Hyperintensionalism and overfitting: a test case

DANIEL KODSI
Trinity College, University of Oxford, UK
daniel.kodsi@philosophy.ox.ac.uk

1. Introduction

When, in general, are some properties identical?[1] A familiar answer is simply: when they are necessarily equivalent. Necessarily, properties that hold of the same things in the same possible circumstances are identical. That claim constitutes the doctrine of *intensionalism*. Correspondingly, *hyperintensionalism* is the negation of intensionalism. Hyperintensional theorising, driven partly by dissatisfaction with intensionalism and partly by exciting new tools for exploring the nature of properties, forms a growing research programme in contemporary metaphysics.

Enthusiasm for hyperintensionalism is not universal, however. In particular, in recent work, Timothy Williamson has argued that hyperintensional theorising is guilty of *overfitting*, a well-known pathology in the natural and social sciences whereby models of some phenomenon are fitted too closely to the data and hence are inadvertently fitted to noise in the data (Williamson 2021, 2024). Biases in methods for generating data are a characteristic cause of overfitting. Increases in the complexity (or 'degree of freedom') of models are a characteristic symptom. More holistically, overfitting tends to involve *ad hoc* complications. For a model that has been gerrymandered to fit biased data typically has poor predictive power. It is surprised by new data. Extra parameters must be added to the model on the spot to restore adequacy of fit. There is nothing in the background guiding the changes.

Williamson argues that hyperintensional theorising generally fits the profile of overfitting. This paper argues, more narrowly, that the hyperintensional theory developed by Cian Dorr in 'To be F is to be G' does so (Dorr 2016). Although the focus here is limited, the target is a natural one. Dorr's agenda-setting paper did much not only to generate interest in the topic of the identity-conditions of properties but to encourage exploring it in a *higher-order* formal setting, in which quantification into predicate position replaces first-order quantification over properties. Further, though Dorr's own approach is explicitly exploratory, the theory that he discusses remains (nearly a decade later) one of the best-developed forms of hyperintensionalism in a higher-order setting. As an experiment in higher-order hyperintensionalism, it is among the most rigorous to date. It is also sometimes advertised as illustrating the feasibility of

[1] This paper uses 'property' as a catch-all term for properties, propositions (or states of affairs), and relations. This terminological choice is elaborated in §2 below.

alternative approaches to intensionalism.[2] That makes it an appropriate test case of the prospects for contemporary forms of hyperintensionalism.

The plan is as follows. §2 presents the formal background relevant to the critique of Dorr's theory. §3 presents the relevant methodological background. Taken together, they aim to provide one way of understanding the disagreement between intensionalism and hyperintensionalism. §4 is an extended critical discussion of Dorr's theory. §5 draws provisional morals.

2. Formal background

This section explains one way of formalising talk of properties, their identity-conditions and in turn intensionalism. This is in the setting of higher-order logic. Higher-order resources do not provide the only way of formulating intensionalism, but they facilitate precise statements of competing hypotheses about the nature of properties. They have also recently enjoyed increased uptake in metaphysics; Dorr's 'To be F is to be G' is an influential case in point.[3] In any case, this paper works in a higher-order language in order to engage with Dorr's theory on its own terms.

More exactly, this paper proceeds in a simply, relationally typed lambda calculus. For present purposes, the relevant facts about this language are as follows:

- Each term in the language has a *type*. The set of types is defined recursively by the rule that e is a type, and for any finite number $n \geq 0$ and types $\sigma_1, \ldots, \sigma_n$, the sequence $\langle \sigma_1, \ldots, \sigma_n \rangle$ is a type; nothing else is a type.
- The base type e is the type of *singular terms*. The complex types $\langle \sigma_1, \ldots, \sigma_n \rangle$ for $n \geq 0$ are the types of *n-place predicates*. In particular, $\langle\rangle$ (the special case for $n = 0$) is the type of zero-place predicates, or *sentences*, and more generally of *formulae*. Terms of type $\langle \sigma_1, \ldots, \sigma_n \rangle$ combine with terms of type $\sigma_1, \ldots, \sigma_n$ in that order to form formulae.
- The language has infinitely many variables of each type, as well as constants for the truth-functional connectives $\neg$ of type $\langle\langle\rangle\rangle$ and $\wedge$, $\vee$ and $\rightarrow$ and $\leftrightarrow$ of type $\langle\langle\rangle,\langle\rangle\rangle$. It also has a constant $\Box$ of type $\langle\langle\rangle\rangle$, which may informally be understood as expressing metaphysical necessity.

[2] For instance, Skiba 2021 identifies Dorr's theory as one of 'three important approaches' to theorising about higher-order identity. For interesting recent developments in higher-order hyperintensionalism, which take as their respective starting points the notions of essence and aboutness, see Ditter manuscript-a, manuscript-b; Goodman 2024, manuscript.

[3] For an introduction to and exemplifications of higher-order metaphysics, see the essays collected in Fritz and Jones 2024. Williamson 2013 is another important recent precedent.

- The language has a variable-binding operator λ such that, for any $n > 0$, when $z_1, \dots, z_n$ are pairwise distinct variables of types $\sigma_1, \dots, \sigma_n$ and A is a formula (in which $z_1, \dots, z_n$ may but need not occur free), $\lambda z_1, \dots, z_n(A)$ is a term of type $\langle \sigma_1, \dots, \sigma_n \rangle$. For example, where z is a variable of type e, $\lambda z(\mathrm{Tall}(z) \rightarrow \mathrm{Tall}(z))$ is a term of type $\langle e \rangle$, which may be read 'is such that it is tall if it is tall'.

- For every type σ, the language has a constant $\forall_\sigma$ of type $\langle\langle \sigma \rangle\rangle$, which may informally be understood as expressing the higher-order property of holding of every entity of type σ. The constants $\forall_\sigma$ play the role of universal quantifiers. For example, $\forall_e \lambda z(\mathrm{Tall}(z))$ may be read 'the property of being tall holds of every object', or more simply, 'every object is tall'.

For convenience, we make free use of infix notation, typically suppress type subscripts when the context permits and are similarly loose about use–mention. As usual, $\Diamond A$ abbreviates $\neg\Box\neg A$, $\forall_\sigma z\, A$ abbreviates $\forall_\sigma \lambda z(A)$ and $\exists_\sigma z\, A$ abbreviates $\neg\forall_\sigma z\, \neg A$.

The characterisation of the semantic values of predicates and the quantifiers in terms of properties was also done for convenience. The relation between quantification over properties in a natural language such as English and quantification into predicate position in the formal language is notoriously vexed. Plausibly, the former is simply a form of restricted first-order quantification. In contrast, both the first-order universal quantifier $\forall_e$ and higher-order universal quantifiers $\forall_\sigma$ for $\sigma \neq e$ are intended to be in a suitable sense unrestricted. Accordingly, talk of properties in this paper plays a primarily heuristic role, as a means of paraphrasing relevant formal claims, and similarly for talk of propositions and relations, which are assimilated to n-place properties for $n = 0$ and $n \geq 2$, respectively.

We also adopt a simple background higher-order quantified modal logic. Thus, we assume for simplicity and clarity that the standard classical conditions on the first-order universal quantifier are extended to the universal quantifier at each type, as well as that the logic of □ extends the modal system S4.[4] Completing the classical picture, we also adopt the principle of *Extensional β*, governing the behaviour of λ-terms:[5]

$$E\beta \qquad \lambda z_1 \dots z_n(A)(a_1 \dots a_n) \leftrightarrow A[a_1/z_1, \dots, a_n/z_n]$$

[4] In fact, the characteristic axioms of S4—the K axiom ($\Box(A \rightarrow B) \rightarrow \Box A \rightarrow \Box B$), the T axiom ($\Box A \rightarrow A$) and the 4 axiom ($\Box A \rightarrow \Box\Box A$)—can all be derived from intensionalism given the necessitation rule for □. Although it does not entail it, intensionalism also makes the 5 axiom ($\Diamond A \rightarrow \Box\Diamond A$) equivalent to the necessity of propositional distinctness ($A \neq B \rightarrow \Box(A \neq B)$). Its simplifying effect on the logic of □, explored in detail in Bacon and Dorr 2024, is an important benefit of intensionalism.

[5] Unless otherwise specified, in displayed schemas such as *Eβ*, A can be any formula and other terms can be of any type, provided that the result is well-formed (e.g., in *Eβ*, $z_1, \dots, z_n$ can be variables of any types $\sigma_1, \dots, \sigma_n$, respectively).

Here $A[a_1/z_1, \ldots, a_n/z_n]$ is the result of substituting $a_1, \ldots, a_n$ for $z_1, \ldots, z_n$, respectively, in A, provided that no variable free in any of $a_1, \ldots, a_n$ thereby becomes bound. Since the logic is closed under necessitation, *Eβ* ensures that any sentence which consists in a λ-term predicated of some arguments and the sentence which results from applying that λ-term to those arguments are provably necessary equivalent. For example, for any one-place predicates of objects F and G and singular term a, we may derive the necessitated biconditional $\Box(\lambda x(Fx \wedge Gx)a \leftrightarrow (Fa \wedge Ga))$. Informally: necessarily, a is such that it is F and it is G if and only if a is F and a is G.

Given the rule of necessitation for □ and the classical behaviour of the quantifiers, we may also derive from *Eβ* a natural strong property comprehension principle:

Comp $\exists X\, \Box \forall z_1 \ldots \forall z_n\, (Xz_1 \ldots z_n \leftrightarrow A)$, where A can be any formula in which X does not occur free

In effect, *Comp* implies that every meaningful predicate modally corresponds to some property of the relevant type. For instance, the predicate 'Tall' corresponds to some property X such that necessarily, for any object x, Xx if and only if Tall(x).[6]

We are now ready to return to our main topic: the identity-conditions on properties. In the present setting, the relation of identity between properties can be formalised perspicuously using λ-abstraction and higher-order quantification. More exactly, for every type σ, we may define a constant $=_\sigma$ which plays the logical role of identity for entities of type σ by the following schema:

Identity $=_\sigma \coloneqq \lambda xy(\forall Z\, (Zx \leftrightarrow Zy))$

For example, $=_e$ stands for first-order identity, the relation which holds of some objects a and b just in case all the same properties hold of a and b. Analogously, $=_\sigma$ for $\sigma \neq e$ stands for identity for properties of type σ, the higher-order relation which holds of some properties F and G of type σ just in case all the same higher-order properties hold of F and G. We will refer to such higher-order identity relations as *equality relations* and to the subject matter they compose as *equality*, henceforth reserving the unmodified term 'identity' for first-order identity.[7]

[6] For an extended discussion and defence of *Comp*, see Williamson 2013: §6. Unrestricted comprehension principles are one of the distinctive benefits of going higher-order; in a first-order setting, analogues of *Comp* notoriously give rise to Russell's paradox.

[7] Of course, if each $=_\sigma$ is to be appropriately regarded as the logical analogue of identity for entities of type σ, the natural generalisations of the standard schemas governing identity should be valid. These are *Ref* ($a =_\sigma a$) and *LL* ($a =_\sigma b \rightarrow (A \leftrightarrow A[b/a]$). It is easy to see that *Ref* is valid given *Identity* and *Eβ*, given that $\forall Z\, (Za \leftrightarrow Za)$. Equally, *LL* is valid, since $\exists X\, \neg(Xa \leftrightarrow Xb)$ whenever $\neg(A \leftrightarrow A[b/a])$ by *Comp*. Hence given the background logic, each $=_\sigma$ indeed behaves logically as it should.

Finally, with the equality relations so defined, we are in a position to provide a similarly precise and perspicuous formalisation of intensionalism. For present purposes, we understand it as the following simple theory of equality:

$$\textit{Intensionalism} \qquad \Box\forall X\ \forall Y\ (\Box\forall z_1\ \ldots\ \forall z_n\ (Xz_1 \ldots z_n \leftrightarrow Yz_1 \ldots z_n) \rightarrow X = Y)$$

Informally: necessarily, any properties that necessarily hold of all the same things are equal. Since the background logic includes the converse principle (i.e., necessarily, any properties that possibly hold of different things are not equal), it follows that being necessarily equivalent is necessarily equivalent to being equal. Thus, by an application of *Intensionalism* itself, necessary equivalence is *equal* to equality.

Intensionalism is strong and simple. By standard abductive criteria on theory choice, such features count crucially in its favour. Some of intensionalism's virtues will partly emerge in the course of discussing Dorr's theory. First, however, it will be worth putting the disagreement between intensionalism and hyperintensionalism in more general methodological perspective.

3. Methodological background

This section explains one central battleline in the disagreement between intensionalism and hyperintensionalism, with an eye towards underlying methodological issues. Following recent work by Timothy Williamson, to do so, it borrows terms from the natural and social sciences for thinking about the trade-off between the abductive virtue of fit with the evidence (or data) and the more structural abductive virtues of simplicity, elegance and informativeness.

In outline, the dialectical situation to be elaborated is this. Hyperintensionalists may hope to pin on intensionalists a diagnosis of *underfitting* the data. On that diagnosis, intensionalism is simple because it is insensitive to genuine distinctions between properties. Conversely, intensionalists may hope to pin on hyperintensionalists a diagnosis of *overfitting* the data. On that diagnosis, hyperintensional theories are complicated because they are sensitive to spurious distinctions between properties.

Prima facie, the diagnoses of underfitting and overfitting are symmetrical. Further, in scientific practice, both problems happen. Worse, it is often hard to tell which one is happening. Intensionalists and hyperintensionalists should agree that a verdict can only be reached by looking in detail at the alleged counterexamples to intensionalism and the typical results of hyperintensional theorising. More specifically, if hyperintensional theorising results in minimal extra complexity and the data which motivate it withstand critical scrutiny, that is bad news for intensionalists. But if hyperintensional theorising enacts *ad hoc* complications and the data which motivate it can be shown independently unreliable, that is bad news for hyperintensionalists.

The rest of this section provides a simple and relatively familiar illustration of the dialectic just outlined. That will help give a sense of what to expect from new-wave hyperintensional theorising of the kind modelled by Dorr.

In more detail: consider the sentences 'It is raining or it is not raining' and 'It is snowing or it is not snowing'. Let 'R' and 'S' abbreviate 'It is snowing' and 'It is raining', respectively. Then uncontroversially, the propositions $R \vee \neg R$ and $S \vee \neg S$ are necessarily equivalent, since each is necessary. Hence, given intensionalism, $R \vee \neg R$ equals $S \vee \neg S$. More generally, in any normal modal logic, intensionalism entails that every tautology expresses the same proposition.

However, it is hardly obvious that $R \vee \neg R$ equals $S \vee \neg S$. For example, it is tempting to judge that the proposition R is a constituent of $R \vee \neg R$ but that it is not a constituent of $S \vee \neg S$. On the natural formalisation of that pair of claims, they entail that $R \vee \neg R$ and $S \vee \neg S$ do not have all the same properties, and so are not equal.

Such apparent counterexamples to intensionalism are familiar. Just as familiarly, they seem to suggest a *structured* picture of the higher-order realm, on which propositions (and in principle properties of other types) are built up from more basic constituents. More specifically, propositions are conceived as having internal structure analogous to the semantic structure of the sentences which express them. In a higher-order setting, this idea is partly captured by the following constraint:

$$\textit{Structure}_{\textit{M}} \qquad \forall X\, \forall Y\, \forall z_1\, \forall z_2\, (Xz_1 = Yz_2 \rightarrow X = Y \wedge z_1 = z_2)$$

Informally: propositions with the structure of monadic predications are equal only if each consists in the same property predicated of the same entity. For example, if $A \neq B$, then $\lambda X(X \vee \neg X)A \neq \lambda X(X \vee \neg X)B$, even though $\Box(\lambda X(X \vee \neg X)A \leftrightarrow \lambda X(X \vee \neg X)B)$ given the background logic. Similarly, the natural polyadic generalisation of *Structure*$_{M}$ has the desired result that whenever $A \neq B$, $(A \vee \neg A) \neq (B \vee \neg B)$.[8]

However, there is a problem for the structured theory of propositions. Namely, as much recent work in higher-order metaphysics has emphasised, *Structure*$_{M}$ is inconsistent in classical higher-order logic.[9] Very briefly, the inconsistency arises since the instance of *Structure*$_{M}$ in which z_1 and z_2 are variables of type $\langle\rangle$ entails that for any proposition A and properties of propositions O_1 and O_2, whenever O_1does not equal O_2, $O_1 A$ does not equal $O_2 A$. In effect, given *Structure*$_{M}$, there are as many propositions as properties of propositions. Yet given classical higher-order logic, one can show by

[8] Note, however, that the simplest such generalisation also entails $A \neq B \rightarrow (A \vee B) \neq (B \vee A)$: a perhaps less desirable result, though not obviously one that can be avoided without gerrymandering, as readers are invited to test for themselves (for first steps, see Dorr 2016: 122n43; Fritz 2023: 193–4).

[9] For recent formal statements and relevant philosophical discussion of the result, which traces back to Russell 1903 (Appendix B), see Hodes 2015; Dorr 2016: 63–4; Goodman 2017; Dorr et al 2021: §7.4; Fritz 2017, 2023: §3; Williamson 2024: §3.5.

Cantorian reasoning that there are more properties of propositions than propositions. Thus, *Structure*$_M$ is inconsistent on simple and natural background assumptions. That striking result is widely known as the *Russell-Myhill paradox*.

From a formal perspective, it is natural to be interested in which candidate weakenings of either classical higher-order logic or *Structure*$_M$ (or perhaps both) serve to restore consistency. For example, inconsistency still results in a higher-order language without λ-terms if the logic contains an unrestricted comprehension principle for properties or pluralities of propositions (Uzquiano 2015; Fritz et al. 2021). But it is to be expected that continued enquiry will discover candidate restrictions to classical higher-order logic or to conjectured constraints on equality such as *Structure*$_M$ that do suffice to restore consistency. Surveying the space of formally available options is far from the aim of this paper.[10]

Rather, our present interest is in the methodological implication of the Russell-Myhill paradox. To restore consistency, structured hyperintensionalists will have to complicate, in one way or another. For example, if the structured theorist works in a lambda calculus, they might build exceptions into *Eβ* for certain (allegedly) special cases. Each such exception should be registered as an increase in complexity relative to the unrestricted schema, until it is shown (without hand-waving) how the restrictions can fruitfully be unified. Alternatively, the structured theorist might restrict the classical rule of *UI*, to block the derivation of a comprehension principle strong enough to generate the paradox. Other options for restoring consistency will involve increases in complexity elsewhere in the system; none will be complication-free.

The need for structured hyperintensionalists to complicate hardly refutes the claim that there are differences in structure between necessarily equivalent propositions, let alone the minimal claim that there are differences—perhaps cognitive or explanatory, rather than specifically structural ones—between necessarily equivalent properties. But it does support the critique of hyperintensional theorising as an exercise in overfitting, since overfitting characteristically leads to increases in complexity. Relatedly, the inconsistency of *Structure*$_M$ suggests that natural judgements of difference in structure are less than fully reliable: at least some must be mistaken.[11]

One factor that makes the label 'overfitting' especially suggestive is the way in which structured hyperintensionalists need to complicate *in order to* restore consistency: in

[10] For rigorous exploration of the territory and a series of limitative results, see Fritz 2020, 2021, 2022, 2023. For an attempt to avoid the result by significantly restricting the λ-calculus (with corresponding restrictions required to *Comp*), see Bacon 2023.

[11] In principle, the structured theorist might argue that natural judgements of structure really support some weaker, consistent theory (see, e.g., Bacon 2023). But that may well be to underestimate their own ingenuity in constructing such a theory. Williamson 2024: §3.5 sketches alternative paradoxes in the aim of illustrating the internal incoherence of the structured picture.

effect, a bare modicum of fit with the evidence. The complications are made *ad hoc*. It may well be that some ways of complicating are less unnatural than others, given the picture of propositions as structured similarly to sentences. However, that doesn't change the fact that the structured theory of propositions has been taken (badly!) by surprise. The situation is reminiscent of how a model overly sensitive to biases in the data set on which it was initially trained will often do badly when tested against new data. Extra parameters are added on the spot to restore fit. Often, theorists have only the flimsiest sense of why the model's predictions were so off.

It is crucially relevant here that an elegant alternative to the structured theory of propositions is ready to hand: intensionalism. Quite generally in science, the simplicity and explanatory power of well-known live options makes a large difference to abductive assessment. For on an abductive methodology, theories are *compared* on dimensions like simplicity and strength to their competitors. Unsurprisingly, adding epicycles to one theory to avoid a catastrophic clash with observation is most clearly bad practice when its rivals have no need for anything like them.

Indeed, the baseline set by leading theories in a domain helps determine what ideas scientists take seriously in the first place. One way to see the point is by considering increases in scientific knowledge over time (compare Williamson 2017: 342). A ragbag of local generalisations about, say, sexual selection in certain species of bird (for instance, about when and why females prefer flashy males) may be good enough to count as a 'theory' on day one. It will certainly not be treated as one if some way of unifying the chief of the generalisations has long since been known (for instance, in signalling theory). More generally, research programmes in science do not normally try to start from scratch in the way that some research programmes in philosophy do. If a scientific research programme not only tried to start from scratch but required gerrymandering even to get off the ground, it would probably die a swift death.

Such methodological remarks help provide a segue to the second part of this paper. For we now turn from structured hyperintensionalism to Dorr's theory of equality in 'To be F is to be G'. The theory is not the result of adding epicycles to the structured theory of propositions. Indeed, it is partly motivated by the methodological pressure to start somewhere else, given the Russell-Myhill paradox. The trouble is that theorising informed *neither* by intensionalism *nor* by the idea that propositions are like sentences risks being badly unconstrained, so quickly lapsing into overfitting.

4. Dorr's theory of equality

This section critically discusses the theory of equality that Cian Dorr develops in 'To be F is to be G'. Dorr elaborates the theory in a spirit of exploration; it is tempting to say 'experimentation'. At any rate, it is engaged with here as something like an experiment in theory-building. Ultimately, the aim is to bring out some overfitting-related traps for hyperintensional experiments with the logic of equality. The plan is as follows. §4.1 discusses Dorr's intensional commitments: principles which follow from intensionalism, but which he motivates as independently plausible. §4.2 discusses their interaction with Dorr's distinctive hyperintensional commitment. Finally, §4.3 discusses Dorr's motivation for the hyperintensional principle itself.[12]

4.1 Involution + Commutativity + β-Equality

Dorr rejects the structured theory of propositions, most firmly on the basis of the Russell-Myhill paradox. That enables him to accept some consequences of intensionalism ruled out by the structured theory, but which might seem attractive considered in isolation. The question is: which intensional conditions to accept?

Dorr himself provides positive arguments for the following conditions, each of which is easily seen to be entailed by intensionalism (given the background logic):

Involution $\lambda z_1 \dots z_n(A) =_\sigma \lambda z_1 \dots z_n(\neg\neg A)$

Commutativity$_\wedge$ $\lambda z_1 \dots z_n(A \wedge B) =_\sigma \lambda z_1 \dots z_n(B \wedge A)$

Commutativity$_\vee$ $\lambda z_1 \dots z_n(A \vee B) =_\sigma \lambda z_1 \dots z_n(B \vee A)$

β-Equality $\lambda z_1 \dots z_n(A)(a_1 \dots a_n) =_{\langle\rangle} A[a_1/z_1, \dots, a_n/z_n]$[13]

Let *Commutativity* be the conjunction of *Commutativity*$_\wedge$ and *Commutativity*$_\vee$; for simplicity, we may count it as a single principle. Then Dorr in effect takes on three distinct intensional commitments.[14] Each is natural on the assumption that the structured theory of propositions is to be rejected. Without appealing to forms of judgement which ultimately motivate the structured theory of propositions, it is not easy to provide a non-gerrymandered generalisation which rules them out. At the same time, however,

[12] All page numbers in this section refer to Dorr 2016 unless otherwise specified. Some minor notational changes have been made for uniformity. In all quotes, any emphasis is in the original.

[13] Dorr's discussion extends to the stronger *β-conversion* schema, according to which $A \leftrightarrow A'$ is provable whenever A' is the result of substituting $B[a_i/z_i]$ for some constituent of the form $\lambda z_1 \dots z_n(B)(a_1 \dots a_n)$ in A. It too follows from intensionalism in the background logic, albeit less obviously. (Thanks here to Alex Roberts and Tim Williamson.) *β-Equality* is featured in the text for presentational convenience, since it has the various instances of *β-conversion* to be mentioned below as immediate consequences.

[14] In fact, Dorr only accepts *β-Equality* in restricted form. Discussion of that complication is deferred to the following subsection. He also provisionally endorses the De Morgan laws formulated as equations, since they are entailed by *Involution* 'when $\vee$ is interpreted as $\lambda XY\neg(\neg X \wedge \neg Y)$ or when $\wedge$ is interpreted as $\lambda XY\neg(\neg X \vee \neg Y)$, 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' (68).

on the assumption that intensionalism is also to be accepted, it is not easy to provide a non-gerrymandered generalisation which entails them.

In brief, theorists who accept neither the structured theory of propositions nor intensionalism face a dilemma. On the one hand, they risk lacking a principled reason not to adopt some minimal intensional commitments. On the other, each intensional commitment that they do adopt risks counting for them as an increase in complexity, since they are in effect starting from scratch.

Continuing to focus on Dorr's case, we can sharpen the latter horn of this dilemma by considering some potential ways of bringing his different intensional commitments under a unifying generalisation. In the present context, the most salient such generalisation is intensionalism itself, which underpins a uniform potential explanation of all three of *Involution*, *Commutativity* and *β-Equality*. That is, given intensionalism, the reason that equations of the forms entailed by *Involution*, *Commutativity* and *β-Equality* are true is simply that the properties expressed by their flanking terms are necessarily equivalent. Clearly, however, such a uniform explanation of *Involution*, *Commutativity* and *β-Equality* is not available to theorists who do not accept intensionalism.

We may further sharpen the issue by asking whether there is a natural way for Dorr to unify any two of *Involution*, *Commutativity* and *β-Equality*. In particular, let us grant that *β-Equality* is simple and informative enough in its own right not to require further explanation. What, then, about *Involution* and *Commutativity*? Borrowing from Dorr's own discussion, a simple initial observation is that both follow from the more general claim that any truth-functionally equivalent predicates express the same property. Dorr calls this generalisation *Booleanism* (67):

Booleanism $\lambda z_1 \ldots z_n(A) =_\sigma \lambda z_1 \ldots z_n(B)$ whenever $A \leftrightarrow B$ is a theorem of classical propositional logic

As briefly indicated in §3, *Booleanism* follows from adding *Intensionalism* to any normal modal logic. But even considered on its own, *Booleanism* underpins a uniform potential explanation of *Involution* and *Commutativity*. In effect, given *Booleanism*, the reason that equations of the forms entailed by *Involution* and *Commutativity* are true is that their flanking terms are truth-functionally equivalent. That potential explanation too sits somewhat uneasily with the rejection of intensionalism, since *Booleanism* suffices for a version of the idea that necessarily equivalent propositions are equal (though not the significantly stronger principle of *Intensionalism* itself). Once one has gone that far in the direction of intensionalism, why not go all the way?[15]

[15] More precisely, let $\top$ be an arbitrary tautology and let $\Box_\top$ be the property of being equal to $\top$, i.e. $\lambda X(X =_{\langle\rangle} \top)$. Then *Booleanism* suffices to prove the schema $\Box_\top(A \leftrightarrow B) \rightarrow (A =_{\langle\rangle} B)$. Since $\Box_\top$ satisfies natural conditions on being the broadest form of necessity (where something is the broadest form of

In any case, part of Dorr's own interest in 'To be F is to be G' is in exploring consistent alternatives to *Booleanism* conceived as such (rather than as alternatives to intensionalism). Given that aim, the question is what distinguishes the instances of *Booleanism* he accepts from those he does not. Notably, the positive argument that Dorr provides for *Involution* and *Commutativity* fails to suggest an answer.

More specifically, Dorr motivates *Involution* by considering a possible language in which, to express the negation of some formula, it is written upside-down (c.f. Ramsey 1927: 42–3). In such a language, the result of negating A twice would be indiscernible from A. Hence, it would be impossible to truly express $\lambda z_1 \ldots z_n(A) \neq \lambda z_1 \ldots z_n(\neg\neg A)$. The key judgement in support of *Involution* is then that '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' (63). Later, Dorr claims that '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*' (68).

For present purposes, the relevant observation about this form of argument is that it does not suggest any suitable generalisation which discriminates *Involution* and *Commutativity* from other instances of *Booleanism*. The simplest, albeit quite schematic candidate seems to be:

(*) $A = B$ whenever for some possible language L, (i) all relevant truths can be expressed in L and (ii) $A \neq B$ cannot be truly expressed in L.

But as Dorr recognises (ibid), there *are* possible languages in which it is impossible to truly express the negation of any instance of *Booleanism*, and presumably if *Booleanism* is true, such languages are relevantly expressively adequate.[16] Yet there is no simple fallback to (*) which subsumes *Involution* and *Commutativity* and does rules out *Booleanism*. One might consider the result of modifying criterion (i) in (*) to something like 'it is independently plausible that all relevant truths can be expressed in L'. But adding such a complex generalisation stated partly in cognitive terms to one's theory of equality would be to take a step in the wrong direction.

necessity just in case having it entails having every form of necessity), that is tantamount to the principle that broadly necessarily equivalent propositions are equal. Notably, much of the recent literature takes the motivation for *Booleanism* to extend to a stronger principle of *Classicism*, according to which all predicates that are equivalent in classical quantificational logic express the same property. *Classicism* does entail the analogue of *Intensionalism* for $\Box_\top$, though not that $\Box_\top$ is equal to $\Box$ (metaphysical necessity). For discussion and applications of Booleanism and Classicism, and of their relation to intensionalism, see Bacon 2018, 2020; Bacon and Dorr 2024; Dorr et al 2021: §8; Fritz 2023; Goodsell 2021, 2024; Goodsell and Yli-Vakkuri manuscript; Roberts 2023.

[16] The possibility that Dorr mentions is of a language in which all formulae are automatically reduced to disjunctive normal form. In the case of strengthenings of intensionalism on which being metaphysically necessary equals holding at all possible worlds, one might entertain the idea of a language in which all formulae are automatically transformed into pictures of some region of modal space.

In short, while Dorr's mode of argument may have some dialectical force in serving to motivate *Involution* and *Commutativity* specifically, it does not suggest any way of unifying them. More basically, the provisional moral is that a piecemeal approach to the consequences of intensionalism looks uncomfortably (though far from decisively) *ad hoc*. Much more alarm bell-raising is what happens when an informative principle inconsistent with intensionalism is thrown into the mix.

4.2 Involution + Non-Vacuous Beta-Equality + Only Logical Circles

In part given his official aim of exploration, Dorr does not commit to intensionalism. That enables him to accept some principles inconsistent with intensionalism, but which might seem attractive considered in isolation. The question is: which hyperintensional conditions to accept?

Dorr himself initially considers the following hyperintensional condition:

No Circles $A \neq_{\sigma} B$, where A and B can be terms of any type σ such that B properly contains an occurrence of A in which no variable free in A is bound

No Circles indeed has striking consequences. For example, for any sentence A, it entails $A \neq (A \wedge A)$. Instantiating: the proposition that it is raining does not equal the proposition that it is raining and it is raining. In principle, Dorr is prepared to bite the bullet on such inequalities: 'I insist that these claims are *not* obviously false' (77). He makes a local appeal to the structural abductive virtue of strength, in effect explaining that natural pre-theoretical assessments must sometimes be sacrificed to structural abductive virtue. Surprisingly, he does not mention the abductive virtue of simplicity, which scientists often mention in the same breath as strength.

In any case, Dorr does not welcome all of the consequences of *No Circles*. The trouble is that, for any sentence A, *No Circles* also entails $A \neq \neg\neg A$. Moreover, *β-Equality* entails $\lambda X(X)A = A$, as well as $\lambda z_1 \dots z_n(A)(a_1, \dots, a_n) = A$ whenever $z_1 \dots z_n$ are not free in A, but the negations of such equations are instances of *No Circles*. In other words, *No Circles* is inconsistent with Dorr's intensional commitments. As he remarks for precisely that reason: 'we cannot accept *No Circles* as it stands' (73).

We defer critical discussion of *No Circles* itself to the following subsection. In brief, it is intended as a formalisation of the idea that true equations cannot 'run in a circle' (72). To anticipate, it will be argued that the judgements Dorr relies on to motivate it are unreliable. For now, however, we focus on issues related to how Dorr handles the inconsistency between *No Circles* and his intensional commitments (which were motivated quite independently of it). For what he does is add epicycles, by building exceptions into both of *No Circles* and *β-Equality*.

In more detail: instead of accepting *No Circles*, Dorr only accepts a restriction of it to instances in which any proper constituent C of B which does not overlap an occurrence of A expresses a 'logical' entity:

Only Logical Circles $A =_\tau B \rightarrow \text{Logical}_\sigma(C)$, where A, B and C can be terms of any 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 no variables free in A or C is bound in either of these occurrences

The new 'Logical' predicates are introduced by the comment:

> [L]et us help ourselves to a predicate Logical$_\tau$, of type $\langle\tau\rangle$, for every type τ. Logical$_\tau$(x_τ) should be true only if x_τ is the denotation [of] some closed term whose only constants are the logical constants ¬, ∧, [∨], $\forall_\tau$, $\exists_\tau$, and $=_\tau$. While this gloss may not constitute a satisfactory *definition* of 'Logical$_\tau$', it seems to convey an adequate grip on the intended interpretation (74).

Since by assumption, Logical($\lambda X(X)$) and Logical(¬), the effect of the restriction is to enable Dorr to accept equations of the form $A = \neg\neg A$ and $A = \lambda X(X)A$.

By itself, the restriction of *No Circles* to *Only Logical Circles* still does not enable Dorr to accept *β-Equality*, since on the intended interpretation of Logical, in typical equations of the form $\lambda z_1 \ldots z_n(A)(a_1, \ldots, a_n) =_{\langle\rangle} A$, at least some a_i will express a non-logical entity. Hence, instead of accepting *β-Equality*, Dorr also only accepts a restriction of it, to cases of non-vacuous λ-abstraction:

Non-Vacuous β-Equality $\lambda z_1 \ldots z_n(A)(a_1, \ldots, a_n) =_{\langle\rangle} A[a_1/z_1, \ldots, a_n/z_n]$ whenever $z_1 \ldots z_n$ all occur free in A

The result of doing so is to restore consistency (in classical higher-order logic), as Dorr proves in a technical appendix. For theorists invested in the success of his experiment in higher-order hyperintensionalism, that is the good news.

The bad news is that building such exceptions into *No Circles* and *β-Equality* is reminiscent of overfitting. It involves adding complications on the spot exactly in order to restore consistency. There is no background picture on offer which might help to explain how Dorr's intensional commitments ultimately fit together with anything like *No Circles*. Yet the absence of any such picture lends itself to the suspicion that the principles do *not* fit together, and that it is mainly by gerrymandering that outright inconsistency has been avoided.[17]

The rest of this subsection substantiates that suspicion by explaining some problems with the specific restrictions that Dorr posits to *No Circles* and *β-Equality*, as well as why his own attempts to rationalise them do not succeed.

First, it is worth asking what is wrong with handling the apparent counterexamples to *No Circles* simply by stipulating that $\neg$ and $\lambda X(X)$ specifically make for exceptions to *No Circles*, instead of restricting *No Circles* with the help of the Logical predicates. A natural answer is that, in the absence of some account of what is special about $\neg$ and $\lambda X(X)$, such a move is unacceptably *ad hoc*. Just specifying individual exceptions to some theoretical principle in the face of counterexamples is not OK. In any case, Dorr refrains from doing it. However, given that Dorr introduces the Logical predicates primarily by individually listing properties to which they should apply, it is not obvious how using them to restrict *No Circles* differs in spirit from the tactic of listing individual exceptions to *No Circles* itself.

Incidentally, one potential approach to defining the Logical predicates is in effect in terms of the signature of the language. But that approach looks artificially restricted, given that the standard Boolean connectives are not uniquely privileged. For example, consider the binary truth-functional connective $\downarrow$, where '$A \downarrow B$' may informally be read 'Neither A nor B'. Presumably, $\downarrow$ can be introduced to the language as a primitive, in which case it should count as logical. More generally, such an approach just pushes the question back a level. What is so special about the signature of the language?

Second, even if there is some interesting way in which logical entities are unique, what is ultimately wanted is an explanation of why they uniquely make for exceptions to *No Circles*.[18] Dorr is sensitive to that explanatory need. He suggests that '*[n]on-logical* entities are indissoluble, and always make for a genuine increase in complexity when they combine with something else' (74). But no independent content is given to this suggestion, beyond that which can already be inferred from *Only Logical Circles*, i.e. that logical entities uniquely make for exceptions to *No Circles*. Nor do pre-theoretical judgements of comparative 'complexity' discriminate along the envisaged lines. For instance, letting 'R' again abbreviate 'It is raining', intuitively, the propositions $R \wedge R$ and $\neg\neg R$ are each more complex than R, even though, again intuitively, $R \wedge R$ results

[17] An (extreme) analogy with the case of mathematical modelling: bringing one model given by the equation $x = y$ into 'agreement' with a rival model given by the equation $x = -y$ by restricting the domain of the former to $x < 0$ and of the latter to $x \geq 0$ and combining the results.

[18] For instance, perhaps being logical can be understood as being permutation-invariant (Tarski 1983). That answer to the question of what makes logical entities unique has no obvious relevance to the question of why they uniquely make for exceptions to *No Circles*.

in part from combining R with a non-logical entity (since R is non-logical), whereas $\neg\neg R$ does not. At best, Dorr's remark provides an informal gloss on the proposed restriction; it does not provide an independent rationale for it.

Incidentally, the specific way in which Dorr's theory implements the suggestion that, in contrast to non-logical entities, logical entities are soluble cannot be motivated by the familiar idea that logical terms do not add new 'subject matter' (or are 'topic-neutral'). For presumably, provided that logical constants do not add new subject matter, both of the sentences '$R \wedge R$' and '$\neg\neg R$' have the same subject matter as 'R', so if the fact that '$\neg\neg R$' has the same subject matter as 'R' is what enables it to express the same proposition, then '$R \wedge R$' should express the same proposition as well. However, on Dorr's theory, it does not, so his talk of not increasing 'complexity' cannot be reduced to not adding new subject matter. In effect, Dorr's theory is sensitive to finer-grained syntactic differences than a conception of higher-order identity which proposes to explain why some necessarily equivalent sentences fail to express the same proposition in terms of differences in their non-logical subject matter, though not uniformly sensitive to such fine-grained syntactic differences.[19]

Third, Dorr does not attempt to argue that vacuous λ-abstraction is independently ruled out. He does claim that positive arguments for *β-Equality* extend less readily to vacuous instances.[20] But even if so, that is of limited significance. As the simpler and stronger schema, full *β-Equality* is the default hypothesis. What is needed are positive arguments *against* vacuous instances of it, that do not readily extend to non-vacuous instances. In other words, what explains why vacuous λ-terms uniquely make for exceptions to *β-Equality*?

Dorr is sensitive to the costs of building an exception into *β-Equality* for the special case of vacuous λ-terms. For formal convenience, he suggests that theorists who do not accept full *β-Equality* should work in a language in which vacuous λ-terms are not well-formed. Doing so obviates the need to 'constantly [have] to make exceptions for

[19] For a critique of the idea that logic is in any interesting sense 'non-substantive', see Williamson 2014. In fact, Dorr himself explicitly repudiates the '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' (42). Of course, it may be that although logic is in no way *epistemologically* non-substantive, logical constants are in some way *metaphysically* non-substantial. However, such special treatment of logical constants receives no support from the background higher-order logic, in which just like other closed terms, logical constants are treated categorematically and express properties by *Comp*.

[20] In particular, Dorr argues for *β-Equality* from the behaviour of predicates defined by a stipulation of the form $F(x_1, \dots, x_n) \coloneqq A$, parenthetically adding about the form of argument that it 'seems a lot less compelling in the vacuous case where some of $x_1, \dots, x_n$ do not occur free in A, since we do not normally introduce new predicates by means of stipulations like that' (65; compare Goodman 2024). Of course, we do not 'normally' introduce new predicates by means of stipulations where A is a formula of more than 100,000 characters either, but that is presumably no objection to the corresponding instances of *β-Equality*. The point of prioritising simplicity and elegance in theory choice is to prevent factors like that from making too much of a difference.

the vacuous case' (57). However, that move only shifts the locus of complexity to the syntax of the formal language. By default, a disparate syntactic treatment of vacuous λ-terms and non-vacuous λ-terms is just as *ad hoc* as a disparate semantic treatment. For instance, why should the more complex predicate $\lambda z(\text{There is no God} \wedge z = z)$ be well-formed, but the simpler predicate $\lambda z(\text{There is no God})$ be ill-formed? Note that for some communicative purposes, the vacuous predicate seems exactly well-suited: for instance, if a philosopher of religion is explaining formally why the premise 'There is no God' perforce has some consequences for each of us, namely that if there is no God, each of is such that there is no God. Given that explanatory aim, the non-vacuous predicate $\lambda z(\text{There is no God} \wedge z = z)$ is gratuitously complex; the extra element '$z = z$' is a distraction from the logical point being made.[21]

Fourth, given the default expectation in favour of a uniform treatment of instances of *β-Equality*, it is anomalous that *Only Logical Circles* is consistent with non-vacuous instances of λ-abstraction but not (on the intended interpretation of Logical) with typical vacuous instances. Notably, the simpler *No Circles* is inconsistent both with some instances of *β-Equality* in which all of $z_1 \dots z_n$ occur free in A and with some in which none does. The fact that one can restore consistency with the former equations without thereby restoring consistency with the latter if one is allowed to make exceptions for certain (allegedly) special cases evinces the unnaturalness of a disparate treatment. It does not lend such a treatment independent support.

Fifth, given the default expectation in favour of a uniform treatment of instances of *Booleanism*, it is also anomalous that *Only Logical Circles* is consistent with *Involution* but not (on the intended interpretation of Logical) with many other instances of *Booleanism*, such as *Idempotence*$_\wedge$:

Idempotence$_\wedge$ $$\lambda z_1 \dots z_n(A) =_\sigma \lambda z_1 \dots z_n(A \wedge A)$$

Pre-theoretically, *Idempotence*$_\wedge$ seems at least as plausible as *Involution*. If 'it is raining' means the same thing as 'it is not not raining', why should it mean something different from 'it is raining and it is raining'? Notably, the simpler condition *No Circles* is again inconsistent with some instances of *Involution* and with some instances of *Idempotence*$_\wedge$. The fact that one can restore consistency with the former equations without thereby restoring consistency with the latter if one is allowed to make exceptions for certain (allegedly) special cases evinces the unnaturalness of a disparate treatment. It does not lend such a treatment independent support.

Here is another example in a similar spirit. Suppose that we do adopt ↓ as primitive. Then *Involution* is inconsistent with *Only Logical Circles* and equations of the form

[21] It is also worth noticing that a ban on vacuous λ-terms invalidates natural inference patterns such as that from $\lambda z_1 \dots z_n(A \wedge B)(a_1, \dots, a_n)$ to $A[a_i/z_i] \wedge \lambda z_1 \dots z_n(B)(a_1, \dots, a_n)$. For then, when $z_1 \dots z_n$ all occur free in A, and thus in $A \wedge B$, but do not all occur free in B, $\lambda z_1 \dots z_n(A \wedge B)(a_1, \dots, a_n)$ may be true even though the predicate $\lambda z_1 \dots z_n(B)$ is ill-formed. Thanks here to Tim Williamson.

$\neg A = A \downarrow A$ whenever $\neg$Logical(A). For if $\neg A = A \downarrow A$, then $\neg\neg A = \neg(A \downarrow A)$. But $\neg(A \downarrow A) \neq A$ by *Only Logical Circles* whenever $\neg$Logical(A). Pre-theoretically, however, equations of the form $\neg A = A \downarrow A$ are as plausible as those of the form $A = \neg\neg A$. If 'it is raining' means the same thing as 'it is not not raining', why should 'it is not raining' mean something different from 'it is neither raining nor raining'?

Finally, the combination of *Non-Vacuous β-Equality* with the falsity of $A = A \wedge A$ whenever $\neg$Logical(A) puts further pressure on the proposed rationale for the restriction of *No Circles* to *Only Logical Circles*. For then, for any sentence A, the equation $\lambda X(X)A = A$ is true by *Non-Vacuous β-Equality*, but the equation $\lambda X(X \wedge X)A = A$ is false whenever $\neg$Logical(A), since $\lambda X(X \wedge X)A = A \wedge A$, again by *Non-Vacuous β-Equality*. Such a disparate treatment of equations of the forms $\lambda X(X)A = A$ and $\lambda X(X \wedge X)A = A$ looks anomalous by Dorr's own lights. For if equations of the former form are true because $\lambda X(X)$ 'dissolves' in $\lambda X(X)A$, why aren't equations of the latter form similarly true because $\lambda X(X \wedge X)$ dissolves in $\lambda X(X \wedge X)A$? Of course, Dorr need not think that logical entities always dissolve: presumably, $\neg$ does not dissolve in $\neg A$. However, the burden is on him to explain the relevant difference between $\lambda X(X)$ and $\lambda X(X \wedge X)$, in a way which makes explicit contact with the idea that logical entities are soluble.

To sum up: this subsection has explored tensions resulting from Dorr's attempt to mix and match intensional and hyperintensional principles. Might a committed hyperintensionalist be better off just plumping for *No Circles*? Perhaps—but not by much. For arguably, the main cause of the complications identified is that the judgements which motivate *No Circles* are unreliable.

4.3 No Circles

The argument of the preceding subsection for invidious complexity in Dorr's theory of equality did not depend on invidious complexity in *No Circles* itself. Formally, however, *No Circles* is overtly a restricted generalisation. It is one of many principles one might consider with something like the form '$A \neq B$ when the terms "A" and "B" differ in such-and-so respect'. More informally, when first presented with *No Circles*, a theorist familiar with the usual options might guess it to be a partial articulation of the vision behind the structured theory of propositions, given that it similarly recycles differences in the constituent structure of predicates as differences between the things that those predicates express.

However, Dorr's aim is to carve out an alternative not just to intensionalism but to the structured theory of propositions. Thus his primary argument for *No Circles* does not take anything like the latter for granted. It is simply from assessment of the following set of equations, which we follow him in presenting informally (70):

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*.

In commenting on these equations, Dorr writes:

> *Grue* and *Bleen* are uncontroversial: just look at the passages of 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 (70–1).
>
> 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 together, since that would be *circular* (71–2).

No Circles is in turn motivated as one way of vindicating the reported judgement—that is, consistently with the basic logic of equality. For as Dorr remarks:

> 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' (73).

Notably, however, Dorr simply proceeds to suggest 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' (ibid), without independently checking that the relevant judgements do *not* over-generalise in some disastrous way.

No doubt much could be said about the nature and limitations of the judgements that Dorr reports. Indeed, Dorr draws plausible connections between them and a cluster of cognitive-explanatory considerations. The dangers of reliance on such considerations in motivating hyperintensionalism is a theme of Williamson 2024; properly elaborating Williamson's arguments here would be too complicated. Hence, we stick to three limited observations.

First, to complete the discussion of the preceding subsection, note that judgements of illicit circularity generalise to combinations of equations formulated in logical terms but otherwise analogous to the combination of *Grue*, *Bleen*, *Green* and *Blue*. For example, suppose that one explains the unfamiliar logical constant ↓ to a student directly in terms of the equation $\downarrow = \lambda XY(\neg(X \vee Y))$. If one then proceeded to explain the familiar logical constants $\neg$ and $\vee$ by the equations $\neg = \lambda X(X \downarrow X)$ and $\vee = \lambda XY((X \downarrow Y) \downarrow (X \downarrow Y))$, it would be quite natural for one's student to complain of illicit circularity and feel a sense of explanatory loss.[22] One can easily construct similar examples using logical operations that are even more difficult to understand on their own terms, independently of the standard Boolean connectives. Such examples reinforce the invidiousness of the restriction of *No Circles* to *Only Logical Circles*.

Second, judgements of illicit circularity are sensitive to more specific cognitive factors than *No Circles* reflects. For one thing, it is not the natural first reaction to an equation such as 'For it to be raining is for it to be raining and raining', presented simply as a candidate for assessment as true or false, to reject it as circular. But for a closer comparison, consider the following three equations, the joint truth of any two of which would constitute a counterexample to *No Circles*:

*Red** To be red is to be a primary colour and not yellow or blue.

*Yellow** To be yellow is to be a primary colour and not blue or red.

*Blue** To be blue is to be a primary colour and not red or yellow.

Considered individually, each of *Red**, *Yellow**, and *Blue** is plausible, even though presumably they stand or fall together. Even considered collectively, their combination need not appear problematic, since none need be conceived as providing an explanation of the meaning of the relevant term. Because 'red', 'yellow' and 'blue' are typically understood to some extent independently of each other, by independently grasped prototypes, *Red**, *Yellow**, and *Blue** are in principle able to constitute a virtuous circle, rather than a vicious one.

By contrast, 'grue' and 'bleen' are typically understood exclusively by their definitions in terms of 'green' and 'blue'. That fact may help explain the viciousness of the combination of *Grue*, *Bleen*, *Green* and *Blue*, since the prospect that 'green' and 'blue' are themselves to be understood in terms of 'grue' and 'bleen' might be felt to endanger understanding of all four terms. After all, if someone proposes to explain something that you thought you understood well in terms of something that you fail to understand, you may start to second-guess how well you understand it. For example, someone who has never before encountered limits may find it counter-productive to be told that the really important thing to understand about the number e is that it equals $\lim_{n\to\infty}(1+\frac{1}{n})^n$. Nevertheless, the equation $e = \lim_{n\to\infty}(1+\frac{1}{n})^n$ in fact constitutes an important mathematical truth.

[22] Thanks here to Alex Roberts.

Third, judgements of illicit circularity also arise more widely than *No Circles* reflects. For one thing, arguments are often judged to be illicitly circular ('question-begging'). It is quite unclear how *No Circles* could help to explain such judgements, not least given their well-known sensitivity to highly specific pragmatic factors. But again, for a closer comparison, consider the following conversational exchange:

A. What is it to be a vixen?
B. To be a vixen is to be a female fox.
A. What is to be a female fox?
B. To be a female fox is to be a vixen.

A natural response by A in the context would be to complain of circularity:

A. If to be a vixen is to be a female fox, then it can't *also* be that to be a female fox is to be a vixen. That would be *circular*!

However, on the intended interpretation of 'to be'-statements, such a response from A is inconsistent with the basic logic of equality. 'To be a vixen is to be a female fox' entails 'to be a female fox is to be a vixen' by the symmetry of equality. Thus judgements of illicit circularity *do* threaten badly to overgeneralise.

Of course, one could put a pragmatic gloss on A's apparent denial of the joint truth of 'To be a vixen is to be a female fox' and 'To be a female fox is to be a vixen': perhaps A is really trying to object to B's joint assertion in the context, rather than to their joint truth. However, it is not clear why such an interpretation is to be preferred to one which takes A's assertion at face value.[23] In any case, it looks *ad hoc* to treat A's judgement as having a different source from the judgement that *Green*, *Blue*, *Grue* and *Bleen* cannot all be true 'since that would be circular'. Of course, proponents of *No Circles* could insist that the judgements do have different sources. If I accuse you of wrongly treating similar cases differently, it is always open to you to deny that the cases are similar after all. Probably, with some spare time, you will even be able to devise a complex story which rationalises your disparate treatment of them. Nevertheless, experiments with complicating the logic of equality to fit some judgements of illicit circularity should be reserved until after it has been made independently plausible that the judgements that are to be accommodated are more trustworthy than closely related judgements of illicit circularity that are known to be false.

[23] Notably, Dorr spends several pages motivating the principle he calls *Symmetry*, i.e. 'If to be F is to be G, then to be G is to be F'. As he writes: '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' (43). The resistance that Dorr reports is defeasible evidence for taking A's rejection of *Symmetry* at face value.

5. Conclusion

The Russell-Myhill paradox vividly illustrates the potential dangers of hyperintensional experiments with the logic of equality. In following the impulses to depart from intensionalism where they lead, outright inconsistency sometimes results.

This paper has sketched some less dramatic but still serious ways in which even exceptionally rigorous hyperintensional theorising may go wrong. Dorr takes great care to make sure that his theory of equality is consistent. However, evidence has emerged that its consistency is a result of gerrymandering. More generally, the various component parts of Dorr's theory look to have been put together *ad hoc*, in a way inadequately guided by any general picture of the nature of properties. The chief danger with that is of falling victim to some bias or other on reasoning about equality. And indeed, the judgements on which Dorr relies to motivate the principle most distinctive of his theory look unreliable.

One way to think of the issue is in terms of what constraints are in play on theorising about equality. In some contemporary metaphysics, there is increasing convergence on certain *formal* constraints. In particular, much recent work proceeds in a shared framework of classical higher-order logic (sometimes a bit more, sometimes a bit less). That framework conduces to the comparison of rival theories about phenomena of metaphysical interest, such as the nature of properties. When developed in such a setting, some popular ideas about what properties are end up looking much worse: the structured theory of propositions is a case in point. Nevertheless, there is a growing sense that if a theory of properties looks bad in a standard higher-order setting, that is (probably, defeasibly, ...) its problem.

Such points extend to standard higher-order *languages*. What language to use forms part of what is up for abductive comparison. Since *ad hoc* restrictions to a perspicuous and general-purpose language like the lambda calculus incur costs in simplicity and fruitfulness, the prospects are limited for finessing abductive costs elsewhere by banning readily intelligible and generally useful terms. Whatever the original problem was is too likely to resurface as a problem with the ban. Moreover, in light of the replication crisis for disciplines like psychology, an increasingly emphasised danger in science—closely related to overfitting—is of researchers helping themselves to too many degrees of freedom in designing experiments and interpreting their results (Wicherts et al 2016). Philosophers granting themselves too much flexibility in which language to articulate their theories may be doing something similar.

At any rate, following recent work by Williamson, this paper has used the vocabulary of 'overfitting' to articulate a *methodological* constraint on theorising about equality: roughly, that it not enact *ad hoc* complications to fit shaky data. Of course, the extent to which hyperintensional theorising does violate a norm of that kind is controversial, as is the exact force that the norm has in metaphysics. This paper has not even tried to settle either issue. Rather, it has pursued the more modest aim of modelling how the proposed methodological constraint impacts on theorising about equality. Hopefully, its conclusions will resonate independently.[24]

[24] Thanks to Michael Bevan, Cian Dorr, Dominik Ehrenfehls and Kin San for helpful conversation and comments on drafts of this paper, and to Ofra Magidor, Alex Roberts and especially Tim Williamson for helpful comments on multiple drafts.

Bibliography

Bacon, Andrew. 2018. ‘The broadest necessity’. *Journal of Philosophical Logic* 47: 733–83.

Bacon, Andrew. 2020. ‘Logical combinatorialism’. *Philosophical Review*, 129: 537–589.

Bacon, Andrew. 2023. ‘A theory of structured propositions’. *Philosophical Review*.

Bacon, Andrew and Dorr, Cian. 2024. ‘Classicism’. In Fritz, Peter and Jones, Nicholas (eds.). 2024. *Higher-Order Metaphysics*. Oxford: Oxford University Press.

Ditter, Andreas. Manuscript-a. ‘Higher-order essences: logic and semantics’.

Ditter, Andreas. Manuscript-b. ‘The hyperintensionality of essence’.

Dorr, Cian. 2016. ‘To be F is to be G’. *Philosophical Perspectives*, 30: 39–134.

Dorr, Cian, Hawthorne, John, and Juhani Yli–Vakkuri. 2021. *The Bounds of Possibility: Puzzles of Modal Variation*. Oxford: Oxford University Press.

Fritz, Peter. 2017. ‘How fine-grained is reality?’. *Filosofisk Supplement*, 13: 52–7.

Fritz, Peter. 2020. ‘On higher-order logical grounds’. *Analysis*, 80: 656–666.

Fritz, Peter. 2021. ‘Structure by proxy, with an application to grounding’. *Synthese*, 198: 6045–6063.

Fritz, Peter. 2022. ‘Ground and grain’. *Philosophy and Phenomenological Research*, 105:299–330.

Fritz, Peter. 2023. ‘Operands and instances’. *The Review of Symbolic Logic*, 16: 188—209.

Fritz, Peter. 2023. *The Foundations of Modality*. Oxford: Oxford University Press.

Fritz, Peter and Jones, Nicholas (eds.). 2024. *Higher-Order Metaphysics*. Oxford: Oxford University Press.

Fritz, Peter, Lederman, Harvey, and Gabriel Uzquiano. 2021. ‘Closed structure’. *Journal of Philosophical Logic*, 50: 1249–1291.

Goodman, Jeremy. 2017. ‘Reality is not structured’. *Analysis*, 77: 43–53.

Goodman, Jeremy. 2024. ‘Higher-order logic as metaphysics’. In Fritz, Peter and Jones, Nicholas (eds.). 2024. *Higher-Order Metaphysics*. Oxford: Oxford University Press.

Goodman, Jeremy. Manuscript. ‘A theory of aboutness’.

Goodman, Nelson. 1954. *Fact, Fiction and Forecast*. Cambridge, MA: Harvard University Press.

Goodsell, Zachary. 2021. ‘Arithmetic is determinate’. *Journal of Philosophical Logic*, 51: 127–150.

Goodsell, Zachary. 2024. ‘Arithmetic is necessary’. *Journal of Philosophical Logic*, 53(4).

Goodsell, Zachary and Yli-Vakkuri, Juhani. Manuscript. *Logical Foundations*.

Hodes, Harold. 2015. 'Why ramify?'. *Notre Dame Journal of Formal Logic*, 56: 379–415.

Ramsey, Frank. 1927. 'Facts and propositions'. *Proceedings of the Aristotelian Society*, 7: 153–70.

Roberts, Alexander. 2023. 'Is identity non-contingent?'. *Philosophy and Phenomenological Research*, 106: 3–34.

Bertrand Russell. 1903. *The Principles of Mathematics*. Cambridge: Cambridge University Press.

Skiba, Lukas. 2021. 'Higher-order metaphysics'. *Philosophy Compass*, 16: 1–11.

Uzquiano, Gabriel. 2015. 'A neglected resolution of Russell's paradox of propositions'. *Review of Symbolic Logic*, 8: 328–344.

Wicherts JM, Veldkamp CLS, Augusteijn HEM, Bakker M, van Aert RCM and van Assen MALM .2016. 'Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid *p*-hacking'. *Front. Psychol.* 7:1832. doi: 10.3389/fpsyg.2016.01832

Williamson, Timothy. 2013. *Modal Logic as Metaphysics*. Oxford: Oxford University Press.

Williamson, Timothy. 2014. 'Logic, metalogic and neutrality'. *Erkenntnis*, 79: 211–231.

Williamson, Timothy. 2017. 'Semantic paradoxes and abductive methodology'. In Bradley Armour-Garb (ed.), *Reflections on the Liar*. Oxford: Oxford University Press.

Williamson, Timothy. 2021. 'Degrees of freedom: is good philosophy bad science?' *Disputatio*, 13: 73–94.

Williamson, Timothy. 2024. *Overfitting and Heuristics in Philosophy*. Oxford: Oxford University Press.