Notre Dame Journal of Formal Logic
Volume 58, Number 4, 2017

Prospects for a Naive Theory of Classes

Hartry Field, Harvey Lederman, and Tore Fjetland Øgaard

Abstract The naive theory of properties states that for every condition there is a property instantiated by exactly the things which satisfy that condition. The naive theory of properties is inconsistent in classical logic, but there are many ways to obtain consistent naive theories of properties in nonclassical logics. The naive theory of classes adds to the naive theory of properties an extensionality rule or axiom, which states roughly that if two classes have exactly the same members, they are identical. In this paper we examine the prospects for obtaining a satisfactory naive theory of classes. We start from a result by Ross Brady, which demonstrates the consistency of something resembling a naive theory of classes. We generalize Brady's result somewhat and extend it to a recent system developed by Andrew Bacon. All of the theories we prove consistent contain an extensionality rule or axiom. But we argue that given the background logics, the relevant extensionality principles are too weak. For example, in some of these theories, there are universal classes which are not declared coextensive. We elucidate some very modest demands on extensionality, designed to rule out this kind of pathology. But we close by proving that even these modest demands cannot be jointly satisfied. In light of this new impossibility result, the prospects for a naive theory of classes are bleak.

1 Introduction

Let L be any first-order language, and let L^+ be the result of adding to it a 1-place predicate "Class," a 2-place predicate $\in$, an abstraction operator $\{ : \}$, and a new conditional operator $\Rightarrow$. For any formula A, $\{x : A\}$ is a term whose free variables are those that are free in A except for x. The *naive theory of classes (over L)* consists of the following principles.

Received April 13, 2014; accepted October 7, 2014
First published online June 6, 2017
2010 Mathematics Subject Classification: Primary 3E70
Keywords: naive class theory, naive comprehension, extensionality axiom, nonclassical logic
 10.1215/00294527-2017-0010

Abstraction Schema:

$$\forall u_1, \ldots, \forall u_n \forall z \big[z \in \{x : A(x; u_1, \ldots, u_n)\} \Leftrightarrow A(z; u_1, \ldots, u_n) \big]. \quad (1)$$

Class Abstracts:

$$\forall u_1, \ldots, \forall u_n \mathrm{Class}\big(\{x : A\big(x, u_1, \ldots, u_n\}\big)\big). \quad (2)$$

Extensionality Rule:

$$\mathrm{Class}(a) \wedge \mathrm{Class}(b) \wedge \forall u (u \in a \Leftrightarrow u \in b) \vDash \forall z (a \in z \Leftrightarrow b \in z). \quad (3)$$

(The schematic variable A in (1) and (2) is assumed to have as substituends all formulas in the language L^+.) The theory also takes universal instantiation and existential generalization to apply even to abstraction terms: informally speaking, it assumes that these are denoting terms. Given the abstraction schema, the conclusion of the extensionality rule is equivalent to the claim that a and b are intersubstitutable in all contexts. The extensionality rule thus guarantees that we can define identity between classes as coextensiveness without giving up on the substitutivity rule for identity. (In fact, there may be reasons to want a stronger form of extensionality; we will discuss that in due course.)

As is well known, the naive theory of classes is inconsistent in classical logic (when $A \Rightarrow B$ is defined as $\neg A \vee B$). Indeed, Russell's paradox shows the classical inconsistency of the abstraction schema alone, *even without extensionality*.

But there are in the literature a variety of nonclassical logics in which the abstraction schema is demonstrably consistent: in some cases it is negation-consistent (that is, it never entails both a sentence and its negation), but in all cases it is at least Post-consistent (that is, it does not entail everything). These theories, with abstraction but not extensionality, are not theories of "classes," since they do not necessarily allow us to identify all "classes" which have exactly the same members. They are better called naive theories of *properties*, and the abstraction operator and the $\in$ used in the theories should be interpreted as a property-forming operator and the relation of instantiation, respectively.[2] The project of finding a logic for this naive theory of properties—with abstraction but not extensionality—turns out to be essentially the same as the project of finding a logic compatible with the naive theory of truth and satisfaction. The nonclassical logics in the recent literature on naive truth and satisfaction thus provide a number of options for strong logics of naive properties.

But what if we want a naive theory of *classes*, with extensionality as well as the abstraction schema? Brady [5]–[8] has presented several closely related naive theories of classes. The theories are very similar, but there are slight differences from one to another in the conditional that they employ. Brady's technique for proving extensionality to be consistent is quite different from anything used in the standard constructions to show that abstraction is consistent; we can no longer simply carry over proofs from the literature on naive truth and satisfaction.

So the question arises: To what extent can Brady's result be generalized to other logics that are known to be adequate for naive properties (or, for naive truth and satisfaction)? A difficulty in answering this question is that Brady's presentation of his proof is rather opaque. Our first goal in this article is to present his result in a simpler, more accessible way; Sections 2 through 6 will be devoted to this task. This new presentation of Brady may or may not count as a "new proof" of his theorem, but it at least makes it easier to see how Brady's result can be generalized to apply to some other logics in the same vicinity as the ones Brady has considered. Section 5 already

contains two slight generalizations of Brady, demonstrating both how Brady's theorem applies to a logic which includes a noncontraposable conditional, and how it generalizes to "dynamic" variants of his construction. But the most important generalization comes in Section 9, when we show that Brady's result carries over neatly to certain logics which have a modal-like semantics that uses 4-valued, as opposed to 3-valued, worlds. (These will include some logics from Bacon [1], which turn out to be much closer to Brady's than they initially appear to be, and yield a naive theory of classes by essentially Brady's argument.) This last generalization has some significant advantages over Brady's original constructions. Most notably, the resulting logic includes a weakening rule (indeed, a weakening axiom) for a noncontraposable conditional $\rightarrow$ from which the contraposable conditional $\Rightarrow$ (whose biconditional is used in the laws above) is defined.

But we argue that even these improved Brady-like logics are too weak for reasoning about classes. Worse yet, we conclude with an impossibility result which shows fairly decisively that one cannot hope to do significantly better. In the presence of abstraction, extensionality introduces new strength: whereas strong conditional logics can be shown to be consistent with naive theories of truth and properties, the same cannot be said for the naive theory of classes.[3]

2 The Goal

Let L be any first-order language, with primitive logical operators $\neg$, $\wedge$, and $\forall$;[4] for simplicity, we will assume that its only singular terms are variables. We will also assume that it contains an identity predicate, and for convenience, that it contains a primitive sentence $\bot$ to be understood as logically false. We use $\vee$ and $\exists$ as metalinguistic abbreviations with the usual definitions, and define $\top$ as $\neg\bot$. (Primitive sentences can be viewed as 0-place predicates, so this will require no addition to the formation rules below.) Let M be any classical normal model for L in which $\bot$ is false, where "normal" means that the extension in M of the identity predicate is $\{\langle o, o\rangle : o \text{ is in the domain of } M\}$.

Let L^+ be the result of adding a new 1-place predicate "Class," a new 2-place predicate $\in$, a term-forming operator $\{:\}$, a binary operator $\rightarrow$, and for convenience, a set N of primitive names with the same cardinality as the domain D_M of M. The goal will be to extend the classical M to a nonclassical model for L^+, nonclassical in that (though the model is constructed using classical set theory) it validates only the principles of a sublogic of classical logic. It is essential that the sublogics have a primitive conditional: the conditional defined from $\neg$ and $\vee$ in these logics will fail to satisfy either reflexivity ($A \rightarrow A$) or modus ponens; because of this, abstraction and extensionality stated with the defined conditional will either fail or be too weak to be of interest. The challenge is to see whether we can give the new conditional a reasonably strong logic, but one which is still consistent with both abstraction and extensionality.

To meet this challenge, we will consider a family of different methods of constructing, for each classical model M of the old L, a new nonclassical model for L^+; at least until Section 8, the nonclassical models (and resulting logics) will differ only in the treatment of the conditionals (and hence biconditionals) that they employ. (Our focus will be on nonclassical sublogics that have all the classical structural rules. Most will restrict excluded middle while retaining disjunctive syllogism, but

we will later consider some that restrict disjunctive syllogism. They will otherwise be standard for the connectives $\neg$, $\wedge$, and $\vee$, and for the quantifiers: in particular, double negation elimination and all the De Morgan laws hold.) But in every case, the new nonclassical model will have the original M effectively as a submodel; M will model the urelements over which the classes are built.[5] (The construction actually works also from the not quite classical model that has empty domain; this yields pure class theory, i.e., with no urelements.) So the idea is to show that whatever the classical reality, there is a nonclassical extension of it with naive classes. (Many of the classes in the models we construct will behave classically; the nonclassicality arises only for "pathological" classes such as the class of all non-self-membered classes.)

More formally, the formation rules for L^+ are what one would expect:

The 0-*terms* are the variables and the primitive names.

For each n, the *atomic n-formulas* are the result of (for some k) applying a k-place predicate (whether in the ground language or $\in$ or "Class") to k n-terms.

For each n, we build up *n-formulas* from atomic n-formulas using $\neg$, $\wedge$, $\rightarrow$, and $\forall$ (together with a variable) in the usual way.

The *$(n+1)$-terms* are the variables and the primitive names together with anything of form $\{x : A\}$, where x is a variable and A is an n-formula. (So if $m > n$, all n-terms are m-terms and all n-formulas are m-formulas.)

A term is anything that is an n-term for some n, and similarly for formulas. (Intuitively, an n-formula is a formula in which the largest number of nestings of abstracts is at most n.)

Free occurrence of a variable in a term or formula is defined inductively in the obvious way. The free occurrences of variables of the term $\{x : A(x)\}$ are the free occurrences of variables other than x in A. A term or formula with no free occurrences of variables is called *closed*. An *abstract* is a term (not necessarily closed) that is not a primitive name or variable.

We will construct our models in two stages. First, in Sections 3–5 we will construct a preliminary model M^+ whose domain is $D_M \cup \{\text{closed abstracts of } L^+\}$. (Each closed abstract can be regarded as naming itself, and each member of D_M will be regarded as named by a member of N; given this, we can treat quantification as substitutional.) Later, in Section 6 we will move to a "contracted" model $M^+/\approx$ whose domain is $D_M \cup \{\text{equivalence classes of closed abstracts of } L^+\}$, under a suitable equivalence relation $\approx$. It is only in the contracted models that it is appropriate to think of what we are adding to D_M as classes; until we do the contraction, we should think of the new models as containing the things in the old model together with class-representatives, where each class will have many representatives. (So the predicate "Class," which applies to the new objects, is a bit misleading in the case of the uncontracted model.) To repeat, there will be a family of different constructions of models M^+, leading to different logics for $\rightarrow$; and for each M^+, there will be a corresponding $M^+/\approx$.

The model M^+ will be such that the predicates of the original language L are classical and have the same extension as in the original model. (So M^+ is in a sense a nonclassical extension of M.) In particular, in M^+ the identity predicate of L is only a predicate of identity-restricted-to-D_M; so to avoid confusion, we will write

the identity predicate of L as $=_L$. We will eventually want to define identity in L^+ by

$$x = y \text{ iff } (x =_L y) \vee \big[\text{Class}(x) \wedge \text{Class}(y) \wedge \forall u(u \in x \Leftrightarrow u \in y)\big],$$

where $\Leftrightarrow$ is as defined below. This obviously coincides with $=_L$ in the domain of M; later on in Section 6 we will show that it satisfies the principles of identity appropriate to the nonclassical logic. But until we contract M^+, this will not behave as an identity predicate in our model, so to avoid the danger of confusion we prefer to avoid talk of class-identity until we get to $M^+/\approx$.

The goal is that whatever the starting classical model, the new model will "validate" (in a sense to be explained) the abstraction schema and the extensionality rule. Once we have defined identity as above, the latter will give us

$$a = b \vDash \forall z(a \in z \Leftrightarrow b \in z),$$

which, together with abstraction, implies the substitutivity rule for identity.

The conditional $\Rightarrow$ is to be understood as a metalinguistic abbreviation: $A \Rightarrow B$ is short for $(A \rightarrow B) \wedge (\neg B \rightarrow \neg A)$ (and $A \Leftrightarrow B$ for $(A \Rightarrow B) \wedge (B \Rightarrow A)$), so that (given the redundancy of double negation and minimal laws for conjunction) it is guaranteed to be contraposable; that is, $A \Rightarrow B$ is equivalent to $\neg B \Rightarrow \neg A$. We could have assumed that the primitive $\rightarrow$ is itself contraposable, in which case $\Rightarrow$ would coincide with it, and we could state abstraction and extensionality by using the corresponding biconditional $\leftrightarrow$. But it is more general not to make that assumption and to use $\Rightarrow$ and $\Leftrightarrow$ defined as above in formulating abstraction and extensionality. (Brady himself uses only a contraposable conditional in his constructions, so his $\rightarrow$ is our $\Rightarrow$. In this respect our constructions will be more general than Brady's, but they are so closely modeled after his that their conditionals are reasonably called "noncontraposable Brady conditionals.")

What about a conditional form of extensionality? As we will see, one of Brady's constructions delivers (in our $\Rightarrow$ notation)

$$\vDash \forall x \forall y\big[\text{Class}(x) \wedge \text{Class}(y) \wedge \forall u(u \in x \Leftrightarrow u \in y) \Rightarrow \forall z(x \in z \Leftrightarrow y \in z)\big],$$

but what are arguably the more satisfactory ones do not. However, all of them yield

$$\vDash \forall x \forall y\big[\text{Class}(x) \wedge \text{Class}(y) \wedge \forall u(u \in x \Leftrightarrow u \in y) \rightarrow \forall z(x \in z \Leftrightarrow y \in z)\big].$$

This last axiom is one benefit of taking the noncontraposable $\rightarrow$ as primitive: it gives us an axiom form of the substitutivity of identity and not just the weaker rule form. (We will discuss this further in Section 6.) Still, we suspect that getting this stronger axiom form depends on particular features of the models we will consider, and we would be willing to settle for the rule form, if doing so allowed better laws for the conditionals.[6]

Another, more substantial reason for taking a noncontraposable conditional as primitive will emerge in Sections 8 and 9.

3 Static and Dynamic Microconstructions

Brady uses a 3-valued modal-like semantics for his proofs that extensionality is consistent with abstraction. We will work in that framework until Section 8, when we will start to move toward a 4-valued generalization. We call the values 0, 1/2, and 1. The cardinality of the set W_M of worlds depends on the cardinality $|M|$ of the ground model M, but this will not matter for our purposes. One world $@_M$ of W_M is designated, in that an inference is *M-valid* if and only if: if the premises have value 1 at

$@_M$, so does the conclusion. An inference is *valid* if and only if it is M-valid for all ground models M.

The value of a conditional at a given world is determined, at least in part, by the values of its antecedent and consequent at other worlds. In particular, for each world w, facts about the other worlds determine a function v_w (a *prevaluation*) that maps each conditional sentence $A \rightarrow B$ into one of the values in $\{0, 1/2, 1\}$.[7] In this section and the next, we will leave the prevaluation v_w associated with world w a black box.

In this section the task is to show how, given the prevaluation v_w at a world w and the underlying ground model M, we determine the values of every sentence of L^+ at w, in particular, the values of sentences containing $\in$. To this end, we adapt the inductive procedure of Gilmore [16], [17] and Kripke [23]. The valuation procedure (which we will call the *microconstruction*) works by first assigning a value $|A|_{w,\sigma}$ in $\{0, 1/2, 1\}$ to each sentence A at each ordinal level σ; we then show that as σ gets bigger we eventually reach an ordinal Ψ_w after which the values of sentences do not change, and those "fixed-point values" $|A|_{w,\Psi_w}$ will be regarded as the values $|A|_w$ (without an ordinal subscript) of the sentences at w.

In Section 4 we show that this procedure gives a value-theoretic result that we call *microextensionality*. After presenting Brady's method of constructing a "good" prevaluation in Section 5 (we will call this the *macroconstruction*), we will show how the microextensionality theorem can be used to establish the consistency of extensionality with abstraction in Brady's conditional logics.

Since in the present section and the next, our focus is on an arbitrary world whose prevaluation is given, we will use v (as opposed to the v_w of our official notation) for the prevaluation at the world in question. Since at each world the value of every sentence is determined by the associated prevaluation together with the ground model, we will sometimes write $|A|_{v,M}$, or even $|A|_v$ (since we will be holding the classical "ground model" M fixed), instead of $|A|_{w,M}$. Thus, we sometimes also speak of the microconstruction *over* v.

Until Section 8, we will stick to microconstructions that generate the strong Kleene logic for the connectives $\neg$, $\wedge$, and $\forall$ (and $\vee$ and $\exists$, which are defined from the others in the usual way) and that use only *minimal* fixed points of the microconstruction. Even with this restriction, there is more than one possible way to adapt Gilmore and Kripke. The simplest—which Brady uses—we call the *static microconstruction*. "Static" here means that the values of conditionals do not change during the construction: they are simply the values given by the prevaluation v. (Within each static microconstruction, the conditionals behave essentially like atomic formulas.) This static construction goes as follows.

For any ordinal σ:

(1) if p is an atomic k-place predicate of the ground language and $t_1, \ldots, t_k$ are closed terms, then $|p(t_1, \ldots, t_k)|_{v,\sigma}$ is 1 if $\langle t_1, \ldots, t_k \rangle$ is in the extension of p in the ground model M, and 0 otherwise;
(2) $|\text{Class}(t_1)|_{v,\sigma}$ is 1 if t_1 is a closed abstract, 0 if it is a name in N;
(3) if t_1 and t_2 are closed terms and t_2 is not a class abstract, then $|t_1 \in t_2|_{v,\sigma}$ is 0;
(4) $|A \rightarrow B|_{v,\sigma}$ is just $v(A \rightarrow B)$;
(5) $|\neg B|_{v,\sigma}$ is $1 - |B|_{v,\sigma}$;
(6) $|A \wedge B|_{v,\sigma}$ is $\min\{|A|_{v,\sigma}, |B|_{v,\sigma}\}$;

(7) $|\forall x A|_{v,\sigma}$ is $\min\{|A(t/x)|_{v,\sigma} : t$ is a closed term of $L^+\}$;
(8) if t_1 is a closed term and t_2 is $\{x : A(x)\}$, then $|t_1 \in t_2|_{v,\sigma}$ is
  (a) 1 iff $(\exists\rho < \sigma)(\forall\tau$ in the interval $[\rho, \sigma))(|A(t/x)|_{v,\tau} = 1)$,
  (b) 0 iff $(\exists\rho < \sigma)(\forall\tau$ in the interval $[\rho, \sigma))(|A(t/x)|_{v,\tau} = 0)$,
  (c) 1/2 otherwise.

This inductive definition yields a value for every sentence at every σ, with the main induction on σ and a subinduction on the complexity of the sentence.

The important feature of this procedure is that it is "monotonic in the information order": if a sentence gets value 1 or 0 at any σ, it gets that same value at all larger τ's; the only transitions in value as σ increases are from 1/2 to 0 and from 1/2 to 1. Slightly more formally, we define the information ordering as follows.

Definition 3.1 $|A|_{v,\sigma} \leq_K |A|_{v,\tau}$ if and only if: if $|A|_{v,\sigma}$ is 1, then so is $|A|_{v,\tau}$, and if $|A|_{v,\sigma}$ is 0, then so is $|A|_{v,\tau}$.

Then the key monotonicity lemma:

(RM): If $\sigma < \tau$, then for all sentences A, $|A|_{v,\sigma} \leq_K |A|_{v,\tau}$.

(The details of the argument for (RM) are available in many places, e.g., Kripke [23].)[8]

Given (RM), two things follow. First, the rule for $\in$ can be simplified to:

(8*) If t_2 is $\{x : Ax\}$, then $|t_1 \in t_2|_{v,\sigma}$ is
  (a) 1 iff $(\exists\rho < \sigma)(|A(t/x)|_{v,\rho} = 1)$,
  (b) 0 iff $(\exists\rho < \sigma)(|A(t/x)|_{v,\rho} = 0)$,
  (c) 1/2 otherwise.

Second, and of crucial importance, cardinality considerations ensure that there is a point Ψ_v past which σ can never change, yielding "final values" for each sentence at each world. Letting $|A|_v$ abbreviate $|A|_{v,\Psi_v}$, this gives us the crucial "fixed-point condition": for any closed t and any A with no variables beyond x free,

(FP): $|t \in \{x : A\}|_v = |A(t/x)|_v$.

In contrast to the static construction, *dynamic* microconstructions allow for certain changes in the values of conditionals as the microconstruction proceeds. In the dynamic constructions the function v still plays a role in determining the value of conditionals, but unlike in the static construction, it is not the whole story; the value of a conditional is now in part determined by the values of its antecedent and consequent. Of course, one can only allow limited forms of changes during the microconstruction if monotonicity is to be preserved (and to give up monotonicity would be to give up the central idea of the construction). But here is one useful example of a dynamic construction (and the only one we will consider in detail in the present 3-valued setting): we keep everything the same except that we replace the valuation rule for $\rightarrow$ by

$$|A \rightarrow B|_{v,\sigma} = \begin{cases} 1 & \text{iff } v(A \rightarrow B) = 1, \\ 0 & \text{iff } v(A \rightarrow B) = 0, \text{ or } [v(A \rightarrow B) = 1/2 \\ & \quad \text{and } |A|_{v,\sigma} = 1 \text{ and } |B|_{v,\sigma} = 0], \\ 1/2 & \text{otherwise; that is, iff } v(A \rightarrow B) = 1/2 \\ & \quad \text{and } (|A|_{v,\sigma} < 1 \text{ or } |B|_{v,\sigma} > 0). \end{cases}$$

This modification in the rules for the conditional does not prevent us from arguing inductively that once a sentence has value 1 it retains that value throughout the

Kripke–Gilmore construction, and the same for 0. Given this, the fixed-point argument goes through as before; (FP) holds for the dynamic construction as well as the static.

Let a formula of form $t \in \{x : C(x; u_1, \ldots, u_k)\}$ and its corresponding formula $C(t; u_1, \ldots, u_k)$ be *basic equivalents*. Call a prevaluation v *transparent* if for any sentences A and B and any A^* and B^* obtainable from A and B, respectively, by sequences of substitutions of basic equivalents, $v(A \to B) = v(A^* \to B^*)$. Then we have the following.

Intersubstitutivity Corollary: If a sentence A is obtainable from a sentence B by a sequence of substitutions of basic equivalents, and v is transparent, then A and B have the same value at the fixed point over v.

(This is immediate from (FP) and the valuation rules for the static and dynamic microconstructions. Without the transparency assumption about v, all we could conclude from (FP) is that this holds when the only substitutions are outside the scope of an $\to$.)

Now call a prevaluation v *reflexive* if $v(C \to C) = 1$ for every sentence C.

Corollary on Abstraction: If v is transparent and reflexive, it gives value 1 to each instance of the abstraction schema.

(Reflexivity together with transparency lead by the intersubstitutivity corollary to $v(t \in \{x : A\} \to A(t/x)) = v(A(t/x) \to t \in \{x : A\}) = 1$ for each t, and so $|t \in \{x : A\} \leftrightarrow A(t/x)|_v = 1$. Since quantification is treated substitutionally, we get abstraction immediately.)

As we will see, there are plenty of transparent and reflexive prevaluations. A trivial example is the prevaluation that assigns value 1 to every conditional. This trivial prevaluation, in fact, even validates extensionality, in both rule and conditional forms. If the task were merely to validate both abstraction and extensionality, we would be done already. But of course we also want the conditional to obey reasonable laws, for example, modus ponens, which this valuation fails to deliver.

Before describing Brady's treatment of the conditional (which we will call the *macroconstruction*), we prove an important theorem about the microconstructions. This "microextensionality theorem" is really the heart of Brady's extensionality result.

4 The Microextensionality Theorem

Let a and b be closed abstracts.

If v is a prevaluation, call it

- $\langle a, b\rangle$-*congruent* if, for all formulas $C(x)$ and $D(x)$ with no variables other than x free, $v(C(a) \to D(a)) = v(C(b) \to D(b))$;
- $\langle a, b\rangle$-*extensional* if, for every closed L^+ term t, $|t \in a|_v = |t \in b|_v$;
- *strongly* $\langle a, b\rangle$-*congruent* if $|a \in t|_v = |b \in t|_v$ for every closed L^+ term t.

For transparent v, we could equivalently say

- *strongly* $\langle a, b\rangle$-*congruent* if, for all formulas $C(x)$ with no variables other than x free, $|C(a)|_v = |C(b)|_v$.

Thus in the case of transparent v, strong $\langle a, b\rangle$-congruence entails ordinary $\langle a, b\rangle$-congruence, at least in the case of the static microconstruction.

Theorem (Microextensionality theorem) *If v is transparent, $\langle a, b\rangle$-congruent, and $\langle a, b\rangle$-extensional, then it is strongly $\langle a, b\rangle$-congruent.*

This theorem holds for both static and dynamic microconstructions. To establish it, it is convenient to reformulate it. Let $\Sigma_{a,b,v}$ be the set of formulas $A(x)$ with no free variables other than x such that, for the final values $|A(a)|_v$ and $|A(b)|_v$, $|A(a)|_v \neq |A(b)|_v$. (We note that $|A(a)|_v \neq |A(b)|_v$ would not be possible unless x were free in $A(x)$, so we might as well have said that x and only x is free in $A(x)$.) Then a further reformulation of the claim that v is strongly $\langle a, b\rangle$-congruent is that $\Sigma_{a,b,v} = \emptyset$.

Given this way of stating the claim, the microextensionality theorem can be put as: for any prevaluation v that is transparent and $\langle a, b\rangle$-congruent,

(1) If $\Sigma_{a,b,v} \neq \emptyset$, then v is not $\langle a, b\rangle$-extensional.

For the remainder of this section, we will assume that v is a transparent, $\langle a, b\rangle$-congruent prevaluation.

With a few more definitions, we can simplify this statement even further. For each sentence B, let $\mu_v(B)$ be the first level of the Kripke construction at which B assumes its final value in the construction based on the prevaluation v. (So if the final value of B is $1/2$ in that construction, $\mu_v(B)$ is 0.) We will call this the *level* of B (relative to v).

For any σ, let $\Sigma_{a,b,v,\sigma}$ be the set of formulas $A(x)$ in $\Sigma_{a,b,v}$ such that at least one of $A(a)$ and $A(b)$ has level at most σ and also has value in $\{0, 1\}$ at the fixed point of the microconstruction. (The second requirement ensures that, at σ, $A(a)$ and $A(b)$ already have different values even if one of them does not yet have its final value.) Obviously, if $|A(a)|_v \neq |A(b)|_v$, then one of them has value 1 or 0 at some σ, so we can rephrase (1) as:

(1*) If $\exists\sigma(\Sigma_{a,b,v,\sigma} \neq \emptyset)$, then v is not $\langle a, b\rangle$-extensional.

We break the proof of this claim into two lemmas.

Lemma 4.1 $\forall\sigma$[if $\Sigma_{a,b,v,\sigma} \neq \emptyset$, then $\Sigma_{a,b,v,\sigma}$ contains formulas of form $t(x) \in x$].

Proof Assuming that $\exists\sigma(\Sigma_{a,b,v,\sigma} \neq \emptyset)$, let $\delta_{a,b,v}$ be the smallest ordinal σ such that $\Sigma_{a,b,v,\sigma} \neq \emptyset$. When $\sigma < \sigma^*$, $\Sigma_{a,b,v,\sigma} \subseteq \Sigma_{a,b,v,\sigma^*}$ by definition; so the lemma will be established if we establish the instance where σ is $\delta_{a,b,v}$. (For the remainder of the proof, a, b, v will remain fixed, so we will suppress mention of them; thus $|A|$ without any subscripts will mean the value of A at the fixed point of the microconstruction over v.) We establish this claim by establishing its contraposition, which we prove by induction on complexity. That is, we establish that if no formula of form $t(x) \in x$ is in Σ_δ, then $\Sigma_\delta = \emptyset$, making use of the fact that, for all $\rho < \delta$, $\Sigma_\rho = \emptyset$. (That is, for any $\rho < \delta$ and any $B(x)$, if $|B(a)|_\rho \in \{0, 1\}$, then $|B(b)|_\rho = |B(a)|_\rho$, and similarly with a and b reversed.)

Atomic formulas $B(x)$ with at most x free are:

(i) atomic formulas of the ground language;
(ii) formulas of form Class$(t(x))$;
(iii) formulas of form $t(x) \in n$ where n is a name in N;
(iv) formulas of form $t(x) \in \{y : B(x, y)\}$; or
(v) formulas of form $t(x) \in x$.

(Recall that we are allowing that $t(x)$ not contain x free and that $B(x, y)$ not contain x free and/or not contain y free; it is just that no variables other than those displayed can be free in these expressions.)

No formulas of form (i) can be in any Σ_σ: a and b are abstracts, so when x is free in $B(x)$ for ground-language atomic B, $|B(a)| = |B(b)| = 0$ (and when x is not free in $B(x)$, $B(a)$ and $B(b)$ are the same sentence).

Similarly for case (ii): the only terms in which a and b occur are terms for classes, and in that case $|\text{Class}(t(a))| = |\text{Class}(t(b))| = 1$.

Similarly, again, for case (iii): if $n \in N$, then $|t \in n| = 0$ for any closed term t.

As for (iv), suppose that $t(x) \in \{y : B(x, y)\}$ is in Σ_δ; then at least one of $|t(a) \in \{y : B(a, y)\}|_\delta$ and $|t(b) \in \{y : B(b, y)\}|_\delta$ is in $\{0, 1\}$, and we can suppose without loss of generality that the first is. But if $|t(a) \in \{y : B(a, y)\}|_\delta = 1$, then there are $\rho < \delta$ such that $|B(a, t(a))|_\rho = 1$ (and hence $|B(a, t(a))| = 1$). By the choice of δ, this requires that $|B(b, t(b))| = 1$. But then

$$\left|t(a) \in \left\{y : B(a, y)\right\}\right| = \left|t(b) \in \left\{y : B(b, y)\right\}\right|_\delta = 1,$$

which contradicts the supposition that $t(x) \in \{y : B(x, y)\}$ is in Σ_δ. The analogous argument holds for $|t(a) \in \{y : B(a, y)\}|_\delta = 0$; so no formula of form (iv) is in Σ_δ.

Finally, in case (v), by hypothesis no formulas of this form are in Σ_δ.

So, putting the cases together, no atomic formulas are in Σ_δ.

But then the result holds for nonatomic formulas too: for it is clear that

(a) if $B(x)$ is not in Σ_δ, then $\neg B(x)$ cannot be either;
(b) if neither $B(x)$ nor $C(x)$ is in Σ_δ, then $B(x) \wedge C(x)$ cannot be;
(c) if, for all closed terms t, $B(t, x)$ is not in Σ_δ, then $\forall y B(y, x)$ cannot be.

What about the conditional? Here we appeal to the $\langle a, b\rangle$-congruence of v. On the static microconstruction, that directly yields

(d_{static}) $B(x) \rightarrow C(x)$ cannot be in Σ_δ.

But even on the dynamic, the $\langle a, b\rangle$-congruence of v yields that a difference between $|B(a) \rightarrow C(a)|_\delta$ and $|B(b) \rightarrow C(b)|_\delta$ requires a difference either between $|B(a)|_\delta$ and $|B(b)|_\delta$ or between $|C(a)|_\delta$ and $|C(b)|_\delta$, so we have

(d) if $B(x)$ and $C(x)$ are not in Σ_δ, then $B(x) \rightarrow C(x)$ cannot be. □

The proof in fact shows that if $\Sigma_{a,b,v} \neq \emptyset$, then the only atomic members of $\Sigma_{a,b,v,\delta_{a,b,v}}$ are of form $t(x) \in x$.

Lemma 4.2 *Let $\delta_{a,b,v}$ be as in the proof of the preceding lemma. Then if $\Sigma_{a,b,v,\delta_{a,b,v}}$ contains a formula of form $t(x) \in x$, then v is not $\langle a, b\rangle$-extensional (and in particular, either $|t(a) \in a|_v \neq |t(a) \in b|_v$ or $|t(b) \in a|_v \neq |t(b) \in b|_v$).*

Proof (Once again, we suppress mention of a, b, v in the proof.) If $t(x) \in x$ is in Σ_δ, then at least one of $|t(a) \in a|_\delta$ and $|t(b) \in b|_\delta$ is in $\{0, 1\}$, and we can suppose without loss of generality that the first is. So either

(i) $|t(a) \in a|_\delta = 1$ and $|t(b) \in b| < 1$
or
(ii) $|t(a) \in a|_\delta = 0$ and $|t(b) \in b| > 0$.

Since a is a closed abstract, we can write it as $\{y : A(y)\}$.

In case (i), the first conjunct implies that $|A(t(a))|_\rho = 1$ for some $\rho < \delta$; by the choice of δ, Σ_ρ must have been empty, so

$$\big|A\big(t(b)\big)\big| = \big|A\big(t(a)\big)\big| = 1$$

and hence $|t(b) \in a| = 1$. But this and the second conjunct imply that $|t(b) \in a| \neq |t(b) \in b|$, so we have a violation of $\langle a, b\rangle$-extensionality, as required. Case (ii) is similar. (Of course, if $|t(b) \in b|_\delta$ rather than $|t(a) \in a|_\delta$ was in $\{0, 1\}$, it would have been $t(a)$ rather than $t(b)$ that provided the counterinstance to $\langle a, b\rangle$-extensionality.) □

Given Lemmas 4.1 and 4.2, the microextensionality theorem is immediate: when v is transparent and $\langle a, b\rangle$-congruent, $\langle a, b\rangle$-extensionality rules out $\Sigma_{\delta,a,b,v}$ containing any formula of form $t(x) \in x$ (by Lemma 4.2), and that entails that $\Sigma_{\delta,a,b,v}$ is empty (by Lemma 4.1).

At this point we might be tempted to follow a suggestion of Maddy's mentioned in endnote 2: introduce a new primitive $=$ into the language, not entering into the microconstruction, and extend the fixed-point valuation to include it by the condition that $|a = b|_v = 1$ if and only if either $a, b \in N$ and $a = b$ in the original model M, or a, b are closed abstracts and $\forall x(|x \in a|_v = |x \in b|_v)$. (There are several choices for the 0-clause.) The microextensionality theorem demonstrates that the $=$ defined in this way will validate every instance of the schematic rule $a = b, \phi(a) \vDash \phi(b)$.[9] But this trick will not give us a genuinely naive theory of classes: since this manner of introducing $=$ into the language does not allow for its appearance in formulas that occur in the term-forming operator, full abstraction (and even comprehension; see endnote 1) is lost. To have naïveté together with extensionality will take something more.

5 Brady's Macroconstruction(s) and Extensionality

We will now introduce that something more: the macroconstruction for conditionals. We will consider the versions developed Brady [5], [7], along with some minor variants.[10] In each of these constructions, the space of "worlds" is well-ordered: we can label the worlds by an initial segment of the ordinals. For every ordinal α, each "macroconstruction" assigns to the world labeled by α a "prevaluation" v_α, and we will write $|A|_\alpha$ for the value of A at world α, that is, the value of A at the minimal fixed point of the microconstruction over v_α. At any world α, the prevaluation v_α assigns to each conditional a value determined entirely by the values of its antecedent and consequent at worlds prior to it in the well-ordering (i.e., labeled by prior ordinals). So in both static and dynamic constructions, whether a conditional gets value 1 at a world α is determined entirely by the values of its antecedent and consequent at prior worlds. (At stage 0 all conditionals get value 1 in both static and dynamic constructions.) In the static construction, the same is true for value 0; in the dynamic, whether it gets value 0 at α is partly determined by the values of the antecedent and consequent at prior worlds, but may also be determined in part by the values of the antecedent and consequent at α itself. In the static constructions, as we will see, the values of conditionals decrease or remain the same as the ordinals increase.

Brady himself uses static constructions and a contraposable conditional that we will write as $\Rightarrow$. In all cases his 1-clause is

$(1\Rightarrow)$ $\quad |A \Rightarrow B|_\alpha = v_\alpha(A \Rightarrow B) = 1$ iff $(\forall \beta < \alpha)(|A|_\beta \leq |B|_\beta)$.

For the 0-clause (which we will write using v_α so as to allow consideration of dynamic variants) Brady [5] uses

$(0{\Rightarrow_A})$ $v_\alpha(A \Rightarrow B) = 0$ iff $(\exists\beta < \alpha)(|A|_\beta = 1$ and $|B|_\beta = 0)$.

Brady [7], by contrast, takes $\Rightarrow$ to be bivalent; there he uses the 0-clause

$(0{\Rightarrow_B})$ $|A \Rightarrow B|_\alpha = v_\alpha(A \Rightarrow B) = 0$ iff $(\exists\beta < \alpha)(|A|_\beta > |B|_\beta)$, i.e., iff $|A \Rightarrow B|_\alpha \neq 1$.

(In this case, there is no room to distinguish the dynamic construction from the static.) The details of the 0-clause will make no difference to the proof that the extensionality rule is sound on the models: all that matters is that the right-hand side be incompatible with the right-hand side of the 1-clause and have the form $(\exists\beta < \alpha)\Theta(\beta)$, where $\Theta(\beta)$ does not contain α free.

As we mentioned earlier, it is more general to introduce a noncontraposable $\rightarrow$, and define $\Rightarrow$ from it, so that $A \Rightarrow B$ means $(A \rightarrow B) \wedge (\neg B \rightarrow \neg A)$. For this we use as the 1-clause

$(1{\rightarrow})$ $|A \rightarrow B|_\alpha = v_\alpha(A \rightarrow B) = 1$ iff

$$(\forall\beta < \alpha)\big(\text{if } |A|_\beta = 1, \text{ then } |B|_\beta = 1\big).$$

$(1{\rightarrow})$ implies $(1{\Rightarrow})$, given the definition of $\Rightarrow$. As we will see, the extensionality proof will also go through in this more general setting.

What about the 0-clause for $\rightarrow$? Once again, for the extensionality rule it makes no difference, beyond the constraints mentioned in connection with $\Rightarrow$. Here are two possibilities:

$(0{\rightarrow_A})$ $v_\alpha(A \rightarrow B) = 0$ iff $(\exists\beta < \alpha)(|A|_\beta = 1$ and $|B|_\beta = 0)$;
$(0{\rightarrow_B})$ $v_\alpha(A \rightarrow B) = 0$ iff $v_\alpha(A \rightarrow B) \neq 1$, i.e., iff

$$(\exists\beta < \alpha)\big(|A|_\beta = 1 \text{ and } |B|_\beta < 1\big).$$

With $\Rightarrow$ defined as above, these rules induce the corresponding rules for $\Rightarrow$.

A third possibility is

$(0{\rightarrow_C})$ $v_\alpha(A \rightarrow B)$ is never 0, i.e., $v_\alpha(A \rightarrow B) = 1/2$ iff

$$(\exists\beta < \alpha)\big(|A|_\beta = 1 \text{ and } |B|_\beta < 1\big).$$

This would not be very interesting in connection with the static microconstruction, but in connection with the dynamic construction we outlined earlier it yields

$$|A \rightarrow B|_\alpha = \begin{cases} 1 & \text{iff } (\forall\beta < \alpha)(\text{if } |A|_\beta = 1, \text{ then } |B|_\beta = 1), \\ 0 & \text{iff } (\exists\beta < \alpha)(|A|_\beta = 1 \text{ and } |B|_\beta < 1) \\ & \quad \text{and } |A|_\alpha = 1 \text{ and } |B|_\alpha = 0. \end{cases}$$

In this case, $|A \Rightarrow B|_\alpha$ is 0 if and only if $(\exists\beta < \alpha)(|A|_\beta > |B|_\beta)$ and $|A|_\alpha = 1$ and $|B|_\alpha = 0$.

We will confine our consideration of dynamic microconstructions to $(0{\rightarrow_C})$. (A dynamic construction with $(0{\rightarrow_B})$ would coincide with the static, and with $(0{\rightarrow_A})$, the dynamic clause mentioned earlier would not produce interestingly different results.)

The crucial fact about all these Brady-like constructions is that whatever the 0-clause (provided it meets the constraints above), we have

(*) if $\alpha < \beta$, then for all A and B, $v_\beta(A \rightarrow B) \leq v_\alpha(A \rightarrow B)$.

(Here $\leq$ is the normal numerical order, as opposed to the information order $\leq_K$ used in the microconstructions.) In the case of the static constructions this means that $|A \rightarrow B|_\beta \leq |A \rightarrow B|_\alpha$, but in the dynamic with $(0{\rightarrow}_C)$ it is not ruled out that $A \rightarrow B$ has value 0 at one stage and 1/2 at a later stage. (*) implies that we eventually reach a fixed point where increasing α makes no difference to the values of sentences. We take this fixed-point Ω_M to be the previously mentioned @, we define a sentence to be M-valid if it takes value 1 at Ω_M, and we define an inference to be M-valid if it preserves 1 there. A sentence is valid (simpliciter) if it takes value 1 at Ω_M for all ground models M; similarly, an inference is valid if it preserves value 1 at Ω_M for all ground models M. The ordinal Ω_M associated with this fixed-point value may differ for different ground models M but in what follows, we will often speak as if M has been fixed, and refer to the fixed point simply as Ω. At this Ω, the 1-clause for $\rightarrow$ yields

$(BFP1)$ $v_\Omega(A \rightarrow B) = 1$ iff $(\forall\beta)$(if $|A|_\beta = 1$, then $|B|_\beta = 1$).

Since v_Ω is one of the v_β's, we can immediately conclude the following.

Corollary *If* $v_\Omega(A \rightarrow B) = 1$ *and* $|A|_\Omega = 1$, *then* $|B|_\Omega = 1$.

Even in the dynamic, $|A \rightarrow B|_\alpha = 1$ iff $v_\alpha(A \rightarrow B) = 1$, so we can replace the first conjunct of the antecedent by $|A \rightarrow B|_\Omega = 1$. Thus, given our definition of validity, we have

Modus Ponens: $A, A \rightarrow B \vDash B$.

We also get a fixed-point result for 0, though exactly how this works depends on the 0-clause we use. For the three listed we have

$(BFP0_A)$ $v_\Omega(A \rightarrow B) = 0$ iff $(\exists\beta)(|A|_\beta = 1$ and $|B|_\beta = 0)$;

$(BFP0_B)$ $v_\Omega(A \rightarrow B) = 0$ iff $v_\Omega(A \rightarrow B) \neq 1$, i.e., iff

$$(\exists\beta)\big(|A|_\beta = 1 \text{ and } |B|_\beta < 1\big);$$

$(BFP0_C)$ $v_\Omega(A \rightarrow B) \neq 0$.

Since v_Ω is one of the v_β's, a corollary of $(BFP0_A)$ is that if $|A|_\Omega = 1$ and $|B|_\Omega = 0$, then $v_\Omega(A \rightarrow B) = 0$ and hence $|A \rightarrow B|_\Omega = 0$. $(BFP0_B)$ gives the stronger result, that if $|A|_\Omega = 1$ and $|B|_\Omega < 1$, then $v_\Omega(A \rightarrow B) = 0$ and hence $|A \rightarrow B|_\Omega = 0$. $(BFP0_C)$ obviously yields no result guaranteeing $v_\Omega(A \rightarrow B) = 0$; however, given that it will be used only with the dynamic construction, here too we have that if $|A|_\Omega = 1$ and $|B|_\Omega = 0$, then ($v_\Omega(A \rightarrow B) \neq 1$ and hence) $|A \rightarrow B|_\Omega = 0$. So each of the constructions (provided that 0-clause C is used only with the dynamic) validates not only modus ponens but also

Contra-Modus-Ponens: $A, \neg B \vDash \neg(A \rightarrow B)$.

To illustrate just one difference between the different clauses and constructions, note that both the $(0{\rightarrow}_A)$ of Brady [5] and the $(0{\rightarrow}_B)$ of Brady [7] validate the rather odd law $\neg(\top \rightarrow \neg(A \rightarrow B))$: since all conditionals have value 1 at the starting valuation of the macroconstruction, their negations are guaranteed to have value 0, while $\top$ has value 1 there; so $\top \rightarrow \neg(A \rightarrow B)$ will have value 0 from stage 1 onward. $(0{\rightarrow}_C)$ with the dynamic construction avoids this odd consequence: the "bad" starting valuation merely forces $v_\alpha(\top \rightarrow \neg(A \rightarrow B))$ to have value 1/2 when α is at least 1, and the dynamic clause leaves it there unless $A \rightarrow B$ itself gets value 1. So the logic resulting from the dynamic has at least this advantage over the static.

Readers who would like a clearer sense of the mechanics of these constructions may wish to look at the following endnote, where we justify the validity of one important law.[11]

Taking stock, we have given a common two-part fixed-point construction governing a variety of different conditionals (both contraposable and noncontraposable; and with different 0-clauses, some static and some dynamic). The first part is a microconstruction focused on the membership relation; the second part is a macroconstruction for the conditional. The two-part construction turns an arbitrary classical base model for the ground language into a 3-valued modal model for the enlarged language. (So far we are dealing with "uncontracted" models, in which distinct closed abstracts denote different things even if the model declares them coextensive.) We define validity in terms of preservation of the value 1 at the "base world" Ω_M (for every formula relative to an assignment of values to the free variables), in every model M. In Section 4 we proved the important microextensionality theorem for the dynamic and static microconstructions. From this, we get a unified proof of Brady's core result for all of these constructions.

Theorem 5.1 (Brady extensionality theorem) *In all of the constructions, we have the rule*

$$\text{Class}(a) \wedge \text{Class}(b) \wedge \forall u(u \in a \Leftrightarrow u \in b) \vDash \forall z(a \in z \Leftrightarrow b \in z); \tag{4}$$

indeed, we have the following noncontraposable conditional form:

$$\vDash \forall x \forall y\big[\text{Class}(x) \wedge \text{Class}(y) \wedge \forall u(u \in x \Leftrightarrow u \in y) \rightarrow \forall z(x \in z \Leftrightarrow y \in z)\big]. \tag{5}$$

And the version with $(0 \rightarrow_B)$ *even yields*

$$\vDash \forall x \forall y\big[\text{Class}(x) \wedge \text{Class}(y) \wedge \forall u(u \in x \Leftrightarrow u \in y) \Rightarrow \forall z(x \in z \Leftrightarrow y \in z)\big]. \tag{6}$$

Proof To prove (5), we need that if a and b are closed abstracts, then for all α, if $|\forall u(u \in a \Leftrightarrow u \in b)|_\alpha = 1$, then $|\forall z(a \in z \Leftrightarrow b \in z)|_\alpha = 1$. And for each α, the antecedent holds if and only if, for all u and all $\beta < \alpha$, $|u \in a|_{v_\beta} = |u \in b|_{v_\beta}$, that is, if and only if all such v_β are $\langle a, b\rangle$-extensional; whereas the consequent holds if and only if, for all z and all $\beta < \alpha$, $|a \in z|_{v_\beta} = |b \in z|_{v_\beta}$, that is, if and only if all such v_β are strongly $\langle a, b\rangle$-congruent. So what we need is

(*) for all α, if all v_β's with $\beta < \alpha$ are $\langle a, b\rangle$-extensional, then they are all strongly $\langle a, b\rangle$-congruent.

It is clear by induction that each v_β is transparent. So microextensionality guarantees that, for each β and each a and b, if v_β is $\langle a, b\rangle$-congruent and $\langle a, b\rangle$-extensional, then v_β is strongly $\langle a, b\rangle$-congruent. So on the assumption that (i) all v_β's with $\beta < \alpha$ are $\langle a, b\rangle$-extensional, we can conclude that (ii) if all such v_β's are $\langle a, b\rangle$-congruent, then they are all strongly $\langle a, b\rangle$-congruent. And it is immediate from the Brady construction that (iii) for each β, if $(\forall \gamma < \beta)$ (v_γ is strongly $\langle a, b\rangle$-congruent), then v_β is $\langle a, b\rangle$-congruent. Putting (ii) and (iii) together, we have, on assumption (i), that for each $\beta < \alpha$, if $(\forall \gamma < \beta)(v_\gamma$ is strongly $\langle a, b\rangle$-congruent), then v_β is strongly $\langle a, b\rangle$-congruent; so by induction, all such v_β's are strongly $\langle a, b\rangle$-congruent, as desired.

That proves (5), and (4) follows from it. And with rule $(0 \rightarrow_B)$, (6) reduces to (5), since with that rule the antecedent $\text{Class}(x) \wedge \text{Class}(y) \wedge \forall u(u \in x \Leftrightarrow u \in y)$ and the consequent $\forall z(x \in z \Leftrightarrow y \in z)$ can only take on the classical values 0 or 1. □

It is worth noting that in any of these constructions, we have a converse of extensionality, even in $\Rightarrow$ form:

$$\vDash \forall x \forall y \big[\text{Class}(x) \wedge \text{Class}(y) \wedge \forall z(x \in z \Leftrightarrow y \in z) \Rightarrow \forall u(u \in x \Leftrightarrow u \in y)\big]. \tag{7}$$

For any u, simply let z be $\{w : u \in w\}$, and apply the intersubstitutivity corollary, as guaranteed by the transparency of all v, plus (FP). So with the bivalent 0-clause $(0 \rightarrow_B)$, extensionality holds with $\Leftrightarrow$, and with the others it holds with $\rightarrow$ in one direction and with either $\rightarrow$ and $\Rightarrow$ in the reverse.

6 Identity

For the Brady construction, it remains only to introduce identity, show that it obeys reasonable laws, and make explicit the familiar technique of turning the models we have constructed into "normal" models where the identity relation is standard. In the logics with non-bivalent conditionals, the identity predicate will not be bivalent either, so it will be worth making explicit how the normalization technique works there. Most of what follows will be independent of the 0-clause for conditionals.

As anticipated in Section 2, we define $x = y$ as[12]

$$(x =_L y) \vee \big[\text{Class}(x) \wedge \text{Class}(y) \wedge \forall u(u \in x \Leftrightarrow u \in y)\big].$$

Note that by this definition and the properties of the Brady construction, if $\alpha < \beta$ and $|x = y|_\alpha < 1$, then $|x = y|_\beta < 1$.

Observation *For x and y in the domain of M, $=$ coincides with $=_L$. More precisely: if either t_1 or t_2 is a name for a member of the domain of the ground model, then for any ordinal α, $|t_1 = t_2|_\alpha$ is the same as $|t_1 =_L t_2|_\alpha$ (which is the same for all α).*

The proof is trivial, given that $|\text{Class}(t)|_\alpha$ is 0 whenever t denotes an object in the ground model.

It is also clear that if for any α $|t_1 =_L t_2|_\alpha$ is 1, then t_1 and t_2 denote the same object of the domain of the ground model and so, for every β, $|t_1 \in z \Leftrightarrow t_2 \in z|_\beta = 1$. Given this and the Observation, the strong form of extensionality ((5) of Theorem 5.1) becomes

$$\vDash \forall x \forall y \big[x = y \rightarrow \forall z(x \in z \Leftrightarrow y \in z)\big].$$

And since for any formula $A(u)$ with at most u free, there is a set $\{u : A(u)\}$; this immediately yields a strong form of substitutivity

Substitutivity: $\vDash \forall x \forall y[x = y \rightarrow [A(x/u) \Leftrightarrow A(y/u)]]$.

(Here A may contain free variables other than u, in which case the validity of the substitutivity claim as written is tantamount to the validity of its universal closure.)

Note that this form of substitutivity is *much* stronger than the rule form

$$x = y, A(x/u) \vDash A(y/u).^{13}$$

It is also trivial that identity is reflexive and symmetric (in a strong $\Leftrightarrow$ form):

Reflexivity: $\vDash \forall x(x = x)$;
Symmetry: $\vDash \forall x \forall y(x = y \Leftrightarrow y = x)$.

And we have two forms of transitivity:

Transitivity:
(1) $\vDash \forall x \forall y \forall z[(x = y \wedge y = z) \rightarrow x = z]$;
(2) $\vDash \forall x \forall y \forall z[x = y \rightarrow (y = z \Leftrightarrow x = z)]$.

Proof Form (2) follows from Substitutivity, taking $A(u)$ to be "$u = z$" (and using the symmetry of $\Leftrightarrow$). For (1), it suffices that, for any β, if $|x = y|_\beta = |y = z|_\beta = 1$, then $|x = z|_\beta = 1$. And that is immediate: the antecedent entails that, for all $\gamma < \beta$ and all u, $|u \in x|_\gamma = |u \in y|_\gamma$ and $|u \in y|_\gamma = |u \in z|_\gamma$; and that entails that, for all $\gamma < \beta$ and all u, $|u \in x|_\gamma = |u \in z|_\gamma$, and hence that $|x = z|_\beta = 1$. □

(We think it unlikely that these very strong forms of Substitutivity and Transitivity will hold for logics substantially different from those in this article, unless they use two primitive conditionals (i.e., unless the noncontraposable $\rightarrow$ in these laws is not one from which the contraposable $\Rightarrow$ is defined).[14])

In the semantics with the bivalent $\rightarrow$ (i.e., with 0-clause (B)), $\Rightarrow$ is equivalent to $\rightarrow$, so Substitutivity and both forms of Transitivity each hold in the stronger form in which the main conditionals are $\Rightarrow$. With bivalent $\rightarrow$, we also get excluded middle for identity statements.[15]

It remains only to show that we can contract the model M^+ constructed in Section 5 to a normal model $M^+/\approx$, one where

> for any o_1, o_2 in the domain $D^+/\approx$ of $M^+/\approx$, $\langle o_1, o_2\rangle$ satisfies "$x = y$" if and only if $o_1 = o_2$.

The method is basically the one familiar from classical logic, but requires slight care since we must proceed world by world (or, macrostage by macrostage). The elements of $D^+/\approx$ are to be equivalence classes of elements of D^+, using the equivalence relation

$$t_1 \approx t_1' \text{ if and only if } |t_1 = t_1'|_\Omega = 1,$$

where Ω is the Brady fixed point; that is, if and only if, for all α, $|t_1 = t_1'|_\alpha = 1$. Then for each α, we take the 3-valued extension E^*_α of $\in$ at the αth stage of the new model M^* to be given by

$$E^*_\alpha\big([t_1], [t_2]\big) = |t_1 \in t_2|_\alpha.$$

This stipulation is only possible if it makes no difference which representatives of the equivalence classes we use, that is, if whenever $t_1 \approx t_1'$ and $t_2 \approx t_2'$, $|t_1 \in t_2|_\alpha = |t_1' \in t_2'|_\alpha$; but this follows from

$$t_1 = t_1', t_2 = t_2' \vDash t_1 \in t_2 \Leftrightarrow t_1' \in t_2',$$

which is easily seen to follow using two instances of the substitutivity rule. (We do not need to use the axiom form.)

7 The Rule of Weakening

In addition to delivering the naive theory of classes, Brady's logics also have some other very desirable features. The construction in [5] delivers the logic TWQ^d, while the one in [7] delivers the stronger TJQ^d (for a statement of some important laws and rules of these logics, see Appendix A). But the logics have one feature we think

highly undesirable, especially in the context of a naive theory of classes: they fail to deliver the "rule of $\Rightarrow$-weakening"

$$B \vDash A \Rightarrow B,$$

or even the "rule of $\rightarrow$-weakening"

$$B \vDash A \rightarrow B.$$

Now, you might say, who wants that? We should not be able to infer from "I will eat dinner tonight" to "If I die 1 minute from now, then I will eat dinner tonight." But although we take this "Ramsey–Adams phenomenon" seriously, there are three reasons why using it to defend Brady's logic for naive class theory seems to us misguided.

First, the reasons the weakening rule fails for Brady do not seem to have anything to do with the Ramsey–Adams phenomenon: the rule fails for Brady in cases which do not exhibit the characteristic features of the Ramsey–Adams phenomenon, and, conversely, Brady validates instances of rule-weakening in paradigm examples of the Ramsey–Adams phenomenon.

Here is a typical example in which the weakening rule fails in Brady's logics. Although

$$\neg(\top \rightarrow \bot)$$

is valid (as is evident from the validity of Contra-Modus-Ponens),

$$\top \rightarrow \neg(\top \rightarrow \bot)$$

is not; indeed, its negation is. The reason is that at v_0, every conditional has value 1, so $\neg(\top \rightarrow \bot)$ has the "erroneous" value 0 at this undesignated world. This value soon gets "corrected": from v_1 on, $\top \rightarrow \bot$ is corrected to having value 0, so $\neg(\top \rightarrow \bot)$ is corrected to having value 1. But the original "error" leads to permanent errors in the values of many conditionals containing $\neg(\top \rightarrow \bot)$, such as $\top \rightarrow \neg(\top \rightarrow \bot)$: since at v_0 $\top$ has value 1 and $\neg(\top \rightarrow \bot)$ does not, $\top \rightarrow \neg(\top \rightarrow \bot)$ has the "erroneous" value 0 at v_1. And the rules of the Brady construction prevent the value of $\top \rightarrow \neg(\top \rightarrow \bot)$ from recovering: once a conditional gets value 0 it is doomed to stay there, it can never recover from the initial error. (All this holds for Brady's $\Rightarrow$ as well.)

As we said, this example (and the structural features of the model which generate it) seems to have nothing whatever to do with the Ramsey–Adams phenomenon. More generally: the Ramsey–Adams phenomenon motivates allowing failures of rule-weakening for some sentences, but it does not motivate allowing cases where $\vDash A \wedge B$ and $\nvDash (A \rightarrow B)$, never mind the case that we have here, where $\vDash A \wedge B$ and yet $\vDash \neg(A \rightarrow B)$ (if we take A to be $\top$ and B to be $\neg(\top \rightarrow \bot)$). (Certainly the above example from ordinary English is unlike this: "I will die 1 minute from now and I will eat dinner tonight" is unassertable, let alone an obvious logical truth.) The kind of extreme failure of rule-weakening exhibited in Brady's logics simply cannot be explained by reference to the Ramsey–Adams phenomenon.

In the opposite direction, the standard Ramsey–Adams examples such as the one above about eating dinner involve sentences in the ground language, not in the extended vocabulary. But nothing in Brady's constructions suggests that weakening will fail for such sentences. (Indeed, at least in the version of the constructions presented here, rule-weakening holds when we restrict attention to consequents which have vocabulary drawn exclusively from the ground language.) So not only

does Brady admit more failures of the weakening rule than are motivated by appeal to the Ramsey–Adams phenomenon, he also does not give us enough failures of the rule to remove the problematic examples.

A second reason why we should not use the Ramsey–Adams phenomenon to defend the failure of rule-weakening in Brady's logics is that one important application of conditionals is for restricted quantification: we explain "All A are B" as "$\forall x$(if Ax then Bx)." And for this, we need a conditional that does obey rule-weakening, since otherwise we cannot infer "All A are B" from "everything is B." A Ramsey–Adams conditional may be useful for other purposes, but it cannot be used for restricted quantification: in a classical framework we need something more like the material conditional for that, and in a nonclassical framework we need something that reduces to an analogue of the material conditional in classical contexts.[16]

Brady himself acknowledges this. In a more recent joint paper on restricted quantification (see Beall, Brady, Hazen, Priest, and Restall [3]), he and his coauthors propose a second conditional for restricted quantification that does yield rule-weakening. Unfortunately, this second conditional has not even been shown to validate naive comprehension or abstraction, let alone extensionality. But the point remains: in the context of restricted quantification, rule-weakening is a clear and widely accepted desideratum.

Our third and final reason for wanting the weakening rule in the context of naive class theory is that failures of that rule rob extensionality of much of its intuitive force. Consider again the two sentences $\top$ and $\neg(\top \rightarrow \bot)$. Let a be $\{x : \top\}$, and let b be $\{x : \neg(\top \rightarrow \bot)\}$. Since $\top$ and $\neg(\top \rightarrow \bot)$ have value 1 at the Brady fixed point, clearly $\forall x(x \in a)$ and $\forall x(x \in b)$ do too: a and b are universal classes. But because $\top \rightarrow \neg(\top \rightarrow \bot)$ does not have value 1 at the fixed point of the macroconstruction, neither does $\forall x(x \in a \rightarrow x \in b)$, so the Brady theory will not declare these two universal classes coextensive. (Indeed, since the negation of $\top \rightarrow \neg(\top \rightarrow \bot)$ does have value at this fixed point, it will actually declare the two universal classes noncoextensive!) Thus in the absence of a weakening rule, extensionality is not nearly as strong as we would have intuitively expected; it seems almost not to deserve the name.

8 Positive Brady Logic and Positive Bacon Logic

We take the observations of the previous section, especially the second and third, to show that if a useful naive theory of classes is possible, it will have to be in a logic that (among other things) validates a rule of weakening. (Of course, weakening alone is not enough to be useful: we noted before that we can get abstraction and extensionality trivially, by declaring all conditionals valid, and it is clear that the same tactic would give us weakening. But that is obviously not very interesting—to mention just one pathology, it would yield a drastic failure of modus ponens.)

The literature contains several proposals for naive theories of truth/satisfaction/properties in logics that do contain a rule of weakening: for instance, the revision constructions of Field [12] or the fixed-point construction for the "fibers" in Field [13]. But Brady's proof cannot be adapted to these other constructions, at least not without substantial modification of those constructions.[17]

But can we modify Brady's construction in such a way as to get weakening for one or both of the conditionals $\rightarrow$ and $\Rightarrow$? One strategy that might initially appear

promising, but is not, is to use the basic Brady construction but with a different v_0 as the starting point. In the constructions developed by Brady himself that we have considered,[18] v_0 assigns value 1 to every conditional. But what if we modified the construction to use some other transparent prevaluation h as v_0? That is, what if we used h to generate a Brady progression, as follows:

$$v_\alpha(A \to B) = \begin{cases} 1 & \text{iff } h(A \to B) = 1 \\ & \quad \text{and } (\forall \beta < \alpha)(\text{if } |A|_\beta = 1, \text{ then } |B|_\beta = 1), \\ 0 & \text{iff } h(A \to B) = 0 \\ & \quad \text{or } (\exists \beta < \alpha)(|A|_\beta = 1 \text{ and } |B|_\beta = 0). \end{cases}$$

(That is for 0-clause $(0 \to_A)$; modifications of the other 0-clauses would be analogous.) We might try using some kind of "super-macroconstruction" (revision-theoretic or fixed-point) to generate a suitable h.

But this approach is not promising even as a way to get a satisfactory logic with weakening, let alone one that gives rise to extensionality. For we will have a failure of weakening if there is a conditional $A \to B$ that the fixed point assigns value 0 but h assigns a nonzero value; for then at the fixed point, $\neg(A \to B)$ will get value 1 and $\top \to \neg(A \to B)$ will not. So to get weakening, we need to know at the start which conditionals must get value 0. But if we knew this, there would not be much reason to run the Brady construction at all. And even if we did know this, and still felt like using the Brady construction, getting extensionality by the obvious generalization of the Brady proof would require an h which is $\langle a, b\rangle$-congruent for every pair $\langle a, b\rangle$ for which the fixed-point prevaluation v_Ω is $\langle a, b\rangle$-extensional. There is no obvious procedure for finding a starting h that guarantees either weakening or extensionality, never mind both.

One thing that *does* work—we do not find it appealing as it stands, but it will be a springboard to a better suggestion in the next section—is to eliminate the involutive negation $\neg$ from the language, at least as applied to formulas not in the ground language. (We would then take $\vee$ and $\exists$ as well as $\wedge$ and $\forall$ as primitive.) While it would be possible to keep $\neg$, but with formation rules that allow it to apply only to formulas of the ground language, it is more convenient to drop it entirely, and for each ground predicate p introduce a dual predicate p^*. ($\bot^*$ is our new version of $\top$.) Call the resulting ground language L^{pos}, and call the extended language $L^{+,\text{pos}}$. We take the logic of the ground language to include the LEM-like axiom $p(x_1, \ldots, x_k) \vee p^*(x_1, \ldots, x_k)$ and the explosion like rule $p(x_1, \ldots, x_k), p^*(x_1, \ldots, x_k) \vDash \bot$. This is not really less expressive than keeping $\neg$ throughout the ground language, because the De Morgan laws for the language with $\neg$ would allow one to drive it in to atomic predicates and replace negations of those by their duals.

We do not want to use a similar trick for the extended language; the point was to get rid of involutive negation there. But since we have a primitive absurdity constant $\bot$, we will still be able to define various "negation like" operations applicable to the full language, using the conditional: for example, $A \to \bot$, $A \to (A \to \bot)$, and so on. We will say more about these in a moment.

But first, we need to say something about a choice point for the conditional in the new setting. Since we have no involutive negation, we can no longer define $\Rightarrow$ from $\to$. It would be possible to allow it as an additional primitive, governed by the same

valuation rule as before, but it will be important in what follows that we not do so; the logic we will be calling "positive Brady logic" uses the rule for $\rightarrow$, not for $\Rightarrow$.[19]

The difference between $\Rightarrow$ and $\rightarrow$ matters even in the negation-free setting, because there will still be sentences that get value $1/2$; and if $\Rightarrow$ were in the language, the difference between $1/2$ and 0 at some worlds would make a difference for which sentences received value 1 at other worlds, including the Brady fixed-point Ω. For instance, let a abbreviate $\{x : x \in x\}$. Then $|a \in a|_v$ is $1/2$ whatever the v, given that we are (as we have throughout) using *minimal* fixed points in the microconstruction. So $|a \in a|_v$ and $|\bot|_v$ are both less than 1 at all v, but $|a \in a \Rightarrow a \in a|_\Omega = 1$ while $|a \in a \Rightarrow \bot|_\Omega < 1$. On the other hand, with $\rightarrow$ but not $\Rightarrow$ or $\neg$ in the language, this phenomenon cannot arise: the distinction between a sentence A having value $1/2$ or 0 is then idle in that it makes no difference as to whether compounds containing A end up receiving value 1 at the designated world. (That is so for any of the 0-clauses, and for static or dynamic, which shows that these distinctions do not matter in the context of positive logic.)

To prove this last claim about the language with only $\rightarrow$, we will introduce a 2-valued model for the negation- and $\Rightarrow$-free language and show that the 3-valued and 2-valued models assign the same sentences value 1. The 3-valued model, which we will denote with a superscripted 3, is exactly as before, except that we are now using the restricted language; we can use a static clause for the conditional, or the dynamic clause that allows transitions from $1/2$ to 0. The 2-valued model, denoted by a superscripted 2, uses only two values, 0 and 1. The rules for all connectives except $\in$ remain unchanged; any dynamic element in the conditional becomes redundant, so the 2-valued construction is always static. Regarding $\in$, the 2-valued microconstruction treats 0 much like $1/2$ in the 3-valued: we start all sentences $a \in b$ at value 0, and sentences can pass from 0 to 1, but not in the other direction. (With the obvious modifications, the required monotonicity lemma (RM) goes through as usual.) In the macroconstruction, the 1-clause for $\rightarrow$ is as before; the 0-clause will not matter to the proof. We then have the following.

Theorem 8.1 *For all sentences A of $L^{+,\mathrm{pos}}$ and all α and σ, $|A|^3_{\alpha,\sigma} = 1$ if and only if $|A|^2_{\alpha,\sigma} = 1$.*

Proof The proof is by induction on pairs α and σ, ordered lexicographically, that is, with $\langle \alpha, \sigma \rangle < \langle \beta, \tau \rangle$ if and only if either $\alpha < \beta$, or $\alpha = \beta$ and $\sigma < \tau$. It suffices to show for each α and σ that

(*) $|B|^3_{\alpha,\sigma} = 1$ iff $|B|^2_{\alpha,\sigma} = 1$ whenever B is either of form $C \rightarrow D$ or of form $a \in b$ where b is a class abstract;

for then an obvious subinduction on complexity yields that this is so for all sentences A of any complexity.

Suppose that (*) holds for all pairs prior to $\langle \alpha, \sigma \rangle$ in the lexicographic ordering.

Then if $\sigma > 0$, it clearly holds for sentences of form $C \rightarrow D$: for by the induction hypothesis this holds for the earlier $\langle \alpha, 0 \rangle$; and so $|C \rightarrow D|^3_{\alpha,\sigma} = 1$ if and only if $|C \rightarrow D|^3_{\alpha,0} = 1$ if and only if $|C \rightarrow D|^2_{\alpha,0} = 1$ if and only if $|C \rightarrow D|^2_{\alpha,\sigma} = 1$. Moreover (still assuming $\sigma > 0$), (*) also holds for sentences of form $a \in \{x : C(x)\}$; for $|a \in \{x : C(x)\}|^3_{\alpha,\sigma}$ is 1 if and only if, for some $\rho < \sigma$, $|C(a)|^3_{\alpha,\rho}$ is 1, which by the induction hypothesis holds if and only if, for some $\rho < \sigma$, $|C(a)|^2_{\alpha,\rho}$ is 1, and hence if and only if $|a \in \{x : C(x)\}|^2_{\alpha,\sigma}$ is 1.

So we need only consider pairs of form $\langle\alpha, 0\rangle$. But in that case, $|a \in \{x : C(x)\}|^3_{\alpha,0}$ and $|a \in \{x : C(x)\}|^2_{\alpha,0}$ are 1/2 and 0, respectively, so both differ from 1. And the induction hypothesis tells us that, for all $\beta < \alpha$, $|C|^3_{\beta,\Psi} = 1$ if and only if $|C|^2_{\beta,\Psi} = 1$, and $|D|^3_{\beta,\Psi} = 1$ if and only if $|D|^2_{\beta,\Psi} = 1$; so given the rules for the noncontraposable $\rightarrow$ (static or dynamic), $|C \rightarrow D|^3_{\alpha,0} = 1$ if and only if $|C \rightarrow D|^2_{\alpha,0} = 1$. □

The Brady fixed-point result of course still holds for positive Brady logic: the restricted language and the 2-valued model do not affect the proof.

But positive Brady logic does have a new phenomenon: for every sentence in the language, not just for conditionals, values are nonincreasing as the construction proceeds:

(**) If $\alpha_1 < \alpha_2$, then $|A|_{\alpha_2} \leq |A|_{\alpha_1}$[20]

(where again this is the real $\leq$, not the information ordering $\leq_K$). And that means that in addition to the negation-free laws that are validated by the Brady construction with involutive negation, we get not only the weakening rule but even the stronger conditional form:

(WeakeningAx): $\vDash B \rightarrow (A \rightarrow B)$.

For if any B has value 1 at an α, it has value 1 at all $\beta < \alpha$, and so $A \rightarrow B$ has value 1 at α.

But what about abstraction and extensionality? Since we do not have $\Leftrightarrow$ in the language, we cannot use our earlier formulations of these laws. But since we are working with a 2-valued model, we do not need anything other than $\leftrightarrow$; the reason is that in the restricted value space, $\leftrightarrow$ has the crucial property that $|A \leftrightarrow B|_\Omega = 1$ if and only if for all α, $|A|_\alpha = |B|_\alpha$. Given the "only if" direction of this "if and only if," abstraction follows directly from the fixed-point theorem. And given both directions of it, extensionality reduces to the claim that when a and b are classes and for all α $\forall x(|x \in a|_\alpha = |x \in b|_\alpha)$, then for all α $\forall z(|a \in z|_\alpha = |b \in z|_\alpha)$; and that is proved in precisely the same way as before. (The claim is simply that $\langle a, b\rangle$-extensionality at every macrostage of the Brady construction entails strong $\langle a, b\rangle$-congruence at every stage, and that follows by an obvious induction using microextensionality and the fact that the initial macrostage is (weakly) $\langle a, b\rangle$-congruent whatever the a and b.)

So this approach does yield a logic with weakening, modus ponens, and our two class-theoretic axioms. (Indeed it yields the logic known as TJK$^+$; see Appendix A.) But having no involutive negation is a high price to pay. And while there are other negation like operations in the language, we do not think they are enough to fill the gap. Consider the obvious candidates, the operations definable from $\perp$ and iterated conditionals. It is clear that we can iterate the operation "$A \rightarrow$": for any A and B, define $A \rightarrow^{(n)} B$ inductively in the obvious way:

$A \rightarrow^{(0)} B$ is just B;
$A \rightarrow^{(n+1)} B$ is $A \rightarrow (A \rightarrow^{(n)} B)$.

And, when the ground language includes arithmetic, we can use the truth predicate to "infinitely disjoin" them:[21] putting it loosely, $A \rightarrow^{(\omega)} B$ is $\exists n[\mathrm{True}(\langle A \rightarrow^{(n)} B\rangle)]$.

This is loose, since the preceding did not introduce the numerical superscripts as quantifiable variables; a more accurate statement[22] would be:

> $A \rightarrow^{(\omega)} B$ is $\exists n$ [n is a natural number and $\exists x$[True(x) and x is the nth member of the sequence whose zeroth member is B and whose $(k+1)$st member is obtained from the kth by prefixing "$A \rightarrow$" for every k]].

We can now continue: for example, $A \rightarrow^{(\omega+n+1)} B$ is $A \rightarrow (A \rightarrow^{(\omega+n)} B)$; and we can define $A \rightarrow^{(\omega+\omega)} B$ analogously to how we defined $A \rightarrow^{(\omega)} B$. When we come to limit ordinals of higher complexity it becomes progressively more complicated, but we can continue a long way through the countable ordinals. (Doing it rigorously involves using a system of ordinal notations.) And insofar as we can do it, we have that if $\mu_1 < \mu_2$, then for all α, if $|A \rightarrow^{(\mu_1)} B|_\alpha = 1$, then $|A \rightarrow^{(\mu_2)} B|_\alpha = 1$. So, letting $\neg_{NI^\mu} A$ be $A \rightarrow^{(\mu)} \bot$, we have a transfinite sequence of progressively weaker "noninvolutive negation operators." ($\neg_{NI^0}$ is the maximally strong "negation": applied to anything at all, it yields an equivalent of $\bot$.[23] There is no way to define a weakest member of the sequence: roughly speaking, this is because the first ordinal for which there is no notation is a limit ordinal.[24]) But none of these "negation" operators behave very much like we would want a negation operator to behave: for instance, for each μ, $\neg_{NI^\mu}\neg_{NI^\mu}\top$ is equivalent to $\neg_{NI^\mu}\neg_{NI^\mu}\bot$ (both have value 1 at stages $\alpha < \mu$, 0 for later stages).

Bacon [1] (in the first five sections of his paper, before he introduces involutive negation) gives a construction which, though presented in a superficially quite different manner (see our Appendix B for exposition and discussion), is very similar to the one we have just described for positive Brady logic. In fact, for the positive fragment, Bacon's construction, like the positive Brady we have developed, delivers the logic TJK$^+$, as described in Appendix A. The only real differences between the constructions are:

(i) that instead of going to the fixed point, Bacon goes through only the finite stages of the Brady construction;
(ii) accordingly, he defines validity, not as preservation of value 1 at the fixed point, but preservation of the property of having value 1 at all the finite stages.

So although there is no finite stage where we have that if $|A|_n$ and $|A \rightarrow B|_n$ are both 1 then $|B|_n = 1$, the definition of validity in (ii) still yields modus ponens. And of course the weakening axiom is still valid. Bacon's construction thus preserves some of the nice features of the Brady positive construction as we have developed it. But we are not sure whether Bacon's construction has any advantages over the Brady positive logic construction, and there is what strikes us as a serious disadvantage: it leads to a dramatic failure of the existential quantifier form of reasoning by cases (MR2 in Appendix A). The sentence $\neg_{NI^\omega}\top$, that is, $\top \rightarrow^{(\omega)} \bot$, has value 1 at each finite n, so on Bacon's definition it is a logical truth (or anyway, a consequence of arithmetic). That is odd in itself, but especially so since on unpacking the definition, it is an existential quantification of the form

$$\exists n\big[n \text{ is a natural number and } F(n)\big],$$

where each $F(n)$ is equivalent, given number theory, to $\top \rightarrow^{(n)} \bot$. Each of the $\top \rightarrow^{(n)} \bot$ is inconsistent: each leads to absurdity by repeated applications of modus ponens. This is the promised drastic failure of reasoning by cases: an existential quantification counts for Bacon as a logical truth, even though each of its instances is

logically inconsistent. As we said, we know of no advantages of the Bacon positive logic over the Brady that could compensate for this.

Nonetheless, it is worth noting that the proof of extensionality for positive Brady logic carries over to Bacon's construction: if a and b are class abstracts and for all n, $|\forall u(u \in a \leftrightarrow u \in b)|_n = 1$, then an induction using the microextensionality theorem shows that for all n, $|\forall z(a \in z \leftrightarrow b \in z)|_n = 1$.

9 4-Valued Brady and Bacon Logics

In the final section of his paper Bacon gives a trick for introducing an involutive negation into his construction, while still preserving the weakening rule, indeed even the weakening axiom for the noncontraposable conditional $\rightarrow$. Bacon himself aborts the construction of the conditional valuation after all finite stages, so that his new proposal still suffers from the failure of reasoning by cases just discussed (in fact, the proposal as he states it does not even quite do what he claims it does: he claims it delivers explosion and reflexivity, but as we show in Appendix B, it can do so only if it denies that certain sentences express propositions and restricts the application of the laws to sentences which do express propositions). But Bacon's trick can also be used for Brady's positive logic, where we continue the construction through to the Brady fixed point and define validity in a model in terms of the values at that fixed point. We will develop Bacon's trick in this second setting.

The trick is to assign each sentence A an ordered pair $\langle |A|^+, |A|^- \rangle$ of values in $\{0, 1\}$ at each world; intuitively, we can think of $|A|^+$ as a value for A and $|A|^-$ as a value for its negation. We assume a classical model of the ground language, which takes sentences into $\{\langle 1, 0\rangle, \langle 0, 1\rangle\}$. In the Kripke microconstruction, as usual we take as given a prevaluation v that assigns values $v^+(A \rightarrow B)$ and $v^-(A \rightarrow B)$ in $\{0, 1\}$ for any sentences A and B. Using this v, we assign values as follows:

(1) $|\neg A|^+_{v,\sigma} = |A|^-_{v,\sigma}$ and $|\neg A|^-_{v,\sigma} = |A|^+_{v,\sigma}$;
(2) $|A \wedge B|^+_{v,\sigma} = \min\{|A|^+_{v,\sigma}, |B|^+_{v,\sigma}\}$;
(3) $|A \wedge B|^-_{v,\sigma} = \max\{|A|^-_{v,\sigma}, |B|^-_{v,\sigma}\}$;
(4) $|\forall x A(x)|^+_{v,\sigma} = \min\{|A(t/x)|^+_{v,\sigma} : t \text{ is a closed term of } L^+\}$;
(5) $|\forall x A(x)|^-_{v,\sigma} = \max\{|A(t/x)|^-_{v,\sigma} : t \text{ is a closed term of } L^+\}$;
(6) if t_1 is a closed term and t_2 is a name in the ground language, then $|t_1 \in t_2|_{v,\sigma} = \langle 0, 1\rangle$;
(7) if t_1 is a closed term and t_2 is $\{x : A(x)\}$, then $|t_1 \in t_2|^+_{v,\sigma}$ is 1 iff $(\exists \rho < \sigma)(\forall \tau \text{ in the interval } [\rho, \sigma))(|A(t/x)|^+_{v,\tau} = 1)$;
(8) if t_1 is a closed term and t_2 is $\{x : A(x)\}$, then $|t_1 \in t_2|^-_{v,\sigma}$ is 1 iff $(\exists \rho < \sigma)(\forall \tau \text{ in the interval } [\rho, \sigma))(|A(t/x)|^-_{v,\tau} = 1)$.

What about the conditional? The analogy is not perfect, but we can think of the $+$-term in the valuation of the conditional as similar to the 1-clause in the 3-valued case. So just as the static and dynamic constructions shared a 1-clause, they will share the rule for $|A \rightarrow B|^+$; namely,

$$|A \rightarrow B|^+_{v,\sigma} = v^+(A \rightarrow B).$$

But just as the static and dynamic constructions differed on the 0-clause in the 3-valued case, they will differ here on the relationship between v^- and $|A \rightarrow B|^-_{v,\sigma}$. In the static construction, the relationship is just identity:

(S) $|A \rightarrow B|^-_{v,\sigma} = v^-(A \rightarrow B)$,

whereas dynamic constructions may vary the relationship in one of several ways:

(D_{i}^{-}) $|A \to B|^{-}_{v,\sigma} = 1$ iff $v^{-}(A \to B) = 1$ and $|A|^{+}_{v,\sigma} = 1$; or
(D_{ii}^{-}) $|A \to B|^{-}_{v,\sigma} = 1$ iff $v^{-}(A \to B) = 1$ and $|B|^{-}_{v,\sigma} = 1$; or
(D_{iii}^{-}) $|A \to B|^{-}_{v,\sigma} = 1$ iff $v^{-}(A \to B) = 1$ and $|A|^{+}_{v,\sigma} = 1$ and $|B|^{-}_{v,\sigma} = 1$; or simply
(D_{iv}^{-}) $|A \to B|^{-}_{v,\sigma} = 1$ iff $|A|^{+}_{v,\sigma} = 1$ and $|B|^{-}_{v,\sigma} = 1$.

(With (D_{iv}^{-}), we do not really need v^{-} at all; or we can think of (D_{iv}^{-}) as a special case of (D_{iii}^{-}), where v^{-} always assigns value 1 to everything.) Whichever of these clauses for $|A \to B|^{-}_{v,\sigma}$ we choose, neither the positive nor the negative component of the value of a sentence ever goes from 1 to 0 as the microstage σ increases; so there must be a point Ψ_v after which no sentence changes value, that is, a fixed point in the microconstruction of the valuation for $\in$, which satisfies (FP).

We now do a Brady macroconstruction of v_α, where at each macro-stage α we determine $v_\alpha(A \to B)$ from the values of A and B at the minimal fixed points of the microconstructions at prior stages. As before, we simply write $|A|_\alpha$ for the value of A at the minimal fixed point of the microconstruction over v_α.

In the case of $v^{+}_{\alpha}(A \to B)$—which is just $|A \to B|^{+}_{\alpha}$ even on the dynamic, so we will write it that way—we set

$$|A \to B|^{+}_{\alpha} = 1 \text{ iff } (\forall \beta < \alpha)\big(|A|^{+}_{\beta} \leq |B|^{+}_{\beta}\big).$$

That is for a noncontraposable conditional; for a contraposable conditional it would be

$$|A \Rightarrow B|^{+}_{\alpha} = 1 \text{ iff } (\forall \beta < \alpha)\big(|A|^{+}_{\beta} \leq |B|^{+}_{\beta} \text{ and } |B|^{-}_{\beta} \leq |A|^{-}_{\beta}\big).$$

What about the v^{-}_{α}? As noted earlier, the choice is irrelevant if we use version (D_{iv}^{-}) of the dynamic conditional. For the static, and for the dynamic proposals (D_{i}^{-})–(D_{iii}^{-}), the most natural choice is

(*I*): $v^{-}_{\alpha}(A \to B) = 1$ iff $(\forall \beta < \alpha)(|A|^{+}_{\beta} = |B|^{-}_{\beta} = 1)$;

then for $\Rightarrow$ we have the same condition. (If (I) is chosen, then the $|A \to B|^{-}_{\alpha}$ clauses given in (D_{i}^{-})–(D_{iii}^{-}) can be simplified, using the observation of the next paragraph: that $+$-values of sentences can never pass from 0 to 1 as the macroconstruction proceeds. For instance, given that observation and (I), (D_{i}^{-}) is equivalent to

(I_{i}^{-}): $|A \to B|^{-}_{\alpha} = 1$ iff $|A|^{+}_{\alpha} = 1$ and $(\forall \beta < \alpha)(|B|^{-}_{\beta} = 1)$.)

So at the initial ($\alpha = 0$) stage, both v^{+} and v^{-} assign every conditional 1 (though in the dynamic versions this does not prevent some $|A \to B|^{-}_{v_0}$ from being 0). An easy induction on α, with a subinduction on complexity, shows that as α increases, neither the $+$ nor the $-$ value of any sentence can ever go from 0 to 1. This then leads to a fixed-point Ω_M as in the original Brady construction; every sentence gets one of the values $\langle 1, 0\rangle$, $\langle 0, 1\rangle$, $\langle 0, 0\rangle$, and $\langle 1, 1\rangle$ at the fixed point.[25]

How is validity to be defined? The best choice is to take it as preservation of the property of having $+$-value 1 at the fixed point (i.e., the property of having value either $\langle 1, 0\rangle$ or $\langle 1, 1\rangle$ there). This, however, means that a sentence and its negation can both be designated; the logic will be paraconsistent, in that the existence of such "dialetheias" will not lead to explosion. For instance, let b be the abstract $\{x : x \in x \to x \notin x\}$. By (FP), $b \in b$ is equivalent to $b \in b \to b \notin b$; so letting X abbreviate the sentence $b \in b$, we have that X is equivalent to $X \to \neg X$. On any of

the treatments of the negative component of conditionals we have discussed, an easy induction gives that at any α, X has value $\langle 1, 1\rangle$.[26]

We could avoid dialetheias by taking only the value $\langle 1, 0\rangle$ as designated. But this would have a very serious cost: if X is the sentence above, then $X \rightarrow X$ has value $\langle 1, 1\rangle$ at every α and thus at the fixed-point $\Omega_{\boldsymbol{M}}$, whatever the $\boldsymbol{M}$, so it would not come out designated; the reflexivity law for conditionals, $A \rightarrow A$, would fail. (This choice between going dialetheic and restricting the reflexivity law arises in Bacon's logics too, though he seems to think otherwise; see Appendix B.) A proposal restricting reflexivity in this way is not only intrinsically unattractive, it is a nonstarter for naive class theory: for if reflexivity fails, then abstraction must also fail. (Extensionality will too, though slightly less obviously.)

There is a way to avoid both dialetheias and restrictions on reflexivity: modify the clause for the negative values of conditionals from (I) to

(II): $v^-_\alpha(A \rightarrow B) = 1$ iff $(\forall \beta < \alpha)(|A|^+_\beta = |B|^-_\beta = 1$ and $|A|^-_\beta = |B|^+_\beta = 0)$;

or what is the same thing given the nonincreasing nature of the values,

- if $\alpha = 0$, then $v^-_\alpha(A \rightarrow B) = 1$;
- if $\alpha > 0$, then $v^-_\alpha(A \rightarrow B) = 1$ iff $(\forall \beta < \alpha)(|A|^+_\beta = |B|^-_\beta = 1)$ and $|A|^-_{v_0} = |B|^+_{v_0} = 0$.

However, this too has a high price: when B is itself a conditional, $|B|^+_{v_0}$ is never 0, so for $\alpha > 0$, $v^-_\alpha(A \rightarrow B)$ can never be 1, and thus we no longer get the law

Contra-Modus-Ponens: $A, \neg B \vDash \neg(A \rightarrow B)$.

For instance, $\top$ and $\neg(\top \rightarrow \bot)$ do not together imply $\neg(\top \rightarrow (\top \rightarrow \bot))$ on this account.[27]

Since the cost of giving up either reflexivity or Contra-Modus-Ponens would be high, we conclude that the best version of the construction in this section admits dialetheias. (It is not only dialetheic, it requires restrictions on excluded middle as well, e.g., for the Russell paradox.)[28] The resulting static construction validates a dialetheic version of TJK (without explosion, but with involutive negation) which we call DTJK, and which is described as usual in Appendix A. In the earlier 3-valued construction, the dynamic construction improved the static primarily by invalidating some undesirable laws of the static. Here, the most obvious effect of the dynamic is adding new laws: for instance, no sentence of the form $\neg(A \rightarrow B) \rightarrow \bot$ is valid in the static (since $|A \rightarrow B|^-_0$ is never 0), but $\neg(\bot \rightarrow B) \rightarrow \bot$ is valid in all of the dynamic constructions except for (D^-_{ii}), and $\neg(A \rightarrow \top) \rightarrow \bot$ is valid in all of the dynamic constructions except for (D^-_{i}).

What about extensionality? Unlike the case of positive logic, the $\leftrightarrow$ form of the extensionality rule does not hold in any of the 4-valued constructions. For let c be $\{x : X\}$, and let d be $\{x : \top\}$, where X is as above and $\top$ has value $\langle 1, 0\rangle$ at every world. Then $|\forall x(x \in c \leftrightarrow x \in d)|^+_{\Omega_{\boldsymbol{M}}} = 1$. But if e is $\{x : x \notin x\}$, then $|c \in e|_{\Omega_{\boldsymbol{M}}} = \langle 1, 1\rangle$, while $|d \in e|_{\Omega_{\boldsymbol{M}}} = \langle 0, 1\rangle$, so $|c \in e \rightarrow d \in e|^+_{\Omega_{\boldsymbol{M}}}$ is 0.

But this counterexample does not carry over to the $\Leftrightarrow$ form of the extensionality rule, since $|\forall x(x \in c \Leftrightarrow x \in d)|^+_{\Omega_{\boldsymbol{M}}}$ is 0, not 1 (given that $|\neg(x \in c) \rightarrow \neg(x \in d)|^+_{\Omega_{\boldsymbol{M}}}$ is 0 for any x). Indeed, the $\Leftrightarrow$ form of the extensionality rule *is* valid in the construction (whatever the v^- rule), as is even the axiom

$$\vDash \forall x \forall y\big[\text{Class}(x) \wedge \text{Class}(y) \wedge \forall u(u \in x \Leftrightarrow u \in y) \rightarrow \forall z(x \in z \Leftrightarrow y \in z)\big].$$

To see this, observe first that, for any α, $|A \Leftrightarrow B|^+_\alpha = 1$ only if, for all $\beta < \alpha$, $|A|_\beta = |B|_\beta$. So if $|\forall u(u \in a \Leftrightarrow u \in b)|^+_\alpha = 1$ for classes a and b, then v_β for $\beta < \alpha$ is $\langle a, b\rangle$-extensional, and we can show inductively, as before, that each v_β for $\beta < \alpha$ is strongly $\langle a, b\rangle$-congruent; hence, $|\forall z(x \in z \Leftrightarrow y \in z)|^+_\alpha = 1$. In fact, it is not just extensionality which holds: abstraction clearly holds as well in both its $\leftrightarrow$ form and also in its stronger $\Leftrightarrow$ form. The construction yields a naive theory of classes, whatever the v^- rule.

This is somewhat good news for dialetheists (at least, for those willing to restrict excluded middle as well). As we noted in Section 7, anyone who wants to use a conditional to define universal restricted quantification has reason to want that conditional to obey a rule of weakening: without it, there is no way to get the obviously desirable rule

RQ-Weakening-Rule: Everything is $B \vDash$ All A are B.

As we discussed earlier, the importance of this rule has been recognized even by some of those (whether dialetheists or not) who stress the importance of relevance conditionals for which the rule of weakening fails: they *also* posit nonrelevance conditionals for which the rule holds, to handle restricted quantification (see, e.g., [3]). So the good news is that dialetheists can combine the weakening rule with abstraction and extensionality.[29] In fact, as with the positive logic, we can strengthen the weakening rule to an axiom, and hence with restricted quantification we get

RQ-Weakening-Axiom: $\vDash$ Everything is $B \rightarrow$ all A are B.

As yet, we have no equally good result on extensionality for those who would avoid dialetheia and simply restrict excluded middle.

We ourselves are not entirely enthusiastic about dialetheic theories. One reason is that they inevitably lead to what we see as a Sophie's choice regarding restricted quantification: between giving up one or the other of the attractive principles

Negative Conjunctive RQ-Weakening:

$$\neg\exists x\big(A(x) \wedge \neg B(x)\big) \vDash \text{ All } A \text{ are } B;$$

RQ-Modus Ponens: All A are B, $A(c) \vDash B(c)$.

For the two together yield

$$\neg\exists x\big(A(x) \wedge \neg B(x)\big), A(c) \vDash B(c),$$

and applying this to sentences A and B we get $\neg(A \wedge \neg B), A \vDash B$ and hence

$$\neg A, A \vDash B,$$

which leads immediately to triviality in the presence of a dialetheia.[30]

Still, even setting this general worry about dialetheism aside, there is an additional reason why the construction in this section does not yield a satisfactory logic for a naive theory of classes; and considering this problem will lead us to a much more general problem, which shows that any dialetheic naive theory of classes will be unsatisfactory. The local problem is that the construction does not really avoid the problem for Brady's 3-valued logics that was discussed at the end of Section 7: it will not yield the result that all universal classes are coextensive, and hence will not yield what one would have thought would be a consequence of extensionality, namely,

Intuitive Extensionality for Universal Classes:

$$\forall x(x \in a) \wedge \forall x(x \in b) \vDash \forall z(a \in z \Leftrightarrow b \in z).$$

Recall the example: a is $\{x : \top\}$, b is $\{x : \neg(\top \rightarrow \bot)\}$. The logic has

(*) $\vDash \neg(\top \rightarrow \bot)$

(as presumably any satisfactory logic does), so it yields both $(\forall y)(y \in a)$ and $(\forall y)(y \in b)$. So a and b are on any reasonable criterion universal classes, and any interesting version of naive class theory will declare them identical. We saw that Brady's theory did not declare them identical, because it did not even declare them coextensive, and we blamed that on the absence of (rule-)weakening; but in fact, even the theory of this section, which has weakening for $\rightarrow$, fails to declare these universal classes coextensive. That would require the validity of

(**) $\top \Leftrightarrow \neg(\top \rightarrow \bot)$,

which fails in the theory (since $(\top \rightarrow \bot) \rightarrow \bot$ is invalid in it). Similarly, the failure of (**) shows that the right-hand side of Intuitive Extensionality for Universal Classes fails for a and b, as can be seen by instantiating z with $\{y : \forall x(x \in y \Leftrightarrow \top\}$. Of course, we would have (**) if we had weakening *for the conditional* $\Rightarrow$ *used in the extensionality rule*, but the construction of this section does not deliver that.[31]

In fact, there is a more general argument here, which shows that no naive dialetheic theory with involutive negation that validates modus ponens (and has the full structural rules) can satisfy a sufficiently strong form of the extensionality rule (where a dialetheic theory is one that postulates dialetheias, not merely allows them). The argument has three stages.

First stage: in the presence of involutive negation, we need a contraposable $\Rightarrow$ in the formulation of the extensionality rule, because with involutive negation, extensionality with a noncontraposable $\rightarrow$ simply cannot hold. For suppose there are A and B such that (i) $\vDash A \leftrightarrow B$, but (ii) $\nvDash \neg A \leftrightarrow \neg B$. Pick an arbitrary closed term t, and let c be $\{x : \neg(t \in x)\}$. Then $\vDash \neg A \leftrightarrow \neg B$ is equivalent to $\vDash \neg(t \in \{y : A\}) \leftrightarrow \neg(t \in \{y : B\})$, which in turn is equivalent to $\vDash \{y : A\} \in c \leftrightarrow \{y : B\} \in c$. So if (ii), then $\nvDash \{y : A\} \in c \leftrightarrow \{y : B\} \in c$. But if (i) holds, then $\vDash \forall x(x \in \{y : A\} \leftrightarrow x \in \{y : B\})$, so we have a violation of extensionality for $\rightarrow$.

Second stage: no dialetheic theory (assuming it validates modus ponens) can validate even the following restricted version of weakening for $\Rightarrow$:

Special Weakening: If $\vDash A$, then $\vDash \top \Rightarrow A$.

For suppose $\vDash B$ and $\vDash \neg B$. Special weakening applied to the first would yield $\vDash \top \Rightarrow B$, and hence by contraposition $\vDash \neg B \Rightarrow \bot$, which with modus ponens yields $\bot$ and hence triviality.

These two stages together show that in a dialetheic context with involutive negation (and modus ponens), we cannot have even special weakening *for the conditional used in the formulation of the extensionality rule.*

Now for the final stage, which is a generalization of the argument using (*) and (**): without special weakening for the conditional used in the formulation of extensionality, extensionality is too weak to deserve the name. To see this, suppose that B is a sentence for which special weakening fails: that is, $\vDash B$ but not $\vDash \top \Rightarrow B$. Since $\vDash B$, the set b defined as $\{x : B\}$ is a universal class. Let a be $\{x : \top\}$; it too is a universal class. But they can only be declared coextensive if $\vDash \top \Leftrightarrow B$, which by hypothesis we do not have. Similarly, by taking z to be $\{y : \forall x(x \in y \Leftrightarrow \top)\}$, we see that the right-hand side of Intuitive Extensionality for Universal Classes fails.

So we conclude that not only the particular dialetheic logic of this section, but any dialetheic logic with modus ponens, must fail to yield a genuinely extensional naive theory of classes.

10 Triviality

But can we do better in some other way? The following theorem shows that, effectively, we cannot.

Theorem 10.1 *Suppose the logic of* $\rightarrow$ *validates the axiom (**)* $(\top \rightarrow \bot) \rightarrow \bot$ *and delivers the following axioms and rules:*

Modus Ponens	$A, A \rightarrow B \vdash B$;
Weakening	$B \vdash A \rightarrow B$;
Prefixing	$A \rightarrow B \vdash (C \rightarrow A) \rightarrow (C \rightarrow B)$;
Suffixing	$A \rightarrow B \vdash (B \rightarrow C) \rightarrow (A \rightarrow C)$;
Quantifier Axiom	$\vdash \forall x A(x) \rightarrow A(a)$;
$\bot$*Df*	$\vdash \bot \rightarrow A$;
Generalization	*If* $\vdash A$, *then* $\vdash \forall x A$;
$\wedge$*-Intro*	$A, B \vdash A \wedge B$.

Then the naive theory of classes stated by using $\rightarrow$ *is Post-inconsistent.*

To show this it will be convenient to introduce some abbreviations. First, we observe that every logic in this article contains the prefixing and suffixing rules; in any such logic, the following three rules can be derived:

Transitivity	$A \rightarrow B, B \rightarrow C \vdash A \rightarrow C$;
Left Elimination	$A \rightarrow (B \rightarrow C), D \rightarrow B \vdash A \rightarrow (D \rightarrow C)$;
Right Elimination	$A \rightarrow (B \rightarrow C), C \rightarrow D \vdash A \rightarrow (B \rightarrow D)$.

For readability, in what follows we will abbreviate $=_{\mathrm{ALT}*}$, introduced in endnote 12, by $\approx$ (read as "is indiscernible from"), so that

$$a \approx b =_{\mathrm{df}} \forall z(a \in z \leftrightarrow b \in z).$$

The crucial observation is that for $\approx$ (though not for $=$ as we have defined it), the suppositions of the theorem guarantee that the following quasisubstitutivity principle holds:

Quasisubstitutivity: $A(a) \vdash a \approx b \rightarrow (\top \rightarrow A(b))$.

This is reminiscent of the substitutivity principles we rejected in endnote 13, but in addition to involving $\approx$ instead of $=$, the new principle has a "$\top \rightarrow$" that is not in the principles we rejected; the following proof of quasisubstitutivity would not go through without it.[32]

Proof

(1) $A(a)$	Assumption;
(2) $\top \rightarrow A(a)$	1, Weakening;
(3) $a \approx b \rightarrow a \in \{x : A(x)\} \leftrightarrow b \in \{x : A(x)\}$	Definition of $\approx$, Quantifier Axiom;
(4) $a \approx b \rightarrow (A(a) \rightarrow A(b))$	3, Abstraction, Left and Right Elimination;
(5) $a \approx b \rightarrow (\top \rightarrow A(b))$	2, 4, Left Elimination. □

For any formula A, let $\mathfrak{p}_A =_{\text{df}} \{x : A\}$, where x is any variable not free in A. Then it is clear that extensionality (together with abstraction and generalization) guarantees the following metarule:

Extensionality Metarule: If $\vdash A \leftrightarrow B$, then $\vdash \mathfrak{p}_A \approx \mathfrak{p}_B$.

Finally, we define the Hinnion class $\mathfrak{h}$ and the Hinnion sentence κ by:

$$\mathfrak{h} =_{\text{df}} \{y : \mathfrak{p}_{y \in y} \approx \mathfrak{p}_\bot\},$$
$$\kappa =_{\text{df}} \mathfrak{h} \in \mathfrak{h}.$$

The Hinnion class can be thought of as a variant of the Curry class, in which the relation $\approx$ between (naive) classes is used in place of the conditional.

The proof of the theorem is then as follows.

Proof

(1) $\kappa \leftrightarrow (\mathfrak{p}_\kappa \approx \mathfrak{p}_\bot)$	Def. of $\mathfrak{h}$, Abstraction and Reflexivity;
(2) $\mathfrak{p}_\kappa \approx \{x : \mathfrak{p}_\kappa \approx \mathfrak{p}_\bot\}$	1, Extensionality Rule;
(3) $(\mathfrak{p}_\kappa \approx \mathfrak{p}_\bot) \rightarrow (\top \rightarrow \{x : \mathfrak{p}_\bot \approx \mathfrak{p}_\bot\} \approx \mathfrak{p}_\bot)$	2, Quasisubstitutivity;
(4) $\mathfrak{p}_\bot \approx \mathfrak{p}_\bot$	$\bot$Df, Abstraction and Reflexivity;
(5) $\mathfrak{p}_\bot \in \{x : \mathfrak{p}_\bot \approx \mathfrak{p}_\bot\}$	4, Abstraction;
(6) $(\{x : \mathfrak{p}_\bot \approx \mathfrak{p}_\bot\} \approx \mathfrak{p}_\bot) \rightarrow (\top \rightarrow \mathfrak{p}_\bot \in \mathfrak{p}_\bot)$	5, Quasisubstitutivity;
(7) $\mathfrak{p}_\bot \in \mathfrak{p}_\bot \rightarrow \bot$	Def. of $\mathfrak{p}_\bot$ + Abstraction;
(8) $(\{x : \mathfrak{p}_\bot \approx \mathfrak{p}_\bot\} \approx \mathfrak{p}_\bot) \rightarrow (\top \rightarrow \bot)$	6, 7, Right Elimination;
(9) $(\top \rightarrow \bot) \rightarrow \bot$	Assumption (**);
(10) $(\{x : \mathfrak{p}_\bot \approx \mathfrak{p}_\bot\} \approx \mathfrak{p}_\bot) \rightarrow \bot$	8, 9, Transitivity;
(11) $(\mathfrak{p}_\kappa \approx \mathfrak{p}_\bot) \rightarrow (\top \rightarrow \bot)$	3, 10, Right Elimination;
(12) $(\mathfrak{p}_\kappa \approx \mathfrak{p}_\bot) \rightarrow \bot$	9, 11, Transitivity;
(13) $\kappa \rightarrow \bot$	1, 12, Transitivity;
(14) $\kappa \leftrightarrow \bot$	13, $\bot$Df;
(15) $\mathfrak{p}_\kappa \approx \mathfrak{p}_\bot$	14, Extensionality Metarule;
(16) $\bot$	12, 15, Modus Ponens;
(17) $\bot \rightarrow A$	Df;$\bot$
(18) A	16, 17, Modus Ponens. □

How bad is this negative result? We argued extensively in Sections 7 and 9 that weakening for the conditional used in the extensionality law was nonnegotiable. Modus ponens, prefixing, suffixing, and $\bot$Df seem to us to have a similarly incontestable status.

One might attempt to respond to the result by giving up on (**); indeed, that is the course followed in the positive Brady and positive Bacon logics of Section 8, where the extensionality rule used $\rightarrow$ rather than $\Rightarrow$ and $\rightarrow$-weakening did hold. (The logic of Section 9 did not have (**), but of course it did not have weakening for $\Rightarrow$ either.) But (**) follows from

(*) $\vdash \neg(\top \rightarrow \bot)$

by weakening from a contraposable $\rightarrow$. So if the weakening rule is assumed to hold for the conditional used in the extensionality rule, there are only two ways we could give up (**):

(i) use a noncontraposable conditional in the extensionality rule
or
(ii) reject (*).

We pointed out near the end of Section 9 that (i) is not viable in the presence of involutive negation. But (ii) is also seriously unattractive: not only does it require rejecting Contra-Modus-Ponens, it requires either rejecting *all* negated conditionals or else rejecting one of the rules $\neg(A \rightarrow B) \vdash \neg(A \rightarrow \bot)$ and $\neg(A \rightarrow \bot) \vdash \neg(\top \rightarrow \bot)$ since these obviously suffice for getting from any given negated conditional to $\neg(\top \rightarrow \bot)$.

So we conclude that the strength of extensionality can be bought only at the price of naïveté, as we have been understanding it.[33] There is, however, a different kind of response to the result, which is to offer an alternative understanding of naïveté. Perhaps the most obvious approach along these lines would be to try replacing the abstraction schema and class-abstracts from Section 1 by the following:

Comprehension Schema:

$$\forall u_1, \ldots, \forall u_n \exists y \big[\mathrm{Class}(y) \wedge \forall x \big(x \in y \Leftrightarrow A(x; u_1, \ldots, u_n)\big)\big].$$

Obviously the abstraction schema and class-abstracts together entail comprehension; but it is not obvious that the move from comprehension to abstraction could be achieved conservatively—even in the presence of extensionality. With or without extensionality, there is of course no problem in conservatively introducing 0-level abstracts, that is, abstracts of form $\{x : A(x; u_1, \ldots, u_n)\}$, where $A(x; u_1, \ldots, u_n)$ itself contains no abstracts. But the full abstraction schema we have been using was not limited in this way.[34]

There are two different ways to understand the comprehension schema, which yield different verdicts on whether abstraction is or is not obtainable by what is understood as a conservative extension of comprehension. The first ("indefinite extensibility") option takes the schema to apply not just to the language in which it is stated (the language L^{+-} that includes "Class," $\in$, and $\rightarrow$ but not the abstraction operator), but to any expansion of that language. In this case, the schema applies to the language with 0-level abstracts, and we can proceed to introduce 1-level abstracts in the same way we introduced 0-level abstracts. Since comprehension applies to the language with 1-level abstracts too, we can introduce 2-level abstracts; and so on. In this way, we get the abstraction schema conservatively from the comprehension schema *so understood*, without any form of extensionality.

A second, "tepid" understanding—which is perhaps the more usual one—is to take the schema as applying only to its own language L^{+-}. In that case, we could still advance beyond 0-level abstracts if our introduction of abstracts at each stage was definitional, for then we would in effect be introducing them into the object language. However, performing this extension conservatively requires extensionality; moreover, it requires a stronger form of extensionality than the rule form we have focused on, and indeed a stronger form than any we have shown to be consistent with comprehension. So if the naive theory of classes were formulated using comprehension instead of abstraction, and comprehension interpreted

in the "tepid" way, then it remains an open question whether the theory is consistent with a conditional logic including the "mandatory" principles we have been discussing.

This leaves us with two open questions, a mathematical one and a philosophical one. The mathematical question concerns whether *comprehension* and extensionality are consistent in a reasonably strong logic (ideally, one of the attractive logics known to be consistent with the naive theory of truth).[35] The philosophical question concerns whether the theory with only comprehension, and not abstraction, can justly be called a naive theory of classes. To put it another way: in accepting comprehension, it seems plausible that the naive theorist should accept the schema as applied to reasonable extensions of the language; but does the addition of $\{ : \}$ count as "reasonable" in the relevant sense? Clearly, if comprehension and extensionality are inconsistent in any logic with weakening for $\Rightarrow$ (for example), the philosophical question would be moot. But if they do turn out to be consistent, there may be interesting questions about the appropriate understanding of "naïveté" in this context, questions we have not attempted to address here.

Even if comprehension plus extensionality are inconsistent in any reasonable logic, this result would *not* show that the logics of naive truth and properties on offer in the literature are unsuited to the development of any extensional set theory at all. A naive theory of classes is *much* stronger than what is required for a set theory which is consonant with the philosophical motivations for preserving naïveté for truth or properties. In fact, it is much stronger even than what is needed to use these nonclassical logics to give a theory which is more satisfactory than standard classical theories which admit quantification over classes as well as over sets.

For now, all we can say is this: If naïveté requires abstraction, then the prospects for the naive theory of classes are dim indeed. If comprehension is enough for naïveté, however, there is still some hope. We have shown that a dialetheic theory will have trouble getting comprehension and extensionality in a reasonable logic, but we have not shown that other logics will face analogous difficulties. Still, we have shown that if comprehension and extensionality are consistent, they must be shown to be so by a model construction which does not also validate abstraction. Since Brady's method, as well as the standard methods from the literature on naive truth and satisfaction, all deliver Comprehension by way of Abstraction, proving these principles consistent will require a completely new approach.

Appendix A: Axioms and Rules

In the following table, we state rules and axioms for an arbitrary conditional $\rightarrow$ (which except in the case of Ext1–Ext5 is *not* assumed to be the "noncontraposable" conditional of the main text).

Ax1	$A \rightarrow A$	reflexivity
Ax2	$A \rightarrow A \vee B$ and $B \rightarrow A \vee B$	
Ax3	$A \wedge B \rightarrow A$ and $A \wedge B \rightarrow B$	
Ax4	$\neg\neg A \leftrightarrow A$	double negation

Ax5	$A \wedge (B \vee C) \rightarrow (A \wedge B) \vee (A \wedge C)$	
Ax6	$(A \rightarrow B) \wedge (A \rightarrow C) \rightarrow (A \rightarrow B \wedge C)$	strong lattice $\wedge$
Ax7	$(A \rightarrow C) \wedge (B \rightarrow C) \rightarrow (A \vee B \rightarrow C)$	strong lattice $\vee$
Ax8	$(A \rightarrow \neg B) \rightarrow (B \rightarrow \neg A)$	contraposition axiom
Ax9	$\neg(A \wedge B) \leftrightarrow (\neg A \vee \neg B)$	De Morgan 1
Ax10	$\neg(A \vee B) \leftrightarrow (\neg A \wedge \neg B)$	De Morgan 2
Ax11	$(A \rightarrow B) \rightarrow ((B \rightarrow C) \rightarrow (A \rightarrow C))$	suffixing axiom
Ax12	$(A \rightarrow B) \rightarrow ((C \rightarrow A) \rightarrow (C \rightarrow B))$	prefixing axiom
Ax13	$(A \rightarrow B) \wedge (B \rightarrow C) \rightarrow (A \rightarrow C)$	conjunctive syllogism
Ax14	$A \rightarrow (B \rightarrow A)$	weakening
Ax15	$(A \rightarrow B) \vee \neg(A \rightarrow B)$	conditional LEM
R1	$A, B \vdash A \wedge B$	adjunction
R2	$A, A \rightarrow B \vdash B$	Modus Ponens
R3	$A, \neg B \vdash \neg(A \rightarrow B)$	Contra-Modus-Ponens
R4	$A, \neg A \vdash B$	explosion
MR1	$\dfrac{A \vdash C \qquad B \vdash C}{A \vee B \vdash C}$	reasoning by cases
Q1	$\forall x A \rightarrow A(x/a)$	a free for x
Q2	$\forall x(A \vee B) \rightarrow A \vee \forall x B$	$x \notin FV\{A\}$
Q3	$\forall x(A \rightarrow B) \rightarrow (A \rightarrow \forall x B)$	$x \notin FV\{A\}$
Q4	$A(x/a) \rightarrow \exists x A$	a free for x
Q5	$A \wedge \exists x B \rightarrow \exists x(A \wedge B)$	$x \notin FV\{A\}$
Q6	$\forall x(B \rightarrow A) \rightarrow (\exists x B \rightarrow A)$	$x \notin FV\{A\}$
RQ	$\dfrac{\Gamma \vdash A(x/y)}{\Gamma \vdash \forall x A}$	$y \notin FV(\Gamma \cup \{\forall x A\})$
MR2	$\dfrac{A(x/y) \vdash B}{\exists x A \vdash B}$	$y \notin FV\{\exists x A, B\}$
Ext1	$\forall x(x \in a \Leftrightarrow x \in b) \vdash \forall x(a \in x \Leftrightarrow b \in x)$	
Ext2	$\forall x(x \in a \Leftrightarrow x \in b) \rightarrow \forall x(a \in x \Leftrightarrow b \in x)$	
Ext3	$\forall x(x \in a \Leftrightarrow x \in b) \Rightarrow \forall x(a \in x \Leftrightarrow b \in x)$	
Ext4	$\forall x(x \in a \leftrightarrow x \in b) \vdash \forall x(a \in x \leftrightarrow b \in x)$	
Ext5	$\forall x(x \in a \leftrightarrow x \in b) \rightarrow \forall x(a \in x \leftrightarrow b \in x)$	

Some differences between our constructions are given in the following table, which is not, however, intended to be comprehensive. Instead, we have merely listed some important and familiar laws and rules validated by the various constructions presented here.

Section		Axioms	Name of Logic
Brady 1983 ([5]) $((0 \Rightarrow_A)$, Section 5)	Logic	For $\rightarrow$: Ax1–Ax7, Ax9–Ax12, R1–R4, Q1–Q6, RQ, MR1–MR2 For $\Rightarrow$: Ax1–Ax12, R1–R4, Q1–Q6, RQ, MR1–MR2	TWQ^d + R3 + R4 (Note: Brady uses "TN^dQ" for TWQ^d + R3)
	Classes	For $\Rightarrow$: Comprehension, Abstraction, Ext1, Ext2	
Brady 2006 ([7]) $((0 \Rightarrow_B)$, Section 5)	Logic	For $\rightarrow$: Ax1–Ax7, Ax9–Ax13, Ax15 R1–R4, Q1–Q6, RQ, MR1–MR2 For $\Rightarrow$: Ax1–Ax13, Ax15 R1–R4, Q1–Q6, RQ, MR1–MR2	TJQ^d + R3 + R4
	Classes	For $\Rightarrow$: Comprehension, Abstraction, Ext1, Ext2, Ext3	
Positive Bacon (Only finite macrostages; Section 8)	Logic	For $\rightarrow$: Ax1–Ax3, Ax5-Ax7, Ax11–Ax14, R1–R2, Q1–Q6, RQ, MR1	TJK^+
	Classes	For $\rightarrow$: Comprehension, Abstraction, Ext4, Ext5	
Positive Brady (Section 8)	Logic	For $\rightarrow$: Ax1–Ax3, Ax5–Ax7, Ax11–Ax14, R1–R2, Q1–Q6, RQ, MR1–MR2	TJK^+ + MR2
	Classes	For $\rightarrow$: Comprehension, Abstraction, Ext4, Ext 5	
4-valued Bacon (Only finite macrostages; Section 9)	Logic	For $\rightarrow$: Ax1–Ax7, Ax9–Ax14, R1–R3, Q1–Q6, RQ, MR1 For $\Rightarrow$: Ax1–Ax13, R1–R3, Q1–Q6, MR1	DTJK-MR2
	Classes	For $\Rightarrow$: Comprehension, Abstraction, Ext1, Ext2	
4-valued Brady (Static version: Section 9)	Logic	For $\rightarrow$: Ax1–Ax7, Ax9–Ax14, R1–R3, Q1–Q6, RQ, MR1–MR2 For $\Rightarrow$: Ax1–Ax13, R1–R3, Q1–Q6, MR1–MR2	DTJK
	Classes	For $\Rightarrow$: Comprehension, Abstraction, Ext1, Ext2	

We have not included the variant of Brady with the $0 \Rightarrow_C$ clause (and dynamic microconstruction): as far as the laws in our list go it is like Brady [5] and distinguished from that logic primarily by *invalidating* certain obviously undesirable laws such as $\neg(\top \rightarrow \neg(A \rightarrow B))$. In the 4-valued constructions, the dynamic variants lead to extra laws: for instance, Bacon in effect uses our (D_{ii}^{-}), which delivers the law $\neg(A \rightarrow \top) \rightarrow \bot$ (as do (D_{iii}^{-}) and (D_{iv}^{-})); and the dynamic variants other than Bacon's yield $\neg(\bot \rightarrow B) \rightarrow \bot$.

Finally, we note a difference between our 4-valued Brady/Bacon-logic and Bacon's claims about his construction. Bacon claims (see [1, p. 101]) (12 in his list of axioms) that when negation is added his own "4-valued" construction delivers R4 (explosion), together with reflexivity. But as we show in the next appendix, this claim must either be restricted or rejected entirely: the logic can achieve explosion plus reflexivity only by denying that certain paradoxical sentences express propositions, and restricting these laws to those sentences which express propositions.

Appendix B: Bacon's Construction

The construction in Bacon [1] appears on its face to be quite different from the one given here; the point of this appendix is to show that they are in fact the same (so

that Bacon's appeal to the Banach fixed-point theorem is not really needed to give his result), and also to show that in his 4-valued construction he does not avoid the choice between accepting dialetheias and restricting reflexivity or Contra-Modus-Ponens.

B.1 Bacon's positive construction We begin with the construction in the first five sections of his paper, which deal with the positive logic (i.e., the logic without involutive negation). Here are some of his definitions.

Definition B.1 A function $f : \mathbb{N} \to \{0, 1\}$ *flatlines* if and only if for some $n \in \mathbb{N}$, $f(m) = 0$ for each $m > n$.

Bacon takes W (the set of worlds) to be $\{f : \mathbb{N} \to \{0, 1\} \mid f \text{ flatlines}\}$, and $f \leq g$ if and only if for each $n \in \mathbb{N}$, $f(n) \leq g(n)$.

These "flatlining functions" are simply the characteristic functions of finite subsets of the natural numbers. Bacon's $\leq$ corresponds to the relation of $\subseteq$ on these subsets.

For the positive logic, Bacon takes the propositions to be the nonempty downward closed subsets of W; so the strongest proposition is not $\emptyset$, but rather $\{\emptyset\}$. We can think of the proposition expressed by a sentence as the set of worlds at which it has value 1; so every sentence, even $\bot$, has value 1 at the minimal world $\emptyset$; it is the "inconsistent world." (So this minimal world is completely uninformative: perhaps it would be more natural to simply drop it, but to stay close to Bacon we will keep it.)

We evaluate conjunction and disjunction by the lattice-theoretic operations of meet ($\sqcap$) and join ($\sqcup$), defined as

$$(f \sqcap g)(n) = \min\big(f(n), g(n)\big), \qquad \text{and}$$
$$(f \sqcup g)(n) = \max\big(f(n), g(n)\big).$$

The quantifiers are evaluated analogously. This generates the usual conditions for the values of sentences at worlds (even the inconsistent world $\emptyset$, since there is as yet no negation in the language).

To handle the conditional, Bacon proceeds as follows.

Definition B.2 The *rank* of an element $f \in W$, $r(f)$, is the smallest n such that $f(m) = 0$ for all $m \geq n$. (Equivalently, the rank of a finite subset f of $\mathbb{N}$ is the smallest $n \in f$ such that all members of f are less than n.)

Definition B.3

$$w^* := \begin{cases} w(n) & \text{if } n < r(w) - 1, \\ 0 & \text{otherwise.} \end{cases}$$

Again, if we think in terms of finite subsets of $\mathbb{N}$, $\emptyset^* = \emptyset$, and otherwise w^* is what results from w by deleting its largest element.

Bacon then defines a ternary accessibility relation R by the following.

Definition B.4 $Rwyz$ if and only if $z \leq w^*$ and $z \leq y$.

In the set representation, this is $z \subseteq w^* \cap y$.

As usual in ternary relation semantics, he lets

$$|A \to B|_w = 1 \text{ iff } \forall y \forall z (\text{if } Rwyz \text{ and } |A|_y = 1, \text{ then } |B|_z = 1).$$

So with this choice of R, we have in the set representation that

$$|A \to B|_w = 1 \text{ iff } \forall y \forall z (\text{if } z \subseteq w^* \cap y \text{ and } |A|_y = 1, \text{ then } |B|_z = 1);$$

or equivalently,

$$|A \to B|_w = 1 \text{ iff } \forall z[\text{if } \exists y(z \subseteq w^* \cap y \text{ and } |A|_y = 1), \text{ then } |B|_z = 1];$$

which, since the propositions are downward closed, amounts to

($) $|A \to B|_w = 1$ iff $\forall z[\text{if } z \subseteq w^* \text{ and } |A|_z = 1, \text{then } |B|_z = 1]$.

(This simplification of the ternary relation semantics is due to the particular choice of R.)

To understand the implications of this, consider first the case where w is $\emptyset$ or a singleton. Then w^* is $\emptyset$, and since every B gets value 1 at $\emptyset$, the right-hand side of ($) always holds: thus at each singleton world, all conditionals get value 1. (Unlike the case of $\emptyset$, $\bot$ and false ground-language sentences get value 0 at singletons.) So there is no difference in values from one singleton world to another. (None for conditionals; and so by induction, none for any other sentences either.) Thus the value of every sentence at a singleton world is just the value of that sentence at level 0 of the positive-logic Brady construction in Section 8.

Similarly, when w is a doubleton world $\{m_1, m_2\}$ where $m_1 < m_2$, ($) yields (since the case where z is $\emptyset$ is trivial):

$$|A \to B|_{\{m_1,m_2\}} = 1 \text{ if and only if: if } |A|_{\{m_1\}} = 1, \text{ then } |B|_{\{m_1\}} = 1.$$

But we saw above that for singleton worlds $\{m_1\}$, values are the same whatever the m_1; this then tells us that for doubletons too, the members do not matter, all that matters is that there be two members. A conditional gets value 1 at a doubleton if and only if, if the antecedent gets value 1 at stage 0 of the positive-logic Brady construction, so does the conclusion. And that is just the condition for it to have value 1 at level 1 of the positive-logic Brady construction.

An obvious induction yields that when w is any finite set of positive cardinality n, the value of a conditional at w is just its value at $n-1$ in the positive-logic Brady construction. So as we said in Section 8, validity in Bacon's positive logic is just preservation of value 1 at all finite levels of the positive-logic Brady construction.

B.2 Bacon's 4-valued construction We turn now to Bacon's introduction of involutive negation. It follows the general plan that we used to extend Brady's logic (no surprise, since we modeled that after Bacon's proposal): a 4-valued semantics obtained by assigning a positive and negative value (each in $\{0, 1\}$) to each conditional, and extending this valuation to other sentences as in Section 9. The positive clause for the conditional is the one for the positive logic, except that here it is stated using $+$ values:

$$|A \to B|^+_w = 1 \text{ iff } \forall y \forall z(\text{if } z \subseteq w^* \cap y \text{ and } |A|^+_y = 1, \text{ then } |B|^+_z = 1),$$

which as above reduces to

($^+$) $|A \to B|^+_w = 1$ iff $\forall z[\text{if } z \subseteq w^* \text{ and } |A|^+_z = 1, \text{ then } |B|^+_z = 1]$.

As for the negative, Bacon agrees that there is a certain amount of choice, but he works with this one:

$$|A \to B|^-_w = 1 \text{ iff } \exists y \exists z[w^* \leq y \text{ and } w \leq z \text{ and } |A|^+_y = 1 \text{ and } |B|^-_z = 1],$$

which simplifies to

$$|A \to B|^-_w = 1 \text{ iff } |A|^+_{w*} = 1 \text{ and } |B|^-_w = 1;$$

this is easily seen to be equivalent to version (I) of the 4-valued Brady when using dynamic clause (D_{ii}) (though note again that Bacon goes only through the finite stages of this construction).

Bacon [1] takes validity to be the preservation of a designated value, but offers three choices for what that is. His preferred choice (see [1, p. 100, (iii)]) is to take a sentence A to be designated if and only if (a) for all w $|A|^+_w$ is 1, and (b) for some w $|A|^-_w$ is 0. Contrary to what Bacon claims, however, this has the consequence that $A \rightarrow A$ is not generally valid: the sentence X of Section 9 illustrates this problem since, as we noted there, $X \rightarrow X$ has value $\langle 1, 1\rangle$ at every stage. Another of Bacon's proposals (see [1, p. 100, (ii)]) is to take A to be designated if and only if for all w, $|A|^+_w$ is 1; this leads to a dialetheic logic similar to that in Section 9 (but which, because it does not proceed to the fixed point, has the additional oddity concerning reasoning by cases discussed in Section 8). The final possibility that he mentions (see [1, p. 100, (i)]) involves taking a sentence A to be designated if and only if (a) for all w, $|A|^+_w$ is 1, and (b*) for all w other than the inconsistent w, $|A|^-_w$ is 0; this has the oddity that no negation of a conditional is ever valid, so we have an even worse failure of Contra-Modus-Ponens than the one contemplated in Section 9.

These problems with Bacon's construction are masked by one particular oddity in his presentation. For some reason that he does not explain, Bacon builds into his proposal (iii) that a sentence does not count as expressing a proposition if it gets value $\langle 1, 1\rangle$ at every world. Similarly, he builds into his proposal (i) (the last one discussed here) that a sentence does not count as expressing a proposition if there are *any* worlds other than $\emptyset$ where it has value $\langle 1, 1\rangle$. So on both of these proposals, the sentence X from Section 9 is claimed not to express a proposition. Bacon does not tell us what to do with sentences which do not express propositions as far as the logic is concerned, but one might attempt to recast Bacon's proposal as one in which some sentences do not express propositions, and the laws are intended to hold only for those sentences which do. But this "fix" goes against the aim of the construction Bacon is engaged in, that of achieving a logic for naive truth: it deals with the paradoxes by taking a "no proposition" view of some of them. (What is more, in the case of proposal (i) the "fix" would not help with the problem about Contra-Modus-Ponens described at the end of the previous paragraph.)

We conclude that the dialetheic version of Bacon's construction is the best and that the construction of Section 9 gives a somewhat improved version of this, without the oddity about reasoning by cases.

In any case, a nice thing about Bacon's logic (which he may not have known) is that it is one in which the naive theory of classes, including extensionality, is consistent. (This virtue is preserved in the "4-valued Brady logic" of Section 9.) And unlike Brady's logics based on the 3-valued space $\{0, 1/2, 1\}$, it has the rule of weakening, indeed the axiom of weakening; we take this to be important at least if the conditional is to be of use for restricted quantification. But as we say in the text, further improvement would be needed before declaring this a satisfactory naive theory of classes, and there is reason to think that no dialethic theory can do the job.

Notes

1. Some may find it more natural to formulate the naive theory of classes without a primitive abstraction operator, replacing the abstraction schema with a comprehension schema

$$\forall u_1, \ldots, \forall u_n \exists y \big[\mathrm{Class}(y) \wedge \forall x \big(x \in y \Leftrightarrow A(x; u_1, \ldots, u_n)\big)\big].$$

Since abstraction clearly implies comprehension, the consistency results for abstraction in Sections 2–9 carry over immediately to comprehension. The difference between these principles will only matter for the negative result in Section 10, so we will postpone detailed discussion of it until then.

2. Some theories of "naive classes" or "naive sets" in the literature, for example, Gilmore [16], White [28] Grišin [18], and Maddy [24], are concerned at best with naive properties in our sense, since these theories do not include rules of extensionality. In the case of Gilmore and Maddy, in fact, we do not strictly have even naive property theory, because there is no $\Rightarrow$ in the language strong enough to deliver the abstraction schema. These theories are nonetheless "naive in spirit," and most naive property theories work by extending Gilmore's construction to include an appropriate $\Rightarrow$.

   Maddy [24, p. 134] at one point suggests adding a primitive $=$ to the language and giving a model-theoretic version of extensionality in terms of it, but even then there is no rule of the language that expresses extensionality. Moreover, when she considers this option, she restricts abstraction so that $=$ cannot appear in the scope of $\{:\}$: her suggestion gets "extensionality in spirit" only by giving up on "naïveté in spirit." (We should also note that while Maddy suggests that the proof that her added $=$ obeys intersubstitutivity is trivial, it certainly did not seem so to us, as the proof of the microextensionality theorem in Section 4 attests.)

   White [28] shows that the addition of an axiom of extensionality turns a naive property theory stated in terms of Łukasiewicz's continuum-valued logic inconsistent, but the naive property theory was already ω-inconsistent, so perhaps this is unsurprising.

   In the case of Grišin, the addition of extensionality is known to lead to (Post-)inconsistency (see Shirahata [27] and especially Cantini [10] for useful discussion); essentially that is because Grišin's logic validates permutation $A \to (B \to C) \vdash B \to (A \to C)$. In fact, as is shown in Øgaard [25, Theorem 15], the rule $A \to (B \to C), B \vdash A \to C$ already suffices for triviality in a naive class theory. The logics we study below will thus not validate this permutation principle. (And, unlike Grišin's, the logics we study will have the full structural rules.)

   There is in addition a series of papers by Hinnion and others (e.g., Forti and Hinnion [14], Hinnion [20], Hinnion and Libert [21], [22]) that show the consistency of various forms of extensionality, but with highly restricted abstraction: it is restricted to formulas in which neither negation, nor a conditional, nor even the abstraction operator itself occur. Because of these restrictions, these theories are not naive in our sense. Similarly, Brady [4] and Brady and Routley [9] restrict abstraction to the $\to$-free language, and so give up on naïveté.

3. Scott [26] showed that if standard ZF is formulated, as it usually is, using the axiom of replacement

$$\forall w_1, \ldots, w_n \forall a \big[(\forall x \in a)(\exists! y)\phi(x, y, w_1, \ldots, w_n) \supset (\exists b)(\forall y)\big(y \in b \equiv (\exists x \in a)\phi(x, y, w_1, \ldots, w_n)\big)\big],$$

   then this $\mathrm{ZF}_{\mathrm{REP}}$ minus extensionality is interpretable in Z and hence has weaker consistency strength than standard ZF. This result might seem to show that extensionality is responsible for adding important strength to standard set theory developed in classical logic. But ZF can equivalently be axiomatized using collection

$$\forall w_1, \ldots, w_n \forall a \big[(\forall x \in a)(\exists y)\phi(x, y, w_1, \ldots, w_n) \supset (\exists b)(\forall x \in a)(\exists y \in b)\phi(x, y, w_1, \ldots, w_n)\big]$$

   together with separation, instead of replacement, and Friedman [15] has shown that any model of the axioms of this $\mathrm{ZF}_{\mathrm{COL}}$ other than extensionality can be transformed into a

model of full ZF, including extensionality. This suggests that Scott's result might be better interpreted as demonstrating the weakness of replacement (as opposed to collection) rather than as illustrating the power of extensionality. (See Hamkins [19] for helpful discussion.) In our nonclassical context, where abstraction is assumed, the situation is very different. We impose no class-theoretic axioms beyond abstraction and extensionality, and since abstraction is known to be consistent in the logics we consider, the inconsistency result is due entirely to the strength of extensionality. This is of particular interest since naive abstraction yields all the standard principles of $\mathrm{ZF_{COL}}$ minus foundation and extensionality: the needed sets can in each case be defined as $\{x : A(x)\}$ for an appropriately chosen A plus parameters. So $\mathrm{ZF_{COL}}$ minus foundation and minus extensionality is consistent with naive abstraction in the known logics for naive properties. Extensionality tips the balance into inconsistency.

4. L might include the term "Set," and a predicate $\in_{\mathrm{set}}$ that is to be distinguished from the $\in$ that is added in the move to L^+. The classes to be introduced are then to be conceived, at least initially, as in all cases distinct from sets. (We might later consider an identification of sets with certain classes, but we see no clearly good way of doing this.)

5. It follows that for any sentence A of the ground language, if we "translate" it into L^+ by restricting all its existential quantifiers by the condition "$\neg\mathrm{Class}(x) \wedge \ldots$" and all its universal quantifiers by "$\mathrm{Class}(x) \vee \ldots$," then the translation has the same value in M^+ that the original had in M.

6. It is natural to put extensionality conditionally: "If a and b are classes with the same members, then they are members of the same classes." But this uses the ordinary English "if... then," which is not well modeled by the noncontraposable Brady $\rightarrow$. A more complicated language with two primitive conditionals might be required to yield an appropriate conditional formulation.

7. Every object in the domain of M^+ will be the denotation (in M^+) of either a new name in N or of a closed class abstract. Were it not for this, we would need to complicate the description of the various v_w's: each would map, into $\{0, 1/2, 1\}$, ordered pairs whose first member is a conditional formula $A \rightarrow B$ and whose second member is an assignment s of objects to variables, with the condition that if s and s^* agree in all variables free in $A \rightarrow B$, then $v_w(\langle A \rightarrow B, s\rangle) = v_w(\langle A \rightarrow B, s^*\rangle)$.

8. We have not added a truth or satisfaction predicate to L^+, but could easily have done so, provided the ground language is rich enough to encode syntax. We could ensure the naïveté of True or *Sat*, relative to a particular coding of syntax, within the same Gilmore–Kripke microconstruction as we used for classes. Focusing just on "True" for simplicity, we would use the following valuation rules:

   If t is not the Gödel number of an L^+ sentence, then $|\mathrm{True}(t)|_{v,\sigma} = 0$;
   If t is the Gödel number of the L^+ sentence A, then

$$\left|\mathrm{True}(t)\right|_{v,\sigma} = \begin{cases} 1 & \text{iff } (\exists\rho < \sigma)(\forall\tau \text{ in the interval } [\rho,\sigma))(|A|_{v,\tau} = 1), \\ 0 & \text{iff } (\exists\rho < \sigma)(\forall\tau \text{ in the interval } [\rho,\sigma))(|A|_{v,\tau} = 0), \\ 1/2 & \text{otherwise.} \end{cases}$$

   The simultaneous construction of valuations for truth and membership clearly will not affect the key monotonicity lemma (RM) in either case. So the construction would give us a fixed point where both truth and class membership are naive.

9. That rule is a rather weak form of substitutivity, and gives rise to transitivity only in the very weak form $a = b, b = c \vDash a = c$, but the absence of a useful conditional in the strong Kleene logic makes these the strongest forms of substitutivity and transitivity available there.

10. Brady [6] is a dialetheic theory, so we discuss it only later, in note 28.

    In a recent paper, Brady [8] uses "metavaluations" to provide some new metalogical results concerning logics which support naive class theory; his results divide into two parts, one concerning what he calls "M1" logics, and the other concerning "M2" logics. In M1 logics, no negations of conditionals are valid (i.e., for every A and B, $\nvdash \neg(A \rightarrow B)$); this renders them of little interest for the present project. (To mention just one reason: a satisfactory logic for reasoning about classes should allow us to say that it is not the case that everything in the universal class belongs to the empty class, that is, it should include $\vdash \neg\forall x(x \in \{y : \top\} \rightarrow x \in \{y : \bot\})$.) The 2013 paper also appears to have a number of errors in the proofs. For instance, Brady claims as one of the main logical achievements of the paper that his construction for M1 logics allows the extension of his consistency proof to logics which include the "special permutation" axiom $(A \rightarrow (B \rightarrow (C \rightarrow D))) \rightarrow (B \rightarrow (A \rightarrow (C \rightarrow D)))$ along with conjunctive syllogism. But this is in fact impossible, since it is shown in Øgaard [25, Corollary 1] that if one adds even the weaker rule form of special permutation $A \rightarrow (B \rightarrow (C \rightarrow D)) \vdash (B \rightarrow (A \rightarrow (C \rightarrow D)))$ to any logic with conjunctive syllogism (satisfying very minimal requirements that Brady's do satisfy), then one can derive the special contraction rule $A \rightarrow (A \rightarrow (A \rightarrow B)) \vdash A \rightarrow (A \rightarrow B)$. As Øgaard remarks, that rule (like ordinary contraction) is enough to trivialize even naive truth theory, which has no analogue of extensionality.

    While Brady offers new metalogical results for both M1 and M2 logics, the M2 logics he studies are essentially the ones from his earlier work. The main difference in the new models is that the conditionals are taken to have starting value 1 if and only if they are theorems of a prespecified deductive system. This modification means that, like some of our "dynamic" constructions, they avoid certain undesirable laws, although Brady does not state which exactly these new nonlaws are. For this reason, and since the errors in the paper make it difficult to know what can be shown using Brady's new technique, we do not consider it further.

11. Consider the prefixing rule for the contraposable conditional $\Rightarrow$:

$$A \Rightarrow B \vDash (C \Rightarrow A) \Rightarrow (C \Rightarrow B).$$

    (The proof of the analogous rule for the noncontraposable conditional is analogous.) If $|A \Rightarrow B|_\Omega = 1$, then $v_\Omega(A \Rightarrow B) = 1$ and hence by $(BFP1)$, for all α, $|A|_\alpha \leq |B|_\alpha$. We want to show that for all α we also have that $|C \Rightarrow A|_\alpha \leq |C \Rightarrow B|_\alpha$; that is, (i) if $|C \Rightarrow A|_\alpha = 1$, then $|C \Rightarrow B|_\alpha = 1$, and (ii) if $|C \Rightarrow B|_\alpha = 0$, then $|C \Rightarrow A|_\alpha = 0$. The antecedent of (i) requires that for all $\beta < \alpha$, $|C|_\beta \leq |A|_\beta$; hence since $|A|_\beta \leq |B|_\beta$ for *any* β, including those which are smaller than α, it is clear that for all $\beta < \alpha$, $|C|_\beta \leq |B|_\beta$, which gives us that $|C \Rightarrow B|_\alpha = 1$, as required. The antecedent of (ii) can never hold for a static conditional with 0-clause (C). For a static with 0-clause (A) it requires that for some earlier β, $|C|_\beta = 1$ and $|B|_\beta = 0$; but since $|B|_\beta \geq |A|_\beta$ for all such β, we also have that $|C|_\beta = 1$ and $|A|_\beta = 0$, guaranteeing that $|C \Rightarrow A|_\alpha = 0$, as required. For a static with 0-clause (B) the proof is exactly analogous. Making the conditional dynamic in any of the ways contemplated will not change anything, since the assumption of the rule still gives us that $v_\Omega(A \Rightarrow B) = 1$, and that guarantees that the value of B is at least that of A at every β.

12. An alternative definition is: $(x =_L y) \vee [\mathrm{Class}(x) \wedge \mathrm{Class}(y) \wedge \forall z(x \in z \Leftrightarrow y \in z)]$. Call this $=_{\mathrm{ALT}}$. It is easy to see that for any α, $|x = y|_\alpha$ is 1 if and only if $|x =_{\mathrm{ALT}} y|_\alpha$ is 1, and that if $|x = y|_\alpha$ is 0, then so is $|x =_{\mathrm{ALT}} y|_\alpha$; but the converse to the latter fails except when one uses the bivalent 0-clause B. We think the definition in the text more intuitive, but the choice between the two will not matter in what follows. (It might seem that we could use the simpler $=_{\mathrm{ALT}*}$: $\forall z(x \in z \Leftrightarrow y \in z)$. But when x and y are distinct objects in the ground model, $|x =_{\mathrm{ALT}*} y|_0 = 1$, which (since $|x =_L y|_0 = 0$) prevents $|x =_{\mathrm{ALT}*} y \rightarrow x =_L y|_\Omega$ from being 1 for such x and y. So if $=_{\mathrm{ALT}*}$ were used for $=$, we would not have the desired $|x = y \Leftrightarrow x =_L y|_\Omega = 1$ for x and y in the ground model.)

13. Some papers in the literature (e.g., Bacon [2], Øgaard [25]) consider another strengthening of substitutivity,

    BadSubst: $A(x/u) \vDash x = y \rightarrow A(y/u)$,

    and derive various paradoxes from the addition of other principles to this. But we think (BadSubst) unacceptable in any context where the bivalence of identity is not assumed. This is clearly so if the language is rich enough to formulate a "determinately operator" D: then an instance of (BadSubst) is $D(x = x) \vDash x = y \rightarrow D(x = y)$, which given $\vDash D(x = x)$ yields $\vDash x = y \rightarrow D(x = y)$, that is, that if x and y are identical they are determinately so. (Evans's argument in [11] against indeterminate identity was basically of this form, though he used a slightly different principle, which follows from this one if $\rightarrow$ is contraposable.) So we do not see these paradoxes as presenting new challenges for naive class theory; in any case, none of the approaches to extensionality contemplated in this paper would validate (BadSubst). In contrast to these earlier results, we think the negative result of the final section is important because it involves principles which we cannot see an independent motivation for denying.

14. We do think that any adequate theory will have rule forms at least as strong as

    Substitutivity Rule: $x = y \vDash A(x/u) \Leftrightarrow A(y/u)$

    (and the generalization that puts any sequence of universal quantifiers over variables other than x and y before the conclusion), and its corollary

    Transitivity Rule: $x = y \vDash \forall z(y = z \Leftrightarrow x = z)$.

    And preferably, it will have basically the axiom forms in the text. But as we say there, this will perhaps require that $\rightarrow$ be replaced with a second primitive conditional (which might be similar to a "Ramsey–Adams conditional" as discussed in the next section).

15. One might wonder what about a form of substitutivity analogous to form (1) of transitivity, that is,

    (StrongSub): $\vDash \forall x \forall y[x = y \wedge A(x/u) \rightarrow A(y/u)]$.

    But this fails even in the bivalent case. For instance, for any x and y, $|x = y|_0$ is 1, so StrongSub will fail for any $A(u)$ for which there are any x and y for which $|A(x)|_0 = 1$ and $|A(y)|_0 < 1$!

16. It might be thought that we need not only rule-weakening, but a conditional form of it, either $\vDash B \rightarrow (A \rightarrow B)$ or an analogue with $\Rightarrow$; for we want not only to be able to infer "All A are B" from "everything is B," but to assert

    If everything is B, then all A are B.

    Perhaps, but we do not take it as obvious that the explicit "if…then" in the displayed sentence is the same as the conditional used for restricted quantification. For example, one might want this second conditional to respect the Ramsey–Adams phenomenon.

17. The following example is a counterexample to extensionality in both constructions, using the abbreviations $a = \{x : \bot\}$, $b = \{x : \top \Rightarrow \bot\}$, and $c = \{x : \forall y(x \in y \Leftrightarrow b \in y)\}$. In both constructions (unlike in Brady's), it is clear that $\forall x(x \in a \Leftrightarrow x \in b)$ comes out valid, so the extensionality rule would require $\forall y(a \in y \Leftrightarrow b \in y)$ to come out valid, which in turn requires $a \in c \Leftrightarrow b \in c$ to come out valid. But it does not.

    For instance, in the revision construction of Field [12], $|b \in c|_\beta$ is 1 whenever $\beta > 0$; which means that for $\alpha > 1$, $|a \in c \Leftrightarrow b \in c|_\alpha$ is 1 if and only if there is an interval prior to α throughout which $|a \in c|_\beta$ is 1. But by choice of c, this just means that for $\alpha > 1$, $|a \in c|_\alpha$ is 1 only if there is an interval prior to α throughout which $|a \in c|_\beta$ is 1. So unless $|a \in c|_\alpha$ is 1 for $\alpha = 1$, then by induction it is not 1 for any bigger α either. But it is not 1 for $\alpha = 1$. To see this, let d be $\{x : (\neg\exists y)(y \in x)\}$. Then $|a \in d|_0 = 1$; but since $|\top \Rightarrow \bot|_0 = 1/2$, $|b \in d|_0 = 1/2$, with the result that $|a \in c|_1 < 1$.

    The reader may verify that the same example (replacing $\Rightarrow$ above with the $\triangleright$ or $\blacktriangleright$ of that paper) is a counterexample to the rule form of extensionality for the "fibers" of Field [13]. Theorem 10.1 will cast further light on this: both logics have attractive features that turn out to preclude extensionality (in the presence of abstraction). Thus the "substantial modification of the constructions" mentioned in the main text would have to result in a significant change of the logic, if the logic was to be used for naive classes. (Naive property theory works fine for both logics as they stand.)

18. Brady [8] does consider an alternative v_0, but not with the intent of achieving weakening. (Indeed, in the case of his M2 logics, our result in Section 10 will show that whatever the corrected version of his construction validates, if it preserves abstraction and extensionality, it will not yield weakening, since his conditionals are contraposable and validate Contra-Modus-Ponens.)

19. It might seem more in the spirit of Brady to take $\Rightarrow$ instead of (or together with) $\rightarrow$ as primitive. But since we do not think the purely positive logic (whether developed with $\Rightarrow$, or $\rightarrow$) is promising, and since our main interest is in laying the groundwork for the 4-valued model in the next section, we will not pursue this approach here.

20. Because of this, we can think of the Brady semantics for the positive logic as assigning to every sentence an ordinal value: $|||A|||$ is the first α for which $|A|_\alpha$ is 0 if there is such an α, and otherwise $|||A|||$ is the fixed-point ordinal $\Omega_{\boldsymbol{M}}$ of the negation-free construction for ground model M. On this semantics, $\Omega_{\boldsymbol{M}}$ is the designated value. (Only sentences with single-bar value 1 at the fixed point get triple-bar value $\Omega_{\boldsymbol{M}}$, because no sentence A can first get (single-bar) value 0 at the fixed-point ordinal: if A first got value 0 at $\Omega_{\boldsymbol{M}}$, then $\top \rightarrow A$ would first get value 0 at $\Omega_{\boldsymbol{M}} + 1$, which is impossible given that $\Omega_{\boldsymbol{M}}$ is a fixed point.)

21. On adding a truth predicate, see footnote 8 above.

22. Still not put perfectly precisely, but we assume clear enough.

23. It is easy to verify that if $(\exists\beta < \mu)(|A|_\beta = 0)$, then $|\neg_{NI^\mu} A|_\alpha$ is 1 for every α, and hence $\neg_{NI^\mu} A$ is equivalent to $\top$; whereas if $(\forall\beta < \mu)(|A|_\beta = 1)$, then $|\neg_{NI^\mu} A|_\alpha$ is 1 for $\alpha < \mu$ and 0 for $\alpha \geq \mu$. (Or in the terminology of footnote 20: If $|||A||| < \mu$, then $|||\neg_{NI^\mu} A||| = \Omega$, and if $|||A||| \geq \mu$, then $|||\neg_{NI^\mu} A||| = \mu$. As μ increases, there are more sentences A for which $|||\neg_{NI^\mu} A||| = \Omega$, and the others get strictly higher value than for smaller μ.) So every sentence in the language is equivalent either to $\top$ or to some $\neg_{NI^\mu}\top$, with $\bot$ equivalent to $\neg_{NI^0}\top$ but also to $\neg_{NI^0}\bot$. It is easy to see that the

system of notation must stop prior to Ω. None of the $\neg_{NI^\mu}$'s lead to paradox because $\neg_{NI^\mu}(\neg_{NI^\mu}\top)$ is equivalent to $\neg_{NI^\mu}\top$: its value is 1 for $\alpha < \mu$, 0 for $\alpha \leq \mu$ (i.e., its triple-bar value is μ).

24. This is not quite accurate, because the notation system need not be restricted to standard Church–Kleene notations; it can be taken to include ones that use a non-bivalent membership predicate, and talk of "the first" F can be problematic in a non-bivalent context. (See [12, Chapter 22] for a treatment of related issues in the context of naive truth and satisfaction.)

25. In the "triple-bar" representation of previous footnotes, every sentence is assigned a pair of ordinals, each at most $\Omega_{\boldsymbol{M}}$, summarizing the values at every stage of the macro-construction.

26. Note that $|X|^+_\alpha = 1$ whenever $(\forall\beta < \alpha)(|X|^-_\beta = 1)$; on the static, $|X|^-_\alpha = 1$ iff $(\forall\beta < \alpha)(|X|^+_\beta = 1)$, and on any of the dynamic proposals, $|X|^-_\alpha = 1$ iff $|X|^+_\alpha = 1$.

27. We think even dialetheists should accept Contra-Modus-Ponens. For instance, if A is a dialetheia, we think $A \rightarrow \neg A$ and $A \rightarrow A$ should be too. But we need not decide this: the point of switching to (*II*) was to avoid dialetheia, so the proponent of this switch will not try to use dialetheia to explain away the resultant failure of Contra-Modus-Ponens.

28. An alternative which does validate excluded middle without restriction can be derived from Brady [6]. There, Brady gives a construction for a dialetheic logic, which differs from his other constructions in three respects. First, all conditionals start by taking value 1/2. Second, he uses the following clauses: $|A \Rightarrow B|_\alpha = 1$ iff $(\forall\beta < \alpha)$(if $|A|_\beta \geq 1/2$, $|B|_\beta = 1$) and $|A \Rightarrow B|_\alpha = 0$ iff $(\exists\beta < \alpha)(|A|_\beta \geq 1/2$ and $|B|_\beta = 0)$, which, as our notation suggests, yields contraposition. Third, and finally, he takes the designated values to be both 1/2 and 1. The construction delivers the logic *DJ* plus Contra-Modus-Ponens. Øgaard [25] notes that in this case we can define a noncontraposable conditional $\mapsto$ which has both weakening and no restrictions on excluded middle, by letting $A \mapsto B = (A \wedge \lambda) \Rightarrow B$, where λ is the Russell sentence $\{x : \neg x \in x\} \in \{x : \neg x \in x\}$. Here, the noncontraposable conditional is defined from the contraposable as opposed to the other way round, but the logic of each is in some respects as good as what we have in the main text. A main loss is that the new conditional does not satisfy suffixing or prefixing (see Section 10).

29. A similar point holds for dialetheists who think we do not need a conditional to express universal restricted quantification but can take that as a primitive quantifier $(\forall x \ni Ax)Bx$ (read as "every A is B"). (When Ax has the form $x \in t$, we will abbreviate $(\forall x \ni Ax)Bx$ as $(\forall x \in t)Bx$; similarly for $\notin$.) For the present construction could be redone in those terms, showing that we can get abstraction and extensionality in the forms

Abstraction:

$$\begin{aligned}\forall u_1, \ldots, \forall u_n\big[&\big(\forall z \in \{x : A(x; u_1, \ldots, u_n)\}\big)A(z; u_1, \ldots, u_n)) \\ &\wedge \big(\forall z \ni A(z; u_1, \ldots, u_n)\big)\big(z \in \{x : A(x; u_1, \ldots, u_n)\}\big) \\ &\wedge \big(\forall z \notin \{x : A(x; u_1, \ldots, u_n)\}\big)\neg A(z; u_1, \ldots, u_n)) \\ &\wedge \big(\forall z \ni \neg A(z; u_1, \ldots, u_n)\big)\big(z \notin \{x : A(x; u_1, \ldots, u_n)\}\big)\big];\end{aligned}$$

Extensionality Rule:

$$\begin{aligned}&\mathrm{Class}(a) \wedge \mathrm{Class}(b) \wedge (\forall u \in a)(u \in b) \wedge (\forall u \in b)(u \in a)\\ &\quad \wedge (\forall u \notin a)(u \notin b) \wedge (\forall u \notin b)(u \notin a)\\ &\quad \vDash \big(\forall z \ni (a \in z)\big)(b \in z) \wedge \big(\forall z \ni (b \in z)\big)(a \in z)\\ &\quad \wedge \big(\forall z \ni (a \notin z)\big)(b \notin z) \wedge \big(\forall z \ni (b \notin z)\big)(a \notin z);\end{aligned}$$

while also having

RQ-Weakening: $\forall x B x \vDash (\forall x \ni A x) B x$.

30. We could have run the same argument using Negative RQ-Weakening, that is, $\neg\exists x A(x) \vDash$ All A are B, instead of Negative Conjunctive RQ-Weakening.

31. Perhaps it is worth noting that the a and b for which the Intuitive Extensionality claim fails are ones for which the defining formulas are not intuitively dialetheias; so there is no escape from the argument by confining the Intuitive Extensionality claim to classes not defined from dialetheias. (Such an attempted escape would in any case require a huge retreat. Someone who is dialetheic about the Russell class will think that it both is and is not universal; if they want extensionality, they should want the Russell class to both be and not be identical to $\{x : \top\}$, and that there both are and are not classes that contain $\{x : \top\}$ but do not contain the Russell class (or conversely). That involves accepting the Intuitive Extensionality claim as stated. A dialetheist who exempts classes defined from dialetheias from "counting" against extensionality and related goals could no more claim success at a naive theory of classes than could an advocate of restrictions on excluded middle who exempted classes defined by non-bivalent formulas from the purview of extensionality.)

32. The principle also holds for $=_{\mathrm{ALT}}$ when as we have been assuming identity in the ground language is classical (i.e., determinate), though it would not hold in a more general setting. However, even there the modified principle

$$\mathrm{Class}(a), \mathrm{Class}(b), A(a) \vdash a =_{\mathrm{ALT}} b \rightarrow \big(\top \rightarrow A(b)\big)$$

would hold; and since the main proof below will only use a, b such that $\vdash \mathrm{Class}(a) \wedge \mathrm{Class}(b)$, this modified principle would suffice for the proof.

33. It is worth noting that the proof targets not merely extensionality, but the extensionality metarule: it thus shows that even theories of naive properties that accept the assumptions of Theorem 10.1 had better not individuate properties so coarsely that if two properties are defined by sentences that are provably coextensive in the full theory of properties, they are identical. For that would keep the indiscernibility of identicals from being valid. (For instance, in the theory of Field [12, Section 16.2], let a be the property $\{x : \bot\}$, and let b be the property $\{x : \top \rightarrow \bot\}$; then a and b are properties defined by logically equivalent sentences, but they should not be declared identical because b definitely instantiates the property $\{x : \forall y(x \in y \leftrightarrow b \in y)\}$ whereas a does not; see endnote 17.) But that does not seem at all devastating for naive properties, for there are independent reasons for holding that properties should be individuated more finely than this. By contrast, the result cuts to the heart of the naive theory of classes.

34. We should note, however, that the argument against dialetheism at the end of Section 9 would remain even if abstraction were weakened to comprehension: for all of the abstracts employed in the argument were based on formulas that did not themselves

contain abstracts, and would thus be available in the theory from comprehension alone by conservative extension.

35. See [14] for an example where a form of abstraction *restricted to positive formulas* is inconsistent with extensionality, but the similarly restricted form of comprehension is in fact consistent with it.

References

[1] Bacon, A., "A new conditional for naive truth theory," *Notre Dame Journal of Formal Logic*, vol. 54 (2013), pp. 87–104. Zbl 1273.03073. MR 3007964. DOI 10.1215/00294527-1731407. 463, 482, 493, 496

[2] Bacon, A., "Paradoxes of logical equivalence and identity," *Topoi*, vol. 34 (2015), pp. 89–98. MR 3316732. 500

[3] Beall, J., R. T. Brady, A. P. Hazen, G. Priest, and G. Restall, "Relevant restricted quantification," *Journal of Philosophical Logic*, vol. 35 (2006), pp. 587–98. Zbl 1111.03020. MR 2252731. 478, 486

[4] Brady, R. T., "The consistency of the axioms of abstraction and extensionality in a three-valued logic," *Notre Dame Journal of Formal Logic*, vol. 12 (1971), pp. 447–53. Zbl 0185.01402. MR 0297535. 497

[5] Brady, R. T., "The simple consistency of a set theory based on the logic CSQ," *Notre Dame Journal of Formal Logic*, vol. 24 (1983), pp. 431–49. Zbl 0488.03026. MR 0717905. DOI 10.1305/ndjfl/1093870447. 462, 471, 472, 473, 476, 493

[6] Brady, R. T., "The non-triviality of dialectical set theory," pp. 437–70 in *Paraconsistent Logic: Essays on the Inconsistent*, edited by G. Priest, R. Routley, and J. Norman, Philosophia, Münich, 1989. Zbl 0691.03037. 499, 502

[7] Brady, R. T., *Universal Logic*, vol. 109 of *CSLI Lecture Notes*, CSLI Publications, Stanford, 2006. Zbl 1234.03010. MR 2440514. 471, 472, 473, 476, 493

[8] Brady, R. T., "The simple consistency of naive set theory using metavaluations," *Journal of Philosophical Logic*, vol. 43 (2014), pp. 261–81. MR 3218862. DOI 10.1007/s10992-012-9262-2. 462, 499, 501

[9] Brady, R. T., and R. Routley, "The non-triviality of extensional dialectical set theory," pp. 415–37 in *Paraconsistent Logic: Essays on the Inconsistent*, edited by G. Priest, R. Routley, and J. Norman, Philosophia, Münich, 1989. Zbl 0691.03037. MR 0782845. DOI 10.1007/BF00935736. 497

[10] Cantini, A., "The undecidability of Grišin's set theory," *Studia Logica*, vol. 74 (2003), pp. 345–68. Zbl 1039.03040. MR 1996.834. DOI 10.1023/A:1025159016268. 497

[11] Evans, G., "Can there be vague objects?," *Analysis*, vol. 38 (1978), p. 208. DOI 10.1093/analys/38.4.208. 500

[12] Field, H. H., *Saving Truth from Paradox*, Oxford University Press, Oxford, 2008. Zbl 1225.03006. MR 2723032. 478, 501, 502, 503

[13] Field, H. H., "Naive truth and restricted quantification: Saving truth a whole lot better," *Review of Symbolic Logic*, vol. 7 (2014), pp. 147–91. Zbl 1329.03015. MR 3244967. DOI 10.1017/S1755020313000312. 478, 501

[14] Forti, M., and R. Hinnion, "The consistency problem for positive comprehension principles," *Journal of Symbolic Logic*, vol. 54 (1989), pp. 1401–18. Zbl 0702.03026. MR 1026606. 497, 504

[15] Friedman, H., "The consistency of classical set theory relative to a set theory with intuitionistic logic," *Journal of Symbolic Logic*, vol. 38 (1973), pp. 315–19. Zbl 0278.02045. MR 0347565. DOI 10.2307/2272068. 497

[16] Gilmore, P. C., "The consistency of a positive set theory," *Technical Report RC-1754*,

IBM Research Report, 24 January 1967. 466, 497

[17] Gilmore, P. C., "The consistency of partial set theory without extensionality," pp. 147–53 in *Axiomatic Set Theory (Los Angeles, 1967)*, American Mathematical Society, Providence, 1974. Zbl 0309.02065. MR 0360271. 466

[18] Grišin, V. N., "Predicate and set-theoretic calculi based on logic without the contraction rules," *Mathematics of the USSSR-Izvestiya*, vol. 45 (1981), pp. 47–68. Zbl 0464.03027. MR 0607576. 497

[19] Hamkins, J., "Is there any research on set theory without extensionality axiom?" preprint, http://mathoverflow.net/questions/168287/is-there-any-research-on-set-theory-without-extensionality-axiom. 498

[20] Hinnion, R., "Naive set theory with extensionality in partial logic and in paradoxical logic," *Notre Dame Journal of Formal Logic*, vol. 35 (1994), pp. 15–40. Zbl 0801.03019. MR 1271696. DOI 10.1305/ndjfl/1040609292. 497

[21] Hinnion, R., and T. Libert, "Positive abstraction and extensionality," *Journal of Symbolic Logic*, vol. 68 (2003), pp. 828–36. Zbl 1056.03027. MR 2000080. DOI 10.2178/jsl/1058448441. 497

[22] Hinnion, R., and T. Libert, "Topological models for extensional partial set theory," *Notre Dame Journal of Formal Logic*, vol. 49 (2008), pp. 39–53. Zbl 1180.03047. MR 2376779. 497

[23] Kripke, S. A., "Outline of a theory of truth," *Journal of Philosophy*, vol. 72 (1975), pp. 690–716. Zbl 0952.03513. 466, 467

[24] Maddy, P., "Proper classes," *Journal of Symbolic Logic*, vol. 48 (1983), pp. 113–39. Zbl 0546.03008. MR 0693255. DOI 10.2307/2273327. 497

[25] Øgaard, T. F., "Paths to triviality," *Journal of Philosophical Logic*, vol. 45 (2016), pp. 237–76. Zbl 06599320. MR 3500972. 497, 499, 500, 502

[26] Scott, D., "More on the axiom of extensionality," pp. 115–31 in *Essays of the Foundations of Mathematics*, Magnes Press, Hebrew University, Jerusalem, 1961. Zbl 0199.01403. MR 0163838. 497

[27] Shirahata, M., "Linear set theory with strict comprehension," pp. 223–45 in *Proceedings of the Sixth Asian Logic Conference (Beijing, 1996)*, edited by C. T. Chong, World Scientific, River Edge, N.J., 1998. Zbl 0990.03038. MR 1789739. 497

[28] White, R. B., "The consistency of the axiom of comprehension in the infinite-valued predicate logic of Łukasiewicz," *Journal of Philosophical Logic*, vol. 8 (1979), pp. 509–34. Zbl 0418.03037. MR 0551284. DOI 10.1007/BF00258447. 497

Acknowledgments

Field and Lederman contributed equally to the first draft of this paper, containing all of Sections 2–9 and Appendix B in essentially their present form. After reading the earlier draft, Øgaard discovered the impossibility result of Section 10 and recognized that incorporating this result required taking abstraction (not comprehension) as primitive throughout the paper. Field and Lederman then jointly wrote Section 10 and rewrote Section 1. All three authors contributed equally to Appendix A. Thanks to an anonymous reviewer for pressing us to clarify a number of points of presentation and substance.

Field
Philosophy Department
New York University
New York, New York
USA
hartry.field@nyu.edu

Lederman
Department of Philosophy
University of Pittsburgh
Pittsburgh, Pennsylvania
USA
harveyslederman@gmail.com

Øgaard
Department of Linguistic and Scandinavian Studies
University of Oslo
Oslo
Norway
toreog@gmail.com