mathematical reasons for choice seem to attach more properly to the weaker schema. Thanks to Jeff Russell for discussion on this point.

Lemma 5.2 (Substitution Lemma)**.** If $A \in \mathscr{L}(V)_\sigma$, and π is a substitution function defined only on variables, then whenever $[\![[\pi]A]\!]^g_\sigma$ is defined, it equals $[\![A]\!]^{g^{\pi,V}}_\sigma$, where for each $v \in V_\rho$, $g^{\pi,V}_\rho(v)$ is $[\![[\pi]v]\!]^{g\restriction FV([\pi]v)}_\rho$—that is, $[\![\pi(v)]\!]^{g\restriction FV(\pi(v))}_\rho$ if π is defined on v and $g_\rho(v)$ otherwise.

Proof. We first show that that the claim is true for all "straightforward" substitution functions, where π is straightforward if none of the variables in its domain is free in any of the terms in its range. We argue by induction on the cardinality of V. *Base case:* if $V = \varnothing$, then $[\pi]A = A$, and $g = g^{\pi,V} = \varnothing$, so $[\![[\pi]A]\!]^g_\sigma = [\![A]\!]_\sigma = [\![A]\!]^{g^{\pi,V}}_\sigma$. *Induction step:* if A is of type e and some variable v is free in A, then A must be v. Then $g^{\pi,V}$ is $v \underset{e}{\mapsto} [\![[\pi]v]\!]^g_\sigma$, and so (by condition (i)) $[\![v]\!]^{g^{\pi,V}}_e = g^{\pi,V}_e(v) = [\![[\pi]v]\!]^g_e$. Otherwise, A is of some terminal type τ. Let $v \in V_\rho$ for some type ρ, and let $V^- = V - \{v\}$. Let π^- be the restriction of π to V^-, and let k and h be the restrictions of g respectively to variables free in $FV([\pi]v)$, and to variables free in $[\pi]u$ for some u in V^-. Note that h^{π^-,V^-} is the restriction of $g^{\pi,V}$ to V^-, so $g^{\pi,V} = h^{\pi^-,V^-} \cup (v \underset{\rho}{\mapsto} [\![[\pi]v]\!]^k)$. Then

$$
\begin{aligned}
[\![A]\!]^{g^{\pi,V}}_\tau &= [\![(\lambda v.A)v]\!]^{g^{\pi,V}}_\tau && \text{by condition (iii)}\\
&= [\![(\lambda v.A)v]\!]^{h^{\pi^-,V^-} \cup (v \underset{\rho}{\mapsto} [\![[\pi]v]\!]^k)}_\tau \\
&= [\![\lambda v.A]\!]^{h^{\pi^-,V^-}}_{\rho\to\tau} @ [\![v]\!]^{v \underset{\rho}{\mapsto} [\![[\pi]v]\!]^k}_\rho && \text{by condition (ii)}\\
&= [\![\lambda v.A]\!]^{h^{\pi^-,V^-}}_{\rho\to\tau} @ [\![[\pi]v]\!]^k_\rho && \text{by condition (i)}\\
&= [\![[\pi^-]\lambda v.A]\!]^h_{\rho\to\tau} @ [\![[\pi]v]\!]^k_\rho && \text{by the induction hypothesis}\\
&= [\![([\pi^-]\lambda v.A)([\pi]v)]\!]^g_\tau && \text{by condition (ii)}\\
&= [\![(\lambda v.[\pi^-]A)[\pi]v]\!]^g_\tau && \text{since } \pi \text{ is straightforward}\\
&= [\![[([\pi]v)/v][\pi^-]A]\!]^g_\tau && \text{by condition (iii)}\\
&= [\![[\pi]A]\!]^g_\tau && \text{since } \pi \text{ is straightforward}
\end{aligned}
$$

Finally, if the lemma holds for all straightforward π, it holds for all π, since for any π and term $A \in \mathscr{L}(V)_\sigma$ we can find two straightforward substitution functions π_1 and π_2 such that $[\pi]A = [\pi_1][\pi_2]A$: just let π_1 bijectively map V to a typed family of variables U none of which is in any term in the range of π, and let $\pi_2(u) = \pi(\pi_1^{-1}(u))$. So $[\![[\pi]A]\!]^g_\sigma = [\![[\pi_1][\pi_2]A]\!]^g_\sigma = [\![A]\!]^{(g^{\pi_1,V})^{\pi_2,U}}_\sigma$, where for any v in V, $(g^{\pi_1,V})^{\pi_2,U}(v) = [\![[\pi_2]v]\!]^{g^{\pi_1}\restriction FV(\pi_2(v))}$. Applying the just-proved result again, we see that this is the same as $[\![[\pi_1][\pi_2]v]\!]^g = [\![[\pi]v]\!]^g$. Thus $(g^{\pi_1,V})^{\pi_2,U} = g^{\pi,V}$, and so $[\![[\pi]A]\!]^g_\tau = [\![A]\!]^{g^{\pi,V}}_\tau$. □

With this lemma in the background we can define our "reinterpretation" operation as follows:

Definition 5.3. Where $\mathfrak{S} = \langle \mathscr{D}, [\![\cdot]\!] \rangle$ is an $\mathscr{L}^\Sigma$-structure (i.e. $\mathscr{T}^\Sigma$-structure or $\mathscr{K}^\Sigma$-structure), and I is a typed function from Σ' to $\mathscr{D}$, $\mathfrak{S}^I$, the ***reinterpretation of $\mathfrak{S}$ by I***,

is the $\mathscr{I}^{\Sigma'}$-structure [$\mathscr{K}^{\Sigma}$-structure] $\langle \mathscr{D}, [\![\cdot]\!]^I \rangle$, where for any $A \in \mathscr{L}^{\Sigma'}(V)$ containing constants $c_1 \ldots c_n$ of types $\sigma_1 \ldots \sigma_n$ and $g \in \mathscr{D}^{[V]}$, $[\![A]\!]^{Ig} = [\![[v_i/c_i]A]\!]^{g \cup [v_i \overset{}{\mapsto}_{\sigma_i} I(c_i)]}$, where $v_1 \ldots v_n$ are distinct variables of types $\sigma_1 \ldots \sigma_n$ not in V (by Lemma 5.2 it doesn't matter which ones we choose).

Proposition 5.4. $\mathfrak{S}^I$ as defined in Definition 5.3 is an $\mathscr{L}^{\Sigma'}$-structure. Moreover, if $\langle \mathscr{D}, [\cdot] \rangle$ is any $\mathscr{L}^{\Sigma'}$-structure such that $[\cdot]$ agrees with $[\![\cdot]\!]$ on all pure terms and $[c] = I(c)$ for each constant c of Σ', $[\cdot] = [\![\cdot]\!]^I$.

Proof. Condition (i) is trivially satisfied since a variable does not contain any constants; condition (iii) is satisfied since if $A \approx_{\beta\eta} B$, $[v_i/c_i]A \approx_{\beta\eta} [v_i/c_i]B$ by Proposition 2.11. For condition (ii), suppose that $[\![A]\!]^{Ig_1} = [\![C]\!]^{Ig_3}$ and $[\![B]\!]^{Ig_2} = [\![D]\!]^{Ig_4}$, where g_1 and g_2 are compatible and g_3 and g_4 are compatible. Let $c_1 \ldots c_n$ and $d_1 \ldots d_m$ be the constants of AB and CD respectively, and $v_1 \ldots v_n$ and $u_1 \ldots u_m$ variables of appropriate types. Then $[\![AB]\!]^{Ig_1 \cup g_2} = [\![[v_i/c_i]AB]\!]^{g_1 \cup g_2 \cup [v_i \overset{}{\mapsto}_{\sigma_i} I(c_i)]} = [\![[u_i/d_i]CD]\!]^{g_3 \cup g_4 \cup [u_i \overset{}{\mapsto}_{\sigma_i} I(d_i)]}$ (by condition (ii) for $[\![\cdot]\!]$) $= [\![CD]\!]^{Ig_3 \cup g_4}$.

To prove the uniqueness claim we argue by induction on the number of constants. The base case is the given fact that $[\![\cdot]\!]^I$ and $[\cdot]$ agree on pure terms. For the induction step, let c be some constant of type ρ in A, let $\mathbf{x} = [c] = [\![c]\!]^I$, and let B be a term not containing c such that $A = [c/v]B$. Then by (iii), $[A]^g = [(\lambda v.B)c]^g = [\lambda v.B]^g @ \mathbf{x} = [\![\lambda v.B]\!]^{Ig} @ \mathbf{x}$ by the induction hypothesis $= [\![(\lambda v.B)c]\!]^{Ig} = [\![A]\!]^{Ig}$. □

The above operations leave the domain $\mathscr{D}$ unchanged. We will also want to be able to take a structure or model and make a new one by "throwing away" some elements. In the case of an $\mathscr{L}$-structure, we can specify the conditions for this to be possible using the concept of a *closed* subcollection of the domain.

Definition 5.5. When $\mathfrak{S} = \langle \mathscr{D}, [\![\cdot]\!] \rangle$ is an $\mathscr{L}$-structure, $\mathscr{C} \subseteq \mathscr{D}$ is ***closed*** iff for any $A \in \mathscr{L}(V)_\sigma$, and any assignment function $g \in \mathscr{C}^{[V]}$, $[\![A]\!]^g_\sigma \in \mathscr{C}_\sigma$.

Definition 5.6. When $\mathfrak{S} = \langle \mathscr{D}, [\![\cdot]\!] \rangle$ is an $\mathscr{L}$-structure and $\mathscr{C} \subseteq \mathscr{D}$ is closed, the ***restriction of $\mathfrak{S}$ to*** $\mathscr{C}$, is the $\mathscr{L}$-structure $\mathfrak{S}^{\mathscr{C}} = \langle \mathscr{C}, [\cdot] \rangle$, where $[A]^g_\sigma$ is $[\![A]\!]^g_\sigma$ for any $A \in \mathscr{L}(V)_\sigma$ and $g \in \mathscr{C}^{[V]}$.

It is trivial to show that $\mathfrak{S}^{\mathscr{C}}$ satisfies the defining conditions to be an $\mathscr{L}$-structure, and that it is a substructure of $\mathfrak{S}$.

In the case of models, things are less straightforward. Typically, if we start with a model $\langle \mathscr{D}, [\![\cdot]\!], |\cdot| \rangle$ and just restrict $\mathscr{D}$ to some closed $\mathscr{C}$ while leaving $[\![\cdot]\!]$ and $|\cdot|$ unchanged (except for restricting them respectively to assignment functions for $\mathscr{C}$ and to $\mathscr{C}_t$), the result will no longer be a model, because $|[\![\exists_\sigma \mathbf{x}]\!]|$ and $|[\![\forall_\sigma \mathbf{x}]\!]|$ will no longer satisfy conditions (iv) and (v). We will need to either adjust $[\![\forall_\sigma]\!]$ and $[\![\exists_\sigma]\!]$ or adjust the valuation $|\cdot|$ to make sure that these conditions are still satisfied.

In practice it is easiest to do the former, since so long as the characteristic function of each $\mathscr{C}_\sigma$ is the extension of some element of $\mathscr{C}_{\sigma\to t}$, we can use these elements to specify the new denotations of the quantifiers as restrictions of the old denotations.

Definition 5.7. If $\mathfrak{M} = \langle \mathscr{D}, [\![\cdot]\!], |\cdot| \rangle$ is a model and $\mathbf{F}$ is a typed family such that, for each σ, $\mathbf{F}_\sigma \in \mathscr{D}_{\sigma\to t}$, the ***extension*** of $\mathbf{F}$ in $\mathfrak{M}$ is the typed collection $|\mathbf{F}|$ such that $|\mathbf{F}|_\sigma = \{\mathbf{y} \in \mathscr{D}_\sigma : |\mathbf{F}_\sigma @ \mathbf{y}| = 1\}$.

$\mathbf{F}$ is ***self-contained*** if $|\mathbf{F}_{\sigma\to t} @ \mathbf{F}_\sigma| = 1$ for every σ.

Definition 5.8. Suppose that $\mathfrak{M} = \langle \mathscr{D}, [\![\cdot]\!], |\cdot| \rangle$ is an $\mathscr{L}$-model, and $\mathbf{F}$ is a self-contained typed family with a closed extension. Then $\mathfrak{M}^{\mathbf{F}}$, the ***restriction of*** $\mathfrak{M}$ ***by*** $\mathbf{F}$, is the triple $\langle |\mathbf{F}|, [\![\cdot]\!]^{\mathbf{F}}, |\cdot|^{\mathbf{F}} \rangle$, where $|\cdot|^{\mathbf{F}}$ is just the restriction of $|\cdot|$ to $|\mathbf{F}|_t$, and $[\![\cdot]\!]^{\mathbf{F}}$ is the restriction to $|\mathbf{F}|$ of the reinterpreted denotation function $[\![\cdot]\!]^{I_{\mathbf{F}}}$ (see Definition 5.3) where $I_{\mathbf{F}}(c) = [\![c]\!]$ for every constant that is not a quantifier, $I_{\mathbf{F}}(\forall_\sigma) = [\![\lambda x^{\sigma\to t}.\forall_\sigma(\lambda y^\sigma.(\mathbf{F}_\sigma y \to xy))]\!]$, and $I_{\mathbf{F}}(\exists_\sigma) = [\![\lambda x^{\sigma\to t}.\exists_\sigma(\lambda y^\sigma.(\mathbf{F}_\sigma y \wedge xy))]\!]$.

Proposition 5.9. $\mathfrak{M}^{\mathbf{F}}$ as defined in Definition 5.8 is a model.

Proof. Since $\mathbf{F}$ is self-contained and $|\mathbf{F}|$ is closed in $\langle \mathscr{D}, [\![\cdot]\!] \rangle$, $I_{\mathbf{F}}(\forall_\sigma)$ and $I_{\mathbf{F}}(\exists_\sigma)$ belong to $|\mathbf{F}|$; it follows that $|\mathbf{F}|$ is closed in $\langle \mathscr{D}, [\![\cdot]\!]^{\mathbf{F}} \rangle$, so $\langle |\mathbf{F}|, [\![\cdot]\!]^{\mathbf{F}} \rangle$ is an $\mathscr{L}$-structure. So it suffices to show that $|\cdot|^{\mathbf{F}}$ obeys conditions (i)–(v). Conditions (i)–(iii) follow immediately from the fact that $[\![\wedge]\!]^{\mathbf{F}} = [\![\wedge]\!]$, $[\![\vee]\!]^{\mathbf{F}} = [\![\vee]\!]$, and $[\![\neg]\!]^{\mathbf{F}} = [\![\neg]\!]$ and that $|\cdot|^{\mathbf{F}}$ agrees with $|\cdot|$ whenever defined. For condition (iv), note that

$$\begin{aligned}
|[\![\exists_\sigma \mathbf{x}]\!]^{\mathbf{F}}|^{\mathbf{F}} &= |[\![\exists_\sigma \mathbf{x}]\!]^{\mathbf{F}}| = |[\![\lambda x^{\sigma\to t}.\exists_\sigma(\lambda y^\sigma.(\mathbf{F}_\sigma y \wedge xy))\mathbf{x}]\!]| \\
&= |[\![\exists_\sigma(\lambda y^\sigma.(\mathbf{F}_\sigma y \wedge \mathbf{x}y))]\!]| = \max\{|[\![(\lambda y^\sigma.(\mathbf{F}_\sigma y \wedge \mathbf{x}y))\mathbf{y}]\!]| : \mathbf{y} \in \mathscr{D}_\sigma\} \\
&= \max\{|[\![\mathbf{F}_\sigma \mathbf{y} \wedge \mathbf{x}\mathbf{y}]\!]| : \mathbf{y} \in \mathscr{D}_\sigma\} = \max\{\min\{|[\![\mathbf{F}_\sigma \mathbf{y}]\!]|, |[\![\mathbf{x}\mathbf{y}]\!]|\} : \mathbf{y} \in \mathscr{D}_\sigma\} \\
&= \max\{|[\![\mathbf{x}\mathbf{y}]\!]| : \mathbf{y} \in \mathscr{D}_\sigma \text{ and } |[\![\mathbf{F}_\sigma \mathbf{y}]\!]| = 1\} = \max\{|[\![\mathbf{x}\mathbf{y}]\!]^{\mathbf{F}}|^{\mathbf{F}} : \mathbf{y} \in |\mathbf{F}|_\sigma\}
\end{aligned}$$

Condition (v) holds for the same reason. □

A6 Applicative notions

When $\mathfrak{S} = \langle \mathscr{D}, [\![\cdot]\!] \rangle$ is an $\mathscr{L}$-structure and $\mathfrak{S}' = \langle \mathscr{D}', [\cdot] \rangle$ is an $\mathscr{L}'$-structure, a ***@-homomorphism*** from $\mathfrak{S}$ to $\mathfrak{S}'$ is a typed function f from $\mathscr{D}$ to $\mathscr{D}'$ such that $f(\mathbf{x})@f(\mathbf{y}) = f(\mathbf{x}@\mathbf{y})$ for every $\mathbf{x} \in \mathscr{D}_{\sigma\to\tau}$ and $\mathbf{y} \in \mathscr{D}_\sigma$. A ***@-isomorphism*** is a bijective @-homomorphism. In this section we will introduce some properties that depend only on the applicative structure induced by an $\mathscr{L}$-structure, and are thus preserved by @-isomorphisms. Here are the first two:

Definition 6.1. $\mathscr{C} \subseteq \mathscr{D}$ is ***@-closed*** iff $\mathbf{x}@\mathbf{y} \in \mathscr{C}_\tau$ whenever $\mathbf{x} \in \mathscr{C}_{\sigma\to\tau}$ and $\mathbf{y} \in \mathscr{C}_\sigma$.

Definition 6.2. $\mathscr{C} \subseteq \mathscr{D}$ is ***inclusive*** iff $\mathbf{x}@\mathbf{y} \in \mathscr{C}_\tau$ whenever $\mathbf{x} \in \mathscr{D}_{\sigma\to\tau}$ and $\mathbf{y} \in \mathscr{C}_\sigma$.

Here are some elementary consequences of these definitions:

Proposition 6.3. If $\mathscr{C}$ is inclusive, then $\mathbf{x}@\mathbf{y} \in \mathscr{C}_\tau$ whenever $\mathbf{x} \in \mathscr{C}_{\sigma\to\tau}$ and $\mathbf{y} \in \mathscr{D}_\sigma$.

Proof. $\mathbf{x}@\mathbf{y} = [\![(\lambda x.x\mathbf{y})\mathbf{x}]\!] = [\![(\lambda x.x\mathbf{y})]\!]\,@\mathbf{x}$. □

Proposition 6.4. If $\mathscr{C}$ is inclusive and $\mathscr{C}'$ is @-closed, $\mathscr{C} \cup \mathscr{C}'$ is @-closed.

Proof. If $\mathbf{x}$ and $\mathbf{y}$ are both in $\mathscr{C} \cup \mathscr{C}'$, then either they are both in $\mathscr{C}'$, in which case $\mathbf{x}@\mathbf{y}$ is in $\mathscr{C}'$ because $\mathscr{C}'$ is @-closed, or else $\mathbf{y}$ is in $\mathscr{C}$ in which case $\mathbf{x}@\mathbf{y}$ is also in $\mathscr{C}$ because $\mathscr{C}$ is inclusive, or else $\mathbf{x}$ is in $\mathscr{C}$ in which case $\mathbf{x}@\mathbf{y}$ is also in $\mathscr{C}$ by Proposition 6.3 □

Proposition 6.5. $\mathscr{C}$ is closed (see Definition 5.5) iff (a) $\mathscr{C}$ is @-closed, and (b) $\mathscr{C}_\sigma$ contains $[\![A]\!]$ for every closed term A of $\mathscr{L}$.

Proof. Suppose (a) and (b) hold, and the range of g is contained in $\mathscr{C}$; then for any term A with free variables $v_1, \ldots, v_n$,

$$[\![A]\!]^g = [\![(\lambda v_1 \ldots \lambda v_n.A)v_1 \ldots v_n]\!]^g = [\![\lambda v_1 \ldots \lambda v_n.A]\!]\,@g(v_1)@\cdots@g(v_n)$$

and so $[\![A]\!]^g$ is in $\mathscr{C}$. □

Note that an even an inclusive $\mathscr{C}$ need not be closed, since it need not contain $[\![A]\!]$ for every closed A—for example, $\varnothing$ is inclusive.

The concept of inclusiveness is of little interest when we are talking about $\mathscr{K}$-structures, because of the following:

Proposition 6.6. If $\langle \mathscr{D}, [\![\cdot]\!] \rangle$ is an $\mathscr{K}$-structure, $\mathscr{C} \subseteq \mathscr{D}$ is inclusive just in case either $\mathscr{C} = \varnothing$, or $\mathscr{C}_\tau = \mathscr{D}_\tau$ for every terminal τ.

Proof. The right to left direction is trivial. For the left to right direction, suppose that $\mathscr{C}$ is inclusive and $\mathbf{y} \in \mathscr{C}_\sigma$ for some σ; then for any $\mathbf{x} \in \mathscr{D}_\tau$, $\mathbf{x} = [\![\lambda u^\sigma.\mathbf{x}]\!]\,@\mathbf{y}$, so $\mathbf{x} \in \mathscr{C}_\tau$. □

By contrast, the domain of an $\mathscr{I}$-structure may have inclusive subcollections that are proper in terminal types. *Term structures* provide a rich source of examples of this. The general concept of a term structure makes sense for $\mathscr{K}$-structures as well (see BBK, sect. 3), but we will only be concerned here with term $\mathscr{I}$-structures whose domains consist of βη-equivalence classes of closed terms of some underlying $\mathscr{I}$-language.

Definition 6.7. When Σ and Σ' are signatures, a ***βη-term structure for $\mathscr{T}^\Sigma$ with base language*** $\mathscr{T}^{\Sigma'}$ is a $\mathscr{T}^\Sigma$-structure $\langle \mathscr{D}, [\![\cdot]\!] \rangle$ such that (a) $\mathscr{D}_\sigma$ is the set of βη-equivalence classes of closed terms in $\mathscr{T}^{\Sigma'}(\emptyset)_\sigma$, and (b) for any pure term $A \in \mathscr{T}^\emptyset(V)_\sigma$, $g \in \mathscr{D}^{[V]}$, and $h \in (\mathscr{T}^{\Sigma'}(\emptyset))^{[V]}$ such that $h_\rho(v) \in g_\rho(v)$ for all $v \in V_\rho$, $[\![A]\!]^g_\sigma$ is the set of all closed terms of $\mathscr{T}^{\Sigma'}(\emptyset)_\sigma$ βη-equivalent to $[h]A$, i.e. the result of substituting for each free variable of A the closed term to which it is mapped by h. (This makes sense since by Proposition 2.11, we will get the same βη-equivalence class whichever h we choose.)

Proposition 6.8. For any Σ', there is a unique βη-term structure for $\mathscr{T}^\emptyset$ with base language $\mathscr{T}^{\Sigma'}$

Proof. $[\![A]\!]^g$ is obviously uniquely pinned down for each term A of $\mathscr{T}^\emptyset$ by condition (b), so we just have to verify that the resulting $[\![\cdot]\!]$ obeys the conditions in the definition of an $\mathscr{L}$-structure.. Condition (i) is obviously satisfied since when $A \in C$, $[\![v]\!]^{v \overset{\sigma}{\mapsto} C}$ is the βη-equivalence class containing A, i.e. C. For condition (ii), suppose g_1 and g_2 are compatible, g_3 and g_4 are compatible, $[\![A]\!]^{g_1} = [\![C]\!]^{g_3}$, and $[\![B]\!]^{g_2} = [\![D]\!]^{g_4}$. Let h_1–h_4 be typed functions with the same domains as g_1–g_4 such that for each v, $h_i(v) \in g_i(v)$, h_1 and h_2 are compatible, and h_3 and h_4 are compatible. Then $[\![AC]\!]^{g_1 \cup g_2}$ is the $\beta\eta$-equivalence class containing $[h_1 \cup h_2]AC = [h_1]A[h_2]C$. But since $[h_1]A \approx_{\beta\eta} [h_3]C$ and $[h_2]B \approx_{\beta\eta} [h_4]D$, $[h_1]A[h_2]C \approx_{\beta\eta} [h_3]C[h_4]D = [h_3 \cup h_4]CD$ by Proposition 2.10; so $[\![AB]\!]^{g_1 \cup g_2} = [\![CD]\!]^{g_3 \cup g_4}$. Finally, condition (iii) holds as an immediate consequence of Proposition 2.11: if $A \approx_{\beta\eta} B$, then $[h]A \approx_{\beta\eta} [h]B$. □

It follows from Proposition 6.8 that so long as $\mathscr{T}^{\Sigma'}(\emptyset)_\sigma$ is nonempty whenever Σ_σ is, there exists at least one βη-term structure for $\mathscr{T}^\Sigma$ with base language $\mathscr{T}^{\Sigma'}$, since we can construct such a structure by starting with the βη-term structure for $\mathscr{T}^\emptyset$ with base language $\mathscr{T}^{\Sigma'}$, taking any typed function I from Σ to the domain of this structure, and using I to extend the denotation function to $\mathscr{T}^\Sigma$ in accordance with Proposition 5.4.

The domain of a βη-term structure with base language $\mathscr{T}^{\Sigma'}$ may have many nontrivial inclusive subcollections. Since we are dealing with a λI-language, for any $\Sigma'' \subseteq \Sigma'$, if some constant of Σ'' occurs in a term A of $\mathscr{T}^{\Sigma'}_\sigma$, it also occurs in any term βη-equivalent to A, and in any term βη-equivalent to BA for any B. Thus, the collection of all βη-equivalence classes containing terms containing constants in Σ'' is an inclusive subcollection of the domain.

The following definition picks out one important inclusive subcollection of the domain of any $\mathscr{L}$-structure in a way that depends only on its induced applicative structure.

Definition 6.9. When $\mathfrak{S} = \langle \mathscr{D}, [\![\cdot]\!] \rangle$ is an $\mathscr{L}$-structure, $\mathbf{z} \in \mathscr{D}_\sigma$ is ***directly circular*** if σ is of the form $\tau \to \tau$ and for some $\mathbf{x} \in \mathscr{D}_\tau$, $\mathbf{x} = \mathbf{z}@\mathbf{x}$. $\mathbf{z} \in \mathscr{D}_\sigma$ is ***circular*** if for some type τ, $\mathbf{x} \in \mathscr{D}_\tau$, and $\mathbf{y} \in \mathscr{D}_{\sigma\to\tau\to\tau}$, $\mathbf{x} = (\mathbf{y}@\mathbf{z})@\mathbf{x}$. Otherwise $\mathbf{z}$ is ***noncircular***. Noncirc($\mathfrak{S}$) is the typed collection of all non-circular elements of $\mathscr{D}$.

Proposition 6.10. Noncirc($\mathfrak{S}$) is inclusive.

Proof. If $\mathbf{z} \in \mathscr{D}_{\sigma\to\tau}$ and $\mathbf{z}' \in \mathscr{D}_\sigma$ are such that $\mathbf{z}@\mathbf{z}'$ is circular, there is some $\rho, \mathbf{x} \in \mathscr{D}_\rho, \mathbf{y} \in \mathscr{D}_{\tau\to\rho\to\rho}$ such that $\mathbf{x} = (\mathbf{y}@(\mathbf{z}@\mathbf{z}'))@\mathbf{x}$, in which case $\mathbf{x} = ([\![\lambda z^\sigma.\mathbf{y}(\mathbf{z}z)]\!]\,@\mathbf{z}')@\mathbf{x}$, so $\mathbf{z}'$ is circular too. □

Proposition 6.11. If $\mathscr{C} \subseteq \mathscr{D}$ is inclusive and contains no directly circular elements, then $\mathscr{C} \subseteq$ Noncirc($\mathfrak{S}$).

Proof. Suppose that $\mathbf{z} \in \mathscr{C}_\sigma$ is circular, i.e. $\mathbf{x} = (\mathbf{y}@\mathbf{z})@\mathbf{x}$ for some $\tau, \mathbf{x} \in \mathscr{D}_\tau, \mathbf{y} \in \mathscr{D}_{\sigma\to\tau\to\tau}$; then if $\mathscr{C}$ is inclusive, $\mathscr{C}_{\tau\to\tau}$ must contain $\mathbf{y}@\mathbf{z}$ which is directly circular. Thus an inclusive collection that contains no directly circular elements must consist entirely on noncircular elements. □

Since the union of any set of inclusive subcollections of $\mathscr{D}$ must itself be inclusive, it follows from the last two results that we can also characterise Noncirc($\mathfrak{S}$) as the largest inclusive subcollection of the domain of $\mathfrak{S}$ that contains no directly circular elements.

$\mathscr{K}$-structures cannot have noncircular elements, since every inclusive subcollection other than $\varnothing$ contains every member of $\mathscr{D}_\tau$ for every terminal τ, and hence in particular contains the directly circular $[\![\lambda p.p]\!]_{t\to t}$. However, when $\mathfrak{S}$ is an $\mathscr{I}$-structure, Noncirc($\mathfrak{S}$) can be nontrivial. We can see this by looking again at βη-term structures.

Proposition 6.12. If $\mathfrak{S}$ is a βη-term structure with base language $\mathscr{I}^{\Sigma'}$, every circular element is a combinator (i.e. identical to $[\![A]\!]$ for some closed term A of $\mathscr{I}^{\varnothing}$).

Proof. We have already seen that the collection of βη-equivalence classes of impure terms is inclusive. So by Proposition 6.11, to show that it is contained in Noncirc($\mathfrak{S}$) it suffices to show that it does not contain any directly circular elements. For any term A of $\mathscr{I}^{\Sigma'}$ and constant c in Σ', define the number of occurrences of c in A, Count(c, A), in the obvious way: Count$(c, A) = 0$ when A is a variable or constant other than c; Count$(c, c) = 1$; Count(c, AB) = Count(c, A) + Count(c, B); Count$(c, \lambda v.A)$ = Count(c, A). A straightforward induction shows that for any terms A, B of $\mathscr{I}^{\Sigma'}$, if v has at least one free occurrence in A, then Count$(c, [B/v]A) \geq$ Count(c, A) + Count(c, B). Since we are dealing with an $\mathscr{I}$-language, any term C that immediately β-reduces to some term D must be of the form $(\lambda v.A)B$ where v has a free occurrence in A, so for every $c \in \Sigma'$, we must have Count$(c, D) \geq$ Count(c, A) + Count(c, B) = Count(c, C). It follows from this that the

same inequality holds whenever C β-reduces in one step to D. Even more obviously, if C α-reduces or η-reduces in one step to D, $\mathrm{Count}(c, C) = \mathrm{Count}(c, D)$ for every c. Hence whenever C βη-reduces to D, $\mathrm{Count}(c, C) \geq \mathrm{Count}(c, D)$. By the Church-Rosser theorem (Proposition 2.13), when A is in βη-normal form and $B \approx_{\beta\eta} A$, B αβη-reduces to A, so $\mathrm{Count}(c, A) \geq \mathrm{Count}(c, B)$ for every c. And by the strong normalisation theorem (Proposition 2.12), every βη-equivalence class of closed terms of $\mathscr{I}^{\Sigma'}$ contains at least one member A in βη-normal form, and hence one for which $\mathrm{Count}(c, A)$ is maximal for every c.

Suppose then that $\mathbf{x}$ is a directly circular element of $\mathscr{D}_{\tau\to\tau}$, i.e. that for some $\mathbf{y} \in \mathscr{D}_\tau$, $\mathbf{y} = \mathbf{x}@\mathbf{y}$. Let $A \in \mathbf{x}$ and $C \in \mathbf{y}$ be in βη-normal form; then we know that $AC \approx_{\beta\eta} C$. But then, for every c, $\mathrm{Count}(c, A) + \mathrm{Count}(c, C) = \mathrm{Count}(c, AC) \leq \mathrm{Count}(c, C)$. This can only be true if, for every c, $\mathrm{Count}(c, A) = 0$: in other words, A is a pure closed term (a member of $\mathscr{I}^{\varnothing}\varnothing_{\tau\to\tau}$). But if so, $\mathbf{x}$ is $[\![A]\!]_{\tau\to\tau}$. □

Since the question whether an element is noncircular depends only on the applicative structure, an @-isomorphism from an $\mathscr{L}$-structure $\mathfrak{S}$ to an $\mathscr{L}'$-structure $\mathfrak{S}'$ will map $\mathrm{Noncirc}(\mathfrak{S})$ to $\mathrm{Noncirc}(\mathfrak{S}')$. More generally, the image of $\mathrm{Noncirc}(\mathfrak{S})$ under a *@-homomorphism* from $\mathfrak{S}$ to $\mathfrak{S}'$ must *contain* $\mathrm{Noncirc}(\mathfrak{S}')$:

Proposition 6.13. Suppose $\mathfrak{S} = \langle \mathscr{D}, [\![\cdot]\!] \rangle$ is an $\mathscr{L}$-structure, $\mathfrak{S}' = \langle \mathscr{D}', [\cdot] \rangle$ is an $\mathscr{L}'$-structure, and f is a @-homomorphism from $\mathfrak{S}$ to $\mathfrak{S}'$. Then for any $\mathbf{x} \in \mathscr{D}_\sigma$, if $f(\mathbf{x}) \in \mathrm{Noncirc}(\mathfrak{S}')_\sigma$, then $\mathbf{x} \in \mathrm{Noncirc}(\mathfrak{S})_\sigma$.

Proof. Let $\boldsymbol{J}$ be the typed collection defined by $\boldsymbol{J}_\sigma = \{\mathbf{x} \in \mathscr{D}_\sigma : f(\mathbf{x}) \in \mathrm{Noncirc}(\mathfrak{S}')_\sigma\}$. $\boldsymbol{J}$ contains no directly circular elements. For suppose that for $\mathbf{x} \in \mathscr{D}_{\tau\to\tau}$ and $\mathbf{y} \in \mathscr{D}_\tau$, $\mathbf{x}@\mathbf{y} = \mathbf{y}$; then $f(\mathbf{x})@f(\mathbf{y}) = f(x@y) = f(\mathbf{y})$, so $f(\mathbf{x})$ is a directly circular element of $\mathscr{D}'_{\tau\to\tau}$ and hence not in $\mathrm{Noncirc}(\mathfrak{S}')_{\tau\to\tau}$, so $\mathbf{x}$ is not in $\boldsymbol{J}_{\tau\to\tau}$. Also, $\boldsymbol{J}$ is inclusive. For suppose $\mathbf{y} \in \boldsymbol{J}_\sigma$ and $\mathbf{x} \in \mathscr{D}_{\sigma\to\tau}$; then $f(\mathbf{x}@\mathbf{y}) = f(\mathbf{x})@f(\mathbf{y})$ must be in $\mathrm{Noncirc}(\mathfrak{S}')$ since $\mathrm{Noncirc}(\mathfrak{S}')$ is inclusive and contains $f(\mathbf{y})$, and so $\mathbf{x}@\mathbf{y}$ is in $\boldsymbol{J}$. $\boldsymbol{J}$ is thus an inclusive collection with no directly circular elements, which means by Proposition 6.11 that it is contained in $\mathrm{Noncirc}(\mathfrak{S})$. □

Corollary 6.14. If $\mathfrak{S} = \langle \mathscr{D}, [\![\cdot]\!] \rangle$ is an $\mathscr{L}$-structure and $\mathscr{C} \subseteq \mathscr{D}$ is closed in $\mathfrak{S}$, then $\mathrm{Noncirc}(\mathfrak{S}) \cap \mathscr{C} \subseteq \mathrm{Noncirc}(()\mathfrak{S}^{\mathscr{C}})$ (where $\mathfrak{S}^{\mathscr{C}}$ is the restriction of $\mathfrak{S}$ to $\mathscr{C}$, see Definition 5.8).

Proof. The identity function on $\mathscr{C}$ is a @-homomorphism from $\mathfrak{S}^{\mathscr{C}}$ to $\mathfrak{S}$. □

A7 Model existence

We now have the ingredients we need to prove our consistency theorem for *OLC*. First, observe that the translation of *OLC* into a functional language is (βη-equivalent to) the following:

$\mathrm{OLC}_{\sigma,\tau}$ $\qquad \forall x^\tau \forall y^{\sigma\to\tau\to\tau} \forall z^\sigma ((x \equiv_\tau yzx) \to \mathrm{Logical}_\sigma\, z)$

(I have made explicit the initial universal quantifiers.) Let an ***admissible*** signature Σ be a logical signature such that for every σ, the constant 'Logical$_\sigma$' belongs to $\Sigma_{\sigma\to t}$. Let Neg be the signature whose only constant is $\neg$.

Then what we want to prove can be stated as follows:

Model existence theorem. *(a) For every admissible* Σ*, there is an extensionally full* $\mathscr{T}^\Sigma$*-model* $\mathfrak{M}$ *in which* $OLC_{\sigma,\tau}$ *is true for every* σ, τ*, and in which for any* $\mathbf{z} \in \mathscr{D}_\sigma$*,* $|[\![\text{Logical}_\sigma\ \mathbf{z}]\!]| = 1$ *only if* $\mathbf{z} = [\![A]\!]$ *for some closed term* A *of* $\mathscr{T}^\varnothing$*.*

(b) For every admissible Σ*, there is an extensionally full* $\mathscr{T}^\Sigma$*-model* $\mathfrak{M}$ *in which* $OLC_{\sigma,\tau}$ *is true for every* σ, τ*, and in which for any* $\mathbf{z} \in \mathscr{D}_\sigma$*,* $|[\![\text{Logical}_\sigma\ \mathbf{z}]\!]| = 1$ *only if* $\mathbf{z} = [\![A]\!]$ *for some closed term* A *of* $\mathscr{T}^{\text{Neg}}$*, and in which* $[\![\lambda p.\neg(\neg p)]\!]_{t\to t} = [\![\lambda p.p]\!]_{t\to t}$*.*

We can simplify our goal a little by observing that in any Leibnizean model, and *a fortiori* in any extensionally full model, $|[\![x \equiv_\tau yzx]\!]^g| = 1$ just in case $g(x) = (g(y)@g(z))@g(x)$, in which case $g(z)$ is circular. So, $\text{OLC}_{\sigma,\tau}$ is true in a model just in case for every $\mathbf{z} \in \mathscr{D}_\sigma$, either $\mathbf{z}$ is noncircular or $|[\![\text{Logical}_\sigma\ \mathbf{z}]\!]| = 1$. Moreover, we know (by Proposition 5.4) that we can extend the denotation function of any populated extensionally full model to interpret terms containing the constants Logical$_\sigma$ (and any other constants) and assign them any extensions we wish. So, we will be done if we can establish the following:

(a) There is a populated, extensionally full $\mathscr{T}^{\text{Log}}$-model $\mathfrak{M}$ in which each $\mathbf{z} \in \mathscr{D}_\sigma$ is either noncircular or identical to $[\![A]\!]$ for some closed term A of $\mathscr{T}^\varnothing$.

(b) There is a populated, extensionally full $\mathscr{T}^{\text{Log}}$-model $\mathfrak{M}$ in which each $\mathbf{z} \in \mathscr{D}_\sigma$ is either noncircular or identical to $[\![A]\!]$ for some closed term A of $\mathscr{T}^{\text{Neg}}$, and in which $[\![\lambda p.\neg(\neg p)]\!]_{t\to t} = [\![\lambda p.p]\!]_{t\to t}$.

This helps clarify why part (a) of the theorem can't be strengthened by including the final clause of part (b). If $[\![\lambda p.\neg(\neg p)]\!] = [\![\lambda p.p]\!]$, $[\![\neg]\!]$ is indirectly circular; but $[\![\neg]\!]$ cannot be a combinator in any model, since $[\![\lambda p.p]\!]$ is the only combinator of type $t \to t$, and the requirement that $|[\![\neg\mathbf{p}]\!]| = 1 - |[\![\mathbf{p}]\!]|$ rules out the possibility that $[\![\neg]\!] = [\![\lambda p.p]\!]$. Note however that there is nothing to stop it from being the case that $[\![\neg]\!] = [\![\lambda p.p]\!]$ in an $\mathscr{L}$-structure; indeed structures where this is the case will be crucial for proving (b).

So we can address both parts simultaneously, let O be either $\varnothing$ or Neg, and let an O-*combinator* in any $\mathscr{L}$-structure be any element of $\mathscr{D}$ denoted by some closed term of $\mathscr{T}^{\text{O}}$. Our strategy will be as follows. Step one is to identify a populated $\mathscr{T}^\Sigma$-structure $\mathfrak{S} = \langle \mathscr{N}, [\cdot] \rangle$ satisfying the condition that every circular element is an O-combinator. We have actually already carried out this step in Proposition 6.12,

where we saw that in a βη-term $\mathscr{J}$-structure, the only circular elements are combinators. Step two is to construct an extensionally full model $\mathfrak{M} = \langle \mathscr{D}, [\![\cdot]\!], |\cdot| \rangle$ for which there is a homomorphism f from $\langle \mathscr{D}, [\![\cdot]\!] \rangle$ to $\mathfrak{S}$. While it will not be true that every circular element of the domain of this model is an O-combinator (since the homomorphism f maps some circular elements that are not O-combinators onto circular O-combinators), this will still be true of all those elements that are denoted by closed terms. So, step three will be to throw away all those elements of $\mathscr{D}$ that are circular but not O-combinators, using the method for throwing things away described in Definition 5.8. The result of this final step will be a model meeting our requirements.[108]

The following construction gives us what we need to implement step two of our strategy.

Proposition 7.1. For any populated structure $\mathfrak{S} = \langle \mathscr{N}, [\cdot] \rangle$ for a logical $\mathscr{J}^{\Sigma}$, there is an extensionally full $\mathscr{J}^{\Sigma}$-model $\mathfrak{M} = \langle \mathscr{D}, [\![\cdot]\!], |\cdot| \rangle$ such that there is an surjective homomorphism from $\langle \mathscr{D}, [\![\cdot]\!] \rangle$ to $\mathfrak{S}$.

Proof. We will identify $\mathscr{D}_e$ with $\mathscr{N}_e$ and take the members of $\mathscr{D}_\tau$ (for terminal τ) to be ordered pairs. The second coordinate of each ordered pair (which we can think of as its "nonlogical content") will be a member of $\mathscr{N}_\tau$, so the homomorphism mapping $\mathscr{D}_\tau$ to $\mathscr{N}_\tau$ can just be the typed function that is the identity on $\mathscr{D}_e$ and maps each ordered pair in $\mathscr{D}_\tau$ to its second coordinate. The first coordinate of each ordered pair will just be an *extension*. In particular, the first coordinates of the ordered pairs in $\mathscr{D}_t$ will each be 0 or 1, so $|\cdot|$ can just be the function mapping each member of $\mathscr{D}_t$ to its first coordinate. We allow the two coordinates to recombine freely.

To make this precise, let the first and second coordinates of an ordered pair p be denoted by $\pi_1(p)$ and $\pi_2(p)$ respectively; then we construct $\mathscr{D}$, $[\![\cdot]\!]$, and $|\cdot|$ respectively as follows:

(1) a. $\mathscr{D}_e = \mathscr{N}_e$.

b. $\mathscr{D}_t = \{0, 1\} \times \mathscr{N}_t$.

c. $\mathscr{D}_{\sigma\to\tau} = \{\pi_1(\mathbf{x}) : \mathbf{x} \in \mathscr{D}_\tau\}^{\mathscr{D}_\sigma} \times \mathscr{N}_{\sigma\to\tau}$.

(2) a. $[\![v]\!]_\sigma^{v\overset{\mathbf{x}}{\mapsto}} = \mathbf{x}$ for all $\mathbf{x} \in \mathscr{D}_\sigma$.

b. $[\![AB]\!]_\tau^{g\cup h} = \langle \pi_1([\![A]\!]_{\sigma\to\tau}^{g})([\![B]\!]_\sigma^{h}), [AB]_\tau^{\pi_2\circ(g\cup h)} \rangle$, whenever $A \in \mathscr{J}^{\Sigma}(V)_{\sigma\to\tau}$, $B \in \mathscr{J}^{\Sigma}(V')_\sigma$, $g \in \mathscr{D}^{[V]}$, $h \in \mathscr{D}^{[V']}$, and g and h are compatible.

c. $[\![\lambda v.A]\!]_{\sigma\to\tau}^{g} = \langle f, [\lambda v.A]_{\sigma\to\tau}^{\pi_2\circ g} \rangle$, whenever $A \in \mathscr{J}^{\Sigma}(V)_\tau$, $g \in \mathscr{D}^{[V-\{v\}]}$, $v \in V_\sigma$, and f is the function such that for all $\mathbf{x} \in \mathscr{D}_\sigma$, $f(\mathbf{x}) = \pi_1([\![A]\!]_\tau^{g\cup(v\overset{\mathbf{x}}{\mapsto})})$.

[108] Why not proceed more straightforwardly, by simply defining a valuation $|\cdot|$ that makes $\langle \mathscr{N}, [\cdot], |\cdot| \rangle$ an extensionally full model? The answer is that there cannot be an extensionally full model $\langle \mathscr{N}, [\cdot], |\cdot| \rangle$ in which $\langle \mathscr{N}, [\cdot] \rangle$ is a βη-term structure. The proof of this result is omitted here.

d. $[\![\neg]\!] = \langle f_\neg, [\neg]\rangle$, $[\![\wedge]\!] = \langle f_\wedge, [\wedge]\rangle$, $[\![\vee]\!] = \langle f_\wedge, [\vee]\rangle$, $[\![\forall_\sigma]\!] = \langle f_{\forall_\sigma}, [\forall_\sigma]\rangle$, and $[\![\exists_\sigma]\!] = \langle f_{\exists_\sigma}, [\exists_\sigma]\rangle$, where for all $\mathbf{p}, \mathbf{q} \in \mathscr{D}_t$ and $\mathbf{x} \in \mathscr{D}_{\sigma\to t}$, $f_\neg(\mathbf{p}) = 1 - \pi_1(\mathbf{p})$, $f_\wedge(\mathbf{p})(\mathbf{q}) = \min\{\pi_1(\mathbf{p}), \pi_1(\mathbf{q})\}$, $f_\vee(\mathbf{p})(\mathbf{q}) = \max\{\pi_1(\mathbf{p}), \pi_1(\mathbf{q})\}$, $f_{\forall_\sigma}(\mathbf{x}) = \min\{\pi_1(\mathbf{x}@\mathbf{y}) : \mathbf{y} \in \mathscr{D}_\sigma\}$, and $f_{\exists_\sigma}(\mathbf{x}) = \max\{\pi_1(\mathbf{x}@\mathbf{y}) : \mathbf{y} \in \mathscr{D}_\sigma\}$.

(3) $|\mathbf{p}| = \pi_1(\mathbf{p})$ for all $\mathbf{p} \in \mathscr{D}_t$.

Note that since clauses (2b) and (2c) just pass the second coordinate of the argument of $[\![\cdot]\!]$ over to $[\cdot]$, a trivial induction suffices to show that $\pi_2([\![A]\!]^g_\sigma) = [A]^{\pi_2\circ g}_\sigma$ whenever $[\![A]\!]^g_\sigma$ is defined.

It follows immediately from clauses (2d) and (3) that $|\cdot|$ obeys all the conditions for $\langle \mathscr{D}, [\![\cdot]\!], |\cdot|\rangle$ to be a model provided that $\langle \mathscr{D}, [\![\cdot]\!]\rangle$ is an $\mathscr{S}^\Sigma$-structure. So, to show that it is a model it suffices to verify that $\langle \mathscr{D}, [\![\cdot]\!]\rangle$ satisfies conditions (i)–(iii) in the definition of an $\mathscr{S}^\Sigma$-structure. Condition (i) is just clause (2a). For condition (ii), suppose that $[\![A]\!]^{g_1}_{\sigma\to\tau} = [\![C]\!]^{g_2}_{\sigma\to\tau}$, $[\![B]\!]^{g_3}_\sigma = [\![D]\!]^{g_4}_\sigma$, g_1 and g_2 are compatible, and g_3 and g_4 are compatible. Then $[A]^{\pi_2\circ g_1}_{\sigma\to\tau} = [C]^{\pi_2\circ g_2}_{\sigma\to\tau}$, $[B]^{\pi_2\circ g_3}_\sigma = [D]^{\pi_2\circ h_4}_\sigma$, $\pi_2\circ g_1$ and $\pi_2\circ g_2$ are compatible, and $\pi_2\circ g_3$ and $\pi_2\circ g_4$ are compatible, so since $\mathfrak{S}$ satisfies condition (ii), $[AC]^{\pi_2\circ(g_1\cup g_2)}_\tau = [AC]^{(\pi_2\circ g_1)\cup(\pi_2\circ g_2)}_\tau = [BD]^{(\pi_2\circ g_3)\cup(\pi_2\circ g_4)}_\tau = [BD]^{\pi_2\circ(g_3\cup g_4)}_\tau$. Thus by (2b), $[\![AC]\!]^{g_1\cup g_2}_\tau = \langle \pi_1([\![A]\!]^{g_1}_{\sigma\to\tau})([\![C]\!]^{g_2}_\sigma), [AC]^{\pi_2\circ(g_1\circ g_2)}_\tau\rangle = \langle \pi_1([\![B]\!]^{g_3}_{\sigma\to\tau})([\![D]\!]^{g_4}_\sigma), [BD]^{\pi_2\circ(g_3\cup g_4)}_\tau\rangle = [\![BD]\!]^{g_3\cup g_4}_\tau$.

For condition (iii), it is enough to show that if A βη-reduces in one step to B, $\pi_1([\![A]\!]^g_\sigma) = \pi_1([\![B]\!]^g_\sigma)$ whenever defined. (The second coordinates are the same since $\mathfrak{S}$ is an $\mathscr{S}^\Sigma$-structure). We show this by an induction using the definition of one-step βη-reduction. *Base case:* A immediately βη-reduces to B. If A immediately η-reduces to B, σ must be of the form $\rho \to \tau$, and A is $\lambda v.(Bv)$ for some $v \in \mathrm{Var}_\rho$ not free in B. So by (2c) and (2b), $\pi_1([\![A]\!]^g_\sigma)$ is the function f such that for any $\mathbf{x} \in \mathscr{D}_\rho$, $f(\mathbf{x}) = \pi_1([\![Bv]\!]^{g\cup v\overset{}{\mapsto}\mathbf{x}}_\tau) = \pi_1([\![B]\!]^g_{\rho\to\tau})([\![v]\!]^{v\overset{}{\mapsto}\mathbf{x}}_\rho) = \pi_1([\![B]\!]^g_{\rho\to\tau})(\mathbf{x})$, i.e. $f = \pi_1([\![B]\!]^g_{\rho\to\tau})$. If A immediately β-reduces to B, then for some types ρ, τ, $v \in \mathrm{Var}_\rho$, $C \in \mathscr{S}^\Sigma(V \cup \langle v, \rho\rangle)_\tau$ and $D \in \mathscr{S}^\Sigma(U)_\rho$, A is $(\lambda v.C)D$ and B is $[D/v]C$. Then by (2c), $\pi_1([\![\lambda v.C]\!]^{g\downarrow V}_{\rho\to\tau})$ is the function f such that for any $\mathbf{x} \in \mathscr{D}_\rho$, $f(\mathbf{x}) = \pi_1([\![C]\!]^{g\downarrow V\cup v\overset{}{\mapsto}\mathbf{x}}_\tau)$. Thus by (2b), $\pi_1([\![A]\!]^g_\tau) = \pi_1([\![\lambda v.C]\!]^{g\downarrow V}_{\rho\to\tau})([\![D]\!]^{g\downarrow U}_\rho) = f([\![D]\!]^{g\downarrow U}_\rho) = \pi_1([\![C]\!]^{g\downarrow V\cup v\overset{}{\mapsto}[\![D]\!]^{g\downarrow U}_\rho}_\tau)$. By the substitution lemma (Lemma 5.2), this is the same as $[\![[D/v]C]\!]^g_\tau = [\![B]\!]^g_\tau$. *Induction step:* suppose that $A, B \in \mathscr{S}^\Sigma(V)_\sigma$ are such that $\pi_1([\![A]\!]^g_\sigma) = \pi_1([\![B]\!]^g_\sigma)$ for every $g \in \mathscr{D}^{[V]}$. Then by the above proof that $[\![\cdot]\!]$ satisfies condition (ii), whenever $C \in \mathscr{S}^\Sigma(V')_{\sigma\to\tau}$ and $h \in \mathscr{D}^{[V\cup V']}$, $\pi_1([\![CA]\!]^h_\tau) = \pi_1([\![CB]\!]^h_\tau)$. Similarly if σ is of the form $\rho \to \tau$ and $C \in \mathscr{S}^\Sigma(V')_\rho$, $\pi_1([\![AC]\!]^h_\tau) = \pi_1([\![BC]\!]^h_\tau)$. And finally whenever $v \in V_\rho$ and $\mathbf{x} \in \mathscr{D}_\rho$, $\pi_1([\![\lambda v.A]\!]^{g-\{v\}}_{\rho\to\sigma})(\mathbf{x}) = [\![A]\!]^{g[v\overset{}{\mapsto}\mathbf{x}]}_\sigma = [\![B]\!]^{g[v\overset{}{\mapsto}\mathbf{x}]}_\sigma = \pi_1([\![\lambda v.B]\!]^{g-\{v\}}_{\rho\to\sigma})(\mathbf{x})$.

So, $\langle \mathscr{D}, [\![\cdot]\!], |\cdot|\rangle$ is a model. To show that it is extensionally full, note that the first coordinate of any member of $\mathscr{D}_{\sigma_1\to\cdots\to\sigma_n\to t}$ belongs to

$$\left(\left(\{0,1\}^{\sigma_n}\right)^{\cdot^{\cdot}}\right)^{\sigma_1}$$

and for every such function, there is a member of $\mathscr{D}_{\sigma_1\to\cdots\to\sigma_n\to t}$ with it as first coordinate and any arbitrary member of $\mathscr{N}_{\sigma_1\to\cdots\to\sigma_n\to t}$ as second coordinate. So, for any $Z \subseteq \mathscr{D}_{\sigma_1}\times\cdots\times\mathscr{D}_{\sigma_n}$, there is an $\mathbf{x} \in \mathscr{D}_{\sigma_1\to\cdots\to\sigma_n\to t}$ such that for any $\langle \mathbf{y}_1, \ldots, \mathbf{y}_n\rangle \in \mathscr{D}_{\sigma_1}\times\cdots\times\mathscr{D}_{\sigma_n}$, $|\,[\![\mathbf{x}\mathbf{y}_1\ldots\mathbf{y}_n]\!]\,| = \pi_1(\mathbf{x})(\mathbf{y}_1)\cdots(\mathbf{y}_n) = 1$ iff $\langle \mathbf{y}_1, \ldots, \mathbf{y}_n\rangle \in Z$.

Finally, we have already noted that the typed function that is the identity on $\mathscr{D}_e$ and maps each member of $\mathscr{D}_\tau$ to its second coordinate satisfies the condition to be a homomorphism from $\mathscr{D}$ to $\mathscr{N}$. This homomorphism is surjective because every element of $\mathscr{N}_\tau$ is the second element of at least one member of $\mathscr{D}_\tau$. □

We can use Proposition 7.1 to construct an extensionally full model in which every circular element denoted by a closed term is is an O-combinator.

Proposition 7.2. When O is ∅ or Neg, there is an extensionally full $\mathscr{F}^{\mathrm{Log}}$-model $\mathfrak{M} = \langle \mathscr{D}, [\![\cdot]\!], |\cdot|\rangle$ such that for any closed A, either $[\![A]\!]$ is in Noncirc($\langle \mathscr{D}, [\![\cdot]\!]\rangle$) or $[\![A]\!] = [\![B]\!]$ for some $B \in \mathscr{F}^{\mathrm{O}}$. Moreover, if O = Neg, $[\![\lambda p.\neg\neg p]\!] = [\![\lambda p.p]\!]$.

Proof. Choose any Σ such that $\mathscr{F}^{\varnothing}(\varnothing)_\sigma$ is a proper subset of $\mathscr{F}^{\varnothing}(\varnothing)_\sigma$ for every type σ. (For example, this will be the case if Σ is populated.) Let $\mathfrak{S}^- = \langle \mathscr{N}, [\cdot]^-\rangle$ be the unique βη-term $\mathscr{F}^{\varnothing}$-structure with base language $\mathscr{F}^{\Sigma}$ (see Proposition 6.8). Choose a typed function I from Log to $\mathscr{N}$ such that if $c \in \mathrm{Log}_\sigma$ and $c \notin \mathrm{O}_\sigma$, $I(c)$ is an impure element of $\mathscr{N}_\sigma$ (a βη-equivalence class each of whose members contain at least one constant in Σ), while if $c \in \mathrm{O}_\sigma$ (i.e. $c = \neg$ and O = Neg and $\sigma = t \to t$), $I(c)$ is the βη-equivalence class containing $\lambda p.p$. Let $\mathfrak{S} = \langle \mathscr{N}, [\cdot]\rangle$ be the $\mathscr{F}^{\mathrm{Log}}$-structure that results from reinterpreting $\mathfrak{S}^-$ with I (see Definition 5.3). Finally let $\mathfrak{M} = \langle \mathscr{D}, [\![\cdot]\!], |\cdot|\rangle$ be an extensionally full $\mathscr{F}^{\mathrm{Log}}$-model, and f a surjective homomorphism from $\langle \mathscr{D}, [\![\cdot]\!]\rangle$ to $\mathfrak{S}$: such an $\mathfrak{M}$ and f exist by Proposition 7.1. Since f is a homomorphism, it is a @-homomorphism, so by Proposition 6.13, $\mathbf{x}$ is in Noncirc($\langle \mathscr{D}, [\![\cdot]\!]\rangle$) whenever $f(\mathbf{x})$ is in Noncirc($\mathfrak{S}$). But for any closed term A of $\mathscr{F}^{\mathrm{Log}}$ that contains at least one constant not in O, $f([\![A]\!]) = [A]$ is a βη-equivalence class of impure terms, and hence in Noncirc($\mathfrak{S}$) by Proposition 6.12; so $[\![A]\!] \in$ Noncirc($\mathscr{D}, [\![\cdot]\!]$).

Moreover, if O is Neg and we built $\mathfrak{M}$ using the ordered-pair construction of Proposition 7.1, $[\![\lambda p.\neg\neg p]\!]$ will be the same as $[\![\lambda p.p]\!]$, since their second coordinates $[\lambda p.\neg\neg p]$ and $[\lambda p.p]$ are the same, and their first coordinates are also the same, namely the function mapping each ordered pair in $\{0,1\}\times\mathscr{N}_t$ to its first coordinate. □

Finally comes step three, where we throw away everything in the domain of $\mathfrak{M}$ that is circular but not an O-combinator. This is possible because the typed collection of elements that are either noncircular or O-combinators is closed, and since $\mathfrak{M}$ is extensionally full, its domain contains a family with this typed collection as its extension.

Proposition 7.3. There exists an extensionally full $\mathscr{F}^{\mathrm{Log}}$-model $\mathfrak{M} = \langle \mathscr{D}, [\![\cdot]\!], |\cdot|\rangle$ such that for every $\mathbf{x} \in \mathscr{D}_\sigma$, either $\mathbf{x} = [\![A]\!]$ for some closed term A of $\mathscr{F}^{\mathrm{O}}$, or $\mathbf{x}$ is in

Noncirc($\langle \mathscr{D}, [\![\cdot]\!] \rangle$). Moreover, if O is Neg, there is a model meeting these conditions in which $[\![\lambda p.\neg\neg p]\!] = [\![\lambda p.p]\!]$.

Proof. Let $\mathfrak{M} = \langle \mathscr{D}, [\![\cdot]\!], |\cdot| \rangle$ be an extensionally full $\mathscr{I}^{\mathrm{Log}}$-model such that for any closed A, $[\![A]\!]$ is either in Noncirc($\langle \mathscr{D}, [\![\cdot]\!] \rangle$) or an O-combinator, and such that if O = Neg, $[\![\lambda p.\neg\neg p]\!] = [\![\lambda p.p]\!]$; we know from Proposition 7.2 that such a model can always be found. Let $\mathscr{C}$ be the subcollection of $\mathscr{D}$ such that $\mathbf{x} \in \mathscr{C}_\sigma$ iff $\mathbf{x}$ is either noncircular or a O-combinator. Since $\mathscr{C}$ is the union of an inclusive collection and an @-closed one, it is @-closed by Proposition 6.4. Moreover, $\mathscr{C}$ contains $[\![A]\!]$ for every closed term A, so by Proposition 6.5, $\mathscr{C}$ is closed. Since $\mathfrak{M}$ is extensionally full, we know that for each type σ, there exists some $\mathbf{x} \in \mathscr{D}_{\sigma\to\tau}$ such that for any $\mathbf{y} \in \mathscr{D}_\sigma$, $|\mathbf{x}@\mathbf{y}| = 1$ iff $y \in \mathscr{C}_\sigma$. For any such $\mathbf{x} \in \mathscr{D}_{\sigma\to\tau}$, there is a noncircular $\mathbf{x}'$ that has the same extension as $\mathbf{x}$: for any $\mathbf{p} \in \mathrm{Noncirc}(\mathfrak{M})_t$, we can take $\mathbf{x}' = [\![\lambda y^\sigma.\mathbf{x}y \wedge (\mathbf{p} \vee \neg\mathbf{p})]\!]$. (Noncirc$(\mathfrak{M})_t$ cannot be empty: since $\exists p(p)$ is not a term of $\mathscr{I}^{\mathrm{O}}$, $[\![\exists p(p)]\!]$ is noncircular.) So we can choose a typed family $\mathbf{F}$ such that for every σ, $\mathbf{F}_\sigma$ is a noncircular element of $\mathscr{D}_{\sigma\to t}$ whose extension is $\mathscr{C}_\sigma$. $\mathbf{F}$ thus has a closed extension; it is also self-contained, since being noncircular, $\mathbf{F}_\sigma$ belongs to $\mathscr{C}_{\sigma\to t}$. So by Proposition 5.9, there is a model $\mathfrak{M}^{\mathbf{F}} = \langle \mathscr{C}, [\![\cdot]\!]^{\mathbf{F}}, |\cdot|^{\mathbf{F}} \rangle$, the restriction of $\mathfrak{M}$ by $\mathbf{F}$.

The restricted structure $\langle \mathscr{C}, [\![\cdot]\!]^{\mathbf{F}} \rangle$ is a @-substructure of $\langle \mathscr{D}, [\![\cdot]\!] \rangle$. So by Corollary 6.14, every noncircular element of $\langle \mathscr{D}, [\![\cdot]\!] \rangle$ that is in $\mathscr{C}$ is a noncircular element of $\langle \mathscr{C}, [\![\cdot]\!]^{\mathbf{F}} \rangle$. And every *circular* element of $\mathscr{D}$ that is in $\mathscr{C}$ is $[\![A]\!]$ for some closed term A of $\mathscr{I}^{\mathrm{O}}$, in which case it is also $[\![A]\!]^{\mathbf{F}}$, since $[\![\cdot]\!]$ and $[\![\cdot]\!]^{\mathbf{F}}$ agree on $\mathscr{I}^{\mathrm{O}}$. So, $\mathfrak{M}^{\mathbf{F}}$ is a model in which every circular element is an O-combinator.

$\mathfrak{M}^{\mathbf{F}}$ is also extensionally full. For suppose $Z \subseteq \mathscr{C}_{\sigma_1} \times \cdots \times \mathscr{C}_{\sigma_n}$. Then $Z \subseteq \mathscr{D}_{\sigma_1} \times \cdots \times \mathscr{D}_{\sigma_n}$, so by the extensional fullness of $\mathfrak{M}$, there is some $\mathbf{x} \in \mathscr{D}_{\sigma_1\to\cdots\to\sigma_n\to t}$ such that $|[\![\mathbf{x}\mathbf{y}_1\ldots\mathbf{y}_n]\!]| = 1$ exactly when $\langle \mathbf{y}_1, \ldots, \mathbf{y}_n \rangle \in Z$. By the fact noted above, this $\mathbf{x}$ has the same extension as some noncircular $\mathbf{x}'$; being noncircular, $\mathbf{x}'$ also belongs to $\mathscr{C}_{\sigma_1\to\cdots\to\sigma_n\to t}$. So for any $\langle \mathbf{y}_1, \ldots, \mathbf{y}_n \rangle \in \mathscr{C}_{\sigma_1} \times \cdots \times \mathscr{C}_{\sigma_n}$, $|[\mathbf{x}'\mathbf{y}_1\ldots\mathbf{y}_n]|^{\mathbf{F}} = |[\![\mathbf{x}'\mathbf{y}_1\ldots\mathbf{y}_n]\!]| = |[\![\mathbf{x}\mathbf{y}_1\ldots\mathbf{y}_n]\!]| = 1$ iff $\langle \mathbf{y}_1, \ldots \mathbf{y}_n \rangle \in Z$.

Finally, if O is Neg, $[\![\lambda p.\neg\neg p]\!]^{\mathbf{F}} = [\![\lambda p.\neg\neg p]\!] = [\![\lambda p.p]\!] = [\![\lambda p.p]\!]^{\mathbf{F}}$, since $[\![\cdot]\!]^{\mathbf{F}}$ coincides with $[\![\cdot]\!]$ on quantifier-free terms. □

This concludes the proof of the theorem.

A8 Extensions

The proof from §A7 can also be extended to show the consistency of *OLC* with several other principles, some of which look like attractive strengthenings of the theory.

(i) Part (b) of the theorem establishes the consistency of *OLC* with Involution. As noted in §7, once we have Involution we will almost certainly want De Morgan too. We can achieve this simply by choosing our βη-term structure $\langle \mathscr{N}, [\cdot] \rangle$ to be

one in which $[\wedge] = [\vee]$. (For example, we could set both $[\wedge]$ and $[\vee]$ to be the βη-equivalence class containing a certain constant $*$ of the base language.) Given that $[\neg] = [\lambda p.p]$, this ensures that $[\lambda p.\lambda q.\neg p \wedge \neg q] = [\lambda p.\lambda q.\neg(p \vee q)]$ and $[\lambda p.\lambda q.\neg p \vee \neg q] = [\lambda p.\lambda q.\neg(p \wedge q)]$. Moreover, the first coordinates of the denotations of each of these pairs are also guaranteed to be the same since they are just the corresponding truth functions. Thus their denotations in $\mathfrak{M}$, and hence also in $\mathfrak{M}^{\mathbf{F}}$, will also be the same.

(ii) For every terminal type τ, we can define "lifted" negation, disjunction, and conjunction operators $\neg_\tau, \wedge_\tau, \vee_\tau$ in the obvious way, i.e. $\neg_t = \neg$; $\wedge_t = \wedge$; $\vee_t = \vee$; and

$$\begin{aligned}
\neg_{\sigma\to\tau} &= \lambda x^{\sigma\to\tau}.\lambda z^{\sigma}.\neg_\tau(xz)\\
\wedge_{\sigma\to\tau} &= \lambda x^{\sigma\to\tau}.\lambda y^{\sigma\to\tau}.\lambda z^{\sigma}.xz \wedge_\tau yz\\
\vee_{\sigma\to\tau} &= \lambda x^{\sigma\to\tau}.\lambda y^{\sigma\to\tau}.\lambda z^{\sigma}.xz \vee_\tau yz
\end{aligned}$$

Using these, we can define operators of "logical equivalence" and "nonlogical equivalence", $\overset{L}{\sim}_\tau$ and $\overset{N}{\sim}_\tau$:

$$\begin{aligned}
\overset{L}{\sim}_\tau &=_{\text{df}} \lambda x^{\tau}.\lambda y^{\tau}.(x \wedge_\tau y) \equiv_\tau (x \vee_\tau y)\\
\overset{N}{\sim}_\tau &=_{\text{df}} \lambda x^{\tau}.\lambda y^{\tau}.(x \wedge_\tau \neg x) \equiv_\tau (y \wedge_\tau \neg y)
\end{aligned}$$

In ordered-pair models based on βη-term structures, $|[\![\mathbf{x} \overset{N}{\sim}_\tau \mathbf{y}]\!]|$ will be 1 exactly when $\pi_2(\mathbf{x}) = \pi_2(\mathbf{y})$, since $\pi_1([\![\mathbf{x} \wedge_\tau \neg\mathbf{x}]\!]) = \pi_1([\![\mathbf{y} \wedge_\tau \neg\mathbf{y}]\!])$ for any $\mathbf{x}, \mathbf{y}$, while $\pi_2([\![\mathbf{x} \wedge_\tau \neg\mathbf{x}]\!]^g) = \pi_2([\![\mathbf{x} \wedge_\tau \neg\mathbf{x}]\!]^g)$ only when $\pi_2(\mathbf{x}) = \pi_2(\mathbf{y})$. Moreover, if we base the model on a βη-term structure in which $[\wedge]$ is identical to $[\vee]$, $|[\![\mathbf{x} \overset{L}{\sim}_\tau \mathbf{y}]\!]|$ will be 1 exactly when $\pi_1(\mathbf{x}) = \pi_1(\mathbf{y})$, since $\pi_2([\![\mathbf{x} \wedge_\tau \mathbf{y}]\!]) = \pi_2([\![\mathbf{x} \vee_\tau \mathbf{y}]\!])$ for any $\mathbf{x}, \mathbf{y} \in \mathscr{D}_\tau$. This means that $\overset{L}{\sim}_\tau$ will behave as an equivalence relation in these models.[109] Moreover, each of the following principles will hold:

Two-dimensionality	$(x \overset{N}{\sim}_\tau y) \wedge (x \overset{L}{\sim}_\tau y) \to (x \equiv_\tau y)$
Weak ∧-Commutativity	$(x \wedge_\tau y) \overset{L}{\sim}_\tau (y \wedge_\tau x)$
Weak ∧∨-Distributivity	$(x \wedge_\tau (y \vee_\tau z)) \overset{L}{\sim}_\tau (x \wedge_\tau y) \vee (x \wedge_\tau z)$
Weak ∧∨-Dissolution	$(x \wedge_\tau (y \vee_\tau \neg_\tau y)) \overset{L}{\sim}_\tau x$

We also get dual versions of the last three. All of this seems rather nice.

However, these models also validate the following principle:

Weak Extensionality	$(p \leftrightarrow q) \to (p \overset{L}{\sim}_t q)$

[109] $\overset{N}{\sim}_\tau$ is automatically an equivalence relation in every model.

And this is just crazy. Snow is white if and only if grass is green, but it is certainly not true that for it to be the case that snow is white and grass is green is for it to be the case that snow is white or grass is green, since the latter but not the former would be the case if snow were white and grass were red. Fortunately, the construction can be modified in such a way as to invalidate *Weak Extensionality* while keeping the things that seemed nice. In the modified construction, an arbitrary complete Boolean algebra $\boldsymbol{B}$ takes over the role of $\{0, 1\}$ in Proposition 7.1, with the meet and join operations $\bigsqcap$ and $\bigsqcup$ of $\boldsymbol{B}$ replacing the maximum and minimum operations in the definitions of $[\![\wedge]\!]$, $[\![\vee]\!]$, $[\![\forall]\!]$, and $[\![\exists]\!]$, and the complementation operation $'$ of $\boldsymbol{B}$ replacing the "subtract from 1" operation in the definition of $[\![\neg]\!]$. The valuation $|\cdot|$ is then fixed by any *ultrafilter* on $\boldsymbol{B}$, i.e. a mapping f from $\boldsymbol{B}$ to $\{0, 1\}$ such that for any $X \subseteq \boldsymbol{B}$, $f(\bigsqcup X) = \max\{f(x) : x \in X\}$ and $f(\bigsqcap X) = \min\{f(x) : x \in X\}$, and for any x in $\boldsymbol{B}$, $f(x') = 1 - f(x)$.[110] This change does not disrupt the result that $|[\![\mathbf{x} \overset{L}{\sim}_\tau \mathbf{y}]\!]|$ is 1 exactly when $\pi_1(\mathbf{x}) = \pi_1(\mathbf{y})$ and $|[\![\mathbf{x} \overset{N}{\sim}_\tau \mathbf{y}]\!]|$ is 1 exactly when $\pi_2(\mathbf{x}) = \pi_2(\mathbf{y})$, since in any Boolean algebra, $a \sqcup b = a \sqcap b$ only when $a = b$, while $a \sqcap a' = b \sqcap b'$ for any a and b. So $\overset{L}{\sim}_\tau$ still behaves as an equivalence relation, *Two-dimensionality* still holds, and we still get the weakened Boolean axioms as consequences of the corresponding identities in $\boldsymbol{B}$.

(iii) To show that *OLC* is consistent with *Functionality*, note first that any βη-term structure whose base language contains infinitely many constants in every type is functional: when $\mathbf{x} \neq \mathbf{y} \in \mathscr{D}_{\sigma\to\tau}$ and $\mathbf{z}$ is the βη-equivalence class of a constant that does not occur in the members of $\mathbf{x}$ or the members of $\mathbf{y}$, $\mathbf{xz} \neq \mathbf{yz}$ (BBK, lemma 3.14). Moreover, if the input structure $\langle \mathscr{N}, [\cdot] \rangle$ is functional, the ordered-pair model constructed at step two will be functional too, since when $\mathbf{x} \neq \mathbf{y} \in \mathscr{D}_{\sigma\to\tau}$, either $\pi_2(\mathbf{x}) \neq \pi_2(\mathbf{y})$, in which case $[\![\mathbf{xz}]\!] \neq [\![\mathbf{yz}]\!]$ for any $\mathbf{z}$ such that $\pi_2(\mathbf{x})@\pi_2(\mathbf{z}) \neq \pi_2(\mathbf{y})@\pi_2(\mathbf{z})$, or else $\pi_1(\mathbf{x}) \neq \pi_1(\mathbf{y})$, in which case there is some $\mathbf{z} \in \mathscr{D}_\sigma$ such that $\pi_1([\![\mathbf{xz}]\!]) = \pi_1(\mathbf{x})(\mathbf{z}) \neq \pi_1(\mathbf{y})(\mathbf{z}) = \pi_1([\![\mathbf{yz}]\!])$. However, if all we do at step three is, as before, to throw away elements in the domain that are circular but not O-combinators, there is no guarantee that the final model will still be functional, since it could happen that although $\mathbf{y} \neq \mathbf{z} \in \mathscr{D}_{\sigma\to\tau}$ do not get thrown away, each $\mathbf{z} \in \mathscr{D}_\sigma$ such that $\mathbf{x}@\mathbf{z} \neq \mathbf{y}@\mathbf{z}$ does get thrown away. So to preserve functionality, we need to do something more complicated at step three instead. Briefly: define a partial equivalence relation $\sim_\sigma$ on each $\mathscr{D}_\sigma$ as follows: $\mathbf{x} \sim_e \mathbf{y}$ iff $\mathbf{x} = \mathbf{y}$ and $\pi_2(\mathbf{x})$ is noncircular; $\mathbf{x} \sim_t \mathbf{y}$ iff $\mathbf{x} = \mathbf{y}$ and $\pi_2(\mathbf{x})$ is noncircular; $\mathbf{x} \sim_{\sigma\to\tau} \mathbf{y}$ iff whenever $\mathbf{z} \sim_\sigma \mathbf{w}$, $\mathbf{xz} \sim_\tau \mathbf{yw}$, and each of $\mathbf{x}$ and $\mathbf{y}$ either has a noncircular second coordinate, or is $[\![A]\!]$ for some closed $\mathscr{T}^{\mathrm{O}}$-term A. Instead of just throwing away all circular ele-

[110] If we require $\boldsymbol{B}$ to be atomic, we can call its atoms the "logically possible worlds" and the unique atom $\mathbf{w}$ such that $|\mathbf{w}| = 1$ the "actual world" of the model.

ments that are not O-combinators, we will instead go further by throwing away all elements $\mathbf{x}$ such that it is not the case that $\mathbf{x} \sim \mathbf{x}$. Using the functionality of $\langle \mathscr{N}, [\cdot] \rangle$ we can show that $\pi_2(\mathbf{x}) = \pi_2(\mathbf{y})$ whenever $\mathbf{x} \sim \mathbf{y}$. Also, $\sim$ is a congruence on this model, since $[\![A]\!] \sim [\![A]\!]$ for every closed $\mathscr{F}^{\mathrm{Log}}$ term A. So we can take the quotient by $\sim$. The resulting model is functional; and since each equivalence class consists of ordered pairs with the same second co-ordinate, there is a homomorphism from it to $\langle \mathscr{N}, [\cdot] \rangle$; thus the only circular elements in the quotient model are those whose second coordinates are circular in $\langle \mathscr{N}, [\cdot] \rangle$, all of which are guaranteed to be denoted by closed $\mathscr{F}^{\mathrm{O}}$-terms.

(iv) Because our models are built on top of βη-term structures, Commutativity fails in them.[111] But we can secure Commutativity if we build the non-logical domains $\mathscr{N}$ using equivalence classes, not under βη-equivalence, but under a weaker equivalence relation that allows substitution of $\varphi * \psi$ for $\psi * \varphi$ as well as substitution of terms one of which immediately βη-reduces to the other. (Here $*$ is the constant whose βη-equivalence class is $[\wedge]$ and $[\vee]$). Similarly, if we want Associativity, we can get it by weakening the equivalence relation to allow substitution of $\varphi * (\psi * \theta)$ for $(\varphi * \psi) * \theta$. Neither modification will affect the proof of Proposition 6.12, since these substitutions do not change the number of occurrences of any constant c in the base language. Because of this, every equivalence class will still contain a term A such that $\mathrm{Count}(c, A) \geq \mathrm{Count}(c, B)$ for every other B in the equivalence class and every constant c, which is what we need to prove that CD is never equivalent to D unless C is pure. By contrast, we cannot similarly extend the list of permitted substitutions to secure Distributivity, since (as pointed out in note 73) a sequence of Distributivity-licensed substitutions can increase the number of occurrences of a constant without limit.[112]

[111] Instead they validate $((p \wedge q) \equiv (q \wedge p)) \rightarrow p \overset{L}{\sim} q$.

[112] Note however that if we move to a transfinite relational type theory including α-adic conjunction and disjunction operators $\wedge^\alpha$ and $\vee^\alpha$ for both finite and transfinite ordinals α, the natural generalisations of commutativity and associativity will be inconsistent with the natural generalisation of *OLC*. Generalised associativity would imply that for any p, $\wedge^2(p, \wedge^\omega(p, p, \ldots)) \equiv \wedge^\omega(p, p, \ldots)$, which is ruled out by *OLC* when p is nonlogical. And generalised commutativity for a conjunction operator $\wedge^{2\omega}$ would imply that for any p, $\lambda x_1 x_2 \ldots (\wedge^{2\omega}(x_1, x_2, \ldots, p, p \ldots)) \equiv \lambda x_1 x_2 \ldots (\wedge^{2\omega}(p, x_1, x_2, \ldots, p, p \ldots))$; but transfinite *OLC* rules this out, since the term on the right can be derived from the term on the left and p (which may be nonlogical) by plugging them into the first two argument places of $\lambda z q x_1 x_2 \ldots z(q, x_1, x_2, \ldots)$. How bad is this? Giving up Associativity seems tolerable, but I admit I am more worried about Commutativity. Section 7 tentatively suggested that the binary version of Commutativity might be supported by Ramsey-style thought experiments involving non-linear languages. However much more work would need to be done to properly flesh out such an argument and generalise it to the transfinite setting, since it is not at all clear how to imagine the nonlinear elements of the language coexisting harmoniously with transfinite applications and abstracts.

(v) Since everything in $\mathscr{D}_t$ is noncircular, our models validate $\varphi \not\equiv \varphi \wedge \varphi$ and $\varphi \not\equiv \varphi \vee \varphi$ without any exception for logical φ. As I mentioned in §8, I am not sure whether this is a good thing or not; the worry is that it may seem invidious for $[\![\neg]\!]$ to be circular when $[\![\wedge]\!]$, $[\![\vee]\!]$, $[\![\forall_\sigma]\!]$, and $[\![\exists_\sigma]\!]$ are noncircular. Either way, it would be interesting to find a model of *OLC* in which the denotations of all logical constants are circular. However, there is no easy way to adapt our construction so as to give the *quantifiers* circular denotations: even if we managed to make them circular in the initial $\mathfrak{M}$, the final pruned model $\mathfrak{M}^{\mathbf{F}}$ modifies the denotations of the quantifiers by adding the restrictions by $\mathbf{F}$, and we needed to make each $\mathbf{F}_\sigma$ noncircular guarantee that $\mathbf{F}$ would be internally closed, as it must be for $\mathfrak{M}^{\mathbf{F}}$ to be a well-defined model. So, if we wanted to show the consistency of *OLC* with schemas like $p \equiv p \wedge \exists q(q)$, we would need a rather different kind of construction.

References

Andrews, Peter (1972). 'General Models and Extensionality'. *Journal of Symbolic Logic* 37, pp. 395–7.

Audi, Paul (2012). 'Grounding: Toward a Theory of the In-Virtue-of Relation'. *Journal of Philosophy* 109.12, pp. 685–711.

Bacon, Andrew (MS). 'The Broadest Necessity'. MS.

Bacon, Andrew, John Hawthorne and Gabriel Uzquiano (2016). 'Higher-Order Free Logic and the Prior-Kaplan Paradox'. *Canadian Journal of Philosophy* 46.4-5, pp. 493–541.

Bacon, Andrew and Jeff Sanford Russell (MS). 'The Logic of Opacity'. MS.

Bealer, George (1982). *Quality and Concept*. Oxford: Oxford University Press.

Benzmüller, Christoph, Chad E. Brown and Michael Kohlhase (2004). 'Higher-Order Semantics and Extensionality'. *Journal of Symbolic Logic* 69.4, pp. 1027–88.

Braun, David (1988). 'Understanding Belief Reports'. *Philosophical Review* 107, pp. 555–95.

Carnap, Rudolf (1928). *Der Logische Aufbau der Welt*. Trans. by R. A. George.

Chierchia, Gennaro (1989). 'Anaphora and Attitudes *De Se*'. In *Language in Context*, ed. Renate Bartsch, J. F. A. K. van Benthem and P. van Emde Boas. Dordrecht: Foris, pp. 1–31.

Correia, Fabrice (2010). 'Grounding and Truth-functions'. *Logique et Analyse* 53, pp. 211–51.

— (2016). 'On the Logic of Factual Equivalence'. *Review of Symbolic Logic* 9, pp. 103–22.

Correia, Fabrice and Alexander Skiles (MS). 'Grounding, Essence, and Identity'. MS.

Cresswell, Maxwell J. (1965). 'Another Basis for S4'. *Logique et Analyse* 8, pp. 191–5.

— (1985). *Structured Meanings*. Cambridge, MA: MIT Press.

Dorr, Cian (2004). 'Non-Symmetric Relations'. In *Oxford Studies in Metaphysics*, vol. 1, ed. Dean Zimmerman. Oxford: Oxford University Press, pp. 155–92.

— (2005). 'What We Disagree About When We Disagree About Ontology'. In *Fictionalist Approaches to Metaphysics*, ed. Mark Kalderon. Oxford: Oxford University Press, pp. 234–86.

— (2007). 'There Are No Abstract Objects'. In *Contemporary Debates in Metaphysics*, ed. Theodore Sider, John Hawthorne and Dean Zimmerman. Malden, Mass.: Wiley-Blackwell, pp. 32–64.

Dorr, Cian (2014a). 'Quantifier Variance and the Collapse Theorems'. *The Monist* 97, pp. 503–70.
— (2014b). 'Review of Rayo 2013'. *Notre Dame Philosophical Reviews* 2014.06.33.
— (2014c). 'Transparency and the Context-sensitivity of Attitude Reports'. In *Empty Representations: Reference and Non-existence*, ed. Manuel García-Carpintero and Genoveva Martí. Oxford: Oxford University Press, pp. 25–66.
Dorr, Cian and John Hawthorne (2013). 'Naturalness'. In *Oxford Studies in Metaphysics*, vol. 8, ed. Karen Bennett and Dean Zimmerman. Oxford: Oxford University Press, pp. 3–77.
— (2014). 'Semantic Plasticity'. *Philosophical Review* 123, pp. 281–338.
Elbourne, Paul (2005). *Situations and Individuals*. Cambridge, MA: MIT Press.
Fine, Kit (2012). 'Guide to Ground'. In *Metaphysical Grounding*, ed. Fabrice Correia and Benjamin Schneider. Cambridge: Cambridge University Press, pp. 37–80.
— (2015). 'Unified Foundations for Essence and Ground'. *Journal of the American Philosophical Association* 1.2, pp. 296–311.
Fritz, Peter and Jeremy Goodman (forthcoming). 'Higher-Order Contingentism, Part 1: Closure and Generation'. *Journal of Philosophical Logic*. Forthcoming.
Goodman, Jeremy (MS). 'How Fine-Grained is Reality?' MS.
— (forthcoming). 'Reality Is Not Structured'. *Analysis*. Forthcoming.
— (2016). 'Quantificational Logic and Necessitism'. *Philosophical Perspectives* this volume.
Goodman, Nelson (1954). *Fact, Fiction and Forecast*. Fourth. Cambridge, MA: Harvard University Press.
Graff, Delia (2001). 'Descriptions as Predicates'. *Philosophical Studies* 102, pp. 1–42.
Henkin, Leon (1950). 'Completeness in the Theory of Types'. *The Journal of Symbolic Logic* 15.
Hindley, J. Roger and Jonathan P. Seldin (2008). *Lambda Calculus and Combinators: An Introduction*. Cambridge: Cambridge University Press.
Hodes, Harold T. (2015). 'Why Ramify?' *Notre Dame Journal of Formal Logic* 56.2, pp. 379–415.
Huntingdon, Edward V. (1904). 'Sets of Independent Postulates for the Algebra of Logic'. *Transactions of the American Mathematical Society* 5.3, pp. 288–309.
Khamara, E.J. (1988). 'Indiscernibles and the Absolute Theory of Space and Time'. *Studia Leibnitiana* 20, pp. 140–59.
Kripke, Saul (1972). *Naming and Necessity*. revised. Cambridge, MA: Harvard University Press.

Kusumoto, Kiyomi (1999). 'Tense in Embedded Contexts'. PhD thesis. University of Massachusetts, Amherst.
Lewis, David (1970). 'General Semantics'. *Synthese* 22, pp. 18–67.
— (1986). *On the Plurality of Worlds*. Oxford: Blackwell.
Muskens, Reinhard (2007). 'Intensional Models for the Theory of Types'. *Journal of Symbolic Logic* 72, pp. 98–118.
Plantinga, Alvin (1983). 'On Existentialism'. *Philosophical Studies* 44, pp. 1–20.
Plate, Jan (2016). 'Logically Simple Properties and Relations'. *Philosopher's Imprint* 16, pp. 1–40.
Prior, A. N. (1964). 'Conjunction and Contonktion Revisited'. *Analysis* 24, pp. 191–5.
— (1971). *Objects of Thought*. Ed. Peter T. Geach and Anthony J. P. Kenny. Oxford: Clarendon Press.
Ramsey, F. P. (1926). 'The Foundations of Mathematics'. *Proceedings of the London Mathematical Society*. Series 2 25, pp. 338–84.
— (1927). 'Facts and Propositions'. *Proceedings of the Aristotelian Society* 7, pp. 153–70.
Rayo, Agustín (2013). *The Construction of Logical Space*. Oxford: Oxford University Press.
Rosen, Gideon (2010). 'Metaphysical Dependence: Grounding and Reduction'. In *Modality: Metaphysics, Logic, and Epistemology*, ed. Bob Hale and Aviv Hoffmann. Oxford: Oxford University Press.
Russell, Bertrand (1903). *The Principles of Mathematics*. London: Routledge.
Salmon, Nathan (1986a). *Frege's Puzzle*. Cambridge, MA: MIT Press.
— (1986b). 'Reflexivity'. *Notre Dame Journal of Formal Logic* 27, pp. 401–29.
— (2010). 'Lambda in Sentences with Designators'. *Journal of Philosophy* 107, pp. 445–68.
Saul, Jennifer (2007). *Simple Sentences, Substitution, and Intuitions*. Oxford: Oxford University Press.
Schiffer, Stephen (1979). 'Naming and Knowing'. In *Midwest Studies in Philosophy II: Contemporary Perspectives in the Philosophy of Language*, ed. Peter French, Theodore E. Uehling Jr. and Howard K. Wettstein. Minneapolis: University of Minnesota Press, pp. 28–41.
Setiya, Kieran (2007). *Reasons Without Rationalism*. Princeton: Princeton University Press.
Shoemaker, Sydney (1998). 'Causal and Metaphysical Necessity'. *Pacific Philosophical Quarterly* 79, pp. 59–77.

Sider, Theodore (2011). *Writing the Book of the World*. Oxford: Oxford University Press.
Simons, Mandy et al. (2010). 'What Projects and Why'. *Proceedings of SALT* 20, pp. 309–27.
Soames, Scott (1987a). 'Direct Reference, Propositional Attitudes, and Semantic Content'. *Philosophical Topics* 15, pp. 47–87.
— (1987b). 'Substitutivity'. In *On Being and Saying: Essays for Richard Cartwright*, ed. Judith Jarvis Thomson. Cambridge: MIT Press, pp. 99–132.
Sorensen, Roy A. (1999). 'Mirror Notation: Symbol Manipulation Without Inscription Manipulation'. *Journal of Philosophical Logic* 28, pp. 141–64.
Stalnaker, Robert C. (1994). 'The Interaction of Modality with Quantification and Identity'. In. *Ways a World Might Be*. Oxford: Oxford University Press, pp. 144–61.
— (1999). *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford: Oxford University Press.
Suszko, Roman (1975). 'Abolition of the Fregean Axiom'. *Lecture Notes in Mathematics* 453, pp. 169–239.
Whitehead, Alfred North and Bertrand Russell (1910). *Principia Mathematica*. Second. Cambridge: Cambridge University Press.
Williamson, Timothy (1985). 'Converse Relations'. *Philosophical Review* 94, pp. 249–62.
— (2003). 'Everything'. In *Philosophical Perspectives 17: Language and Philosophical Linguistics*, ed. John Hawthorne and Dean Zimmerman. Oxford: Blackwell, pp. 415–65.
— (2013). *Modal Logic as Metaphysics*. Oxford: Oxford University Press.
— (2016). 'Abductive Philosophy'. *Philosophical Forum* 47, pp. 263–80.