Saving Truth from Paradox

HARTRY FIELD

OXFORD
UNIVERSITY PRESS

Contents

PART III. PARACOMPLETENESS

PART IV. MORE ON PARACOMPLETE SOLUTIONS

Introduction

1. GRELLING'S PARADOX

The predicate 'red' is true of all and only the red things; and the predicate 'has mass greater than 10 kilograms' is true of all and only things that have mass greater than 10 kilograms. More generally, for any (simple or complex) predicate of our language that one substitutes for the letter '*F*', the following holds:

(TO) '*F*' is true of all and only the things that are *F*.

Or so it seems. But consider the predicate 'is not true of itself'. (TO) yields

'Is not true of itself' is true of all and only those things that are not true of themselves.

But this is inconsistent according to standard (i.e. classical) logic. For it yields

'Is not true of itself' is true of 'is not true of itself' if and only if 'is not true of itself' is not true of itself;

which is to say,

'Is not true of itself' is true of itself if and only if it is not true of itself.

This has the form

B if and only if not-*B*,

which is contradictory according to standard logic.

This is often called Grelling's paradox, or the heterologicality paradox. (Its inventor Grelling used the term 'heterological' as an abbreviation of 'is not true of itself'.)

2. RUSSELL'S PARADOX FOR PROPERTIES

Grelling's paradox concerns linguistic expressions, but it is arguable that underlying it is a more basic paradox concerning properties, one that is in essence due to Russell. The property of being red is instantiated by all and only the red things; the property of having mass greater than 10 kilograms is instantiated by all and only things that have mass greater than 10 kilograms. More generally, for any (simple or complex) intelligible predicate of our language that one substitutes for the letter 'F', the following would seem to hold:

(INST) The property of being F is instantiated by all and only the things that are F.

But again, there's a problem: consider the "Russell property", the property of not instantiating itself. (INST) yields

The Russell property is instantiated by all and only those things that don't instantiate themselves.

In particular,

The Russell property instantiates itself if and only if it doesn't instantiate itself.

Again this has the classically contradictory form

B if and only if not-B.

3. . . . *v*. RUSSELL'S PARADOX FOR SETS

When writers today speak of Russell's paradox they usually mean not quite the paradox above, but a paradox involving the mathematical notion of set. The paradox is analogous, with 'is a member of' instead of 'instantiates': it seems that there is a set whose members are precisely those things that are not members of themselves; but then in anology with the above we argue that this set is a member of itself if and only if it isn't, which is classically contradictory.

Russell's paradox for sets has a standard solution: there simply is no such set. This is not an *ad hoc* solution, for there is a natural picture of sets according to which it is inevitable. On this picture, the sets form a hierarchy, the rank hierarchy. Putting aside the empty set for simplicity, we have at the bottom layer the "rank 1" sets, whose members aren't sets (but things like apples, people, and particles). At the next layer we have "rank 2" sets, all of whose members are either non-sets or sets of rank 1, with at least one of the latter. (The last condition is to ensure that no rank 2 set is also rank 1.) In general, the sets of each layer have members only of preceding layers (and for each preceding layer, they have a member of that layer or higher). This ensures (i) that no set can be a member of itself. That in turn ensures (ii) that there can be no universal set, i.e. no set of which everything is a member; for then it would have to be a member of itself, which we've just seen is ruled out. Finally, (iii) there can be no Russell set, i.e. no set of all things that are not members of themselves; for by (i) it would have to be a universal set, but by (ii) there is no universal set.

This solution to the Russell paradox is motivated by the hierarchical picture of sets. And that picture is not only natural, it is mathematically fruitful: it provides the framework for virtually all of contemporary mathematics. Indeed, there is reason to think that this was the picture underlying set theory from the beginning (in Cantor's work), even prior to the discovery of Russell's paradox.[1] The notion of set is primarily a mathematical notion, so it would be hard to criticize this solution to the paradox.

Kurt Gödel is reported to have remarked "There never were any set-theoretic paradoxes, but the *property-theoretic* paradoxes are still unresolved."[2] The idea behind the first part of his remark is presumably that the notion of set was hierarchical from the start, so that it should have been obvious all along that there was no Russell set (in the mathematical sense of 'set'). The idea behind the second part is that this obvious resolution of Russell's "paradox" for sets simply doesn't carry over to the paradox for properties.

And this seems correct, at least for the notion of property that Gödel had in mind. I make that qualification because the term 'property' is used in more than one way. One use of the term is for *natural properties*. On this use, what properties there are is something discovered in part by science: e.g. science tells us that there is no such property as being de-phlogisticated, but that there are properties like spin and charm. But Gödel presumably had in mind a quite different use of the term 'property'. Properties in this second sense are sometimes called *concepts*, but that term too is ambiguous: it is often used for mental entities, whereas what is intended here is an objective correlate of mental entities. I'll instead use the term *conceptual property*. The whole point of the notion of conceptual property is that there is a conceptual property corresponding to every intelligible predicate. 'Identical' is intelligible, so there is a 2-place conceptual property of *being identical to*. (That is, a conceptual *relation*: relations are multi-place properties.) From this we can form a one-place conceptual property of *being identical to itself*; this is a universal property, so the hierarchical picture simply can't apply to conceptual properties. Similarly, 'instantiates' is intelligible, so there is a conceptual relation of instantiation, and a conceptual property of instantiating itself. And there is another conceptual property of *not* instantiating itself, which is the Russell property. If this is right, then the standard solution to Russell's paradox for sets simply doesn't carry over to the paradox for conceptual properties.[3]

(By the standard solution to Russell's paradox for sets I mean the solution embodied in set theories such as Zermelo-Fraenkel. Russell himself offered a slightly different solution, the theory of types, in which there is no single membership predicate, but a whole hierarchy of different ones. Similarly, one might offer a resolution to the property-theoretic analog according to which there is no single instantiation predicate, but a whole hierarchy of different ones.

[1] See Lavine 1994, pp. 76 ff. [2] See the opening sentence of Myhill 1984.

[3] The paradox Russell himself had in mind was probably closer to the property paradox than to the paradox for mathematical sets. Russell originally raised the paradox in connection with Frege's theory of extensions. Extensions were supposed to be like mathematical sets in obeying an axiom of extensionality: no two extensions were instantiated by exactly the same things. But they were to be like conceptual properties in that every predicate had one, and that they didn't fall into a hierarchy of ranks.

I will consider theories like this in due course,[4] but for now I merely say that they have been almost universally abandoned for sets, and are unattractive for properties for the same reasons and more.)

One possible reaction to this is that we should, with Quine (1960, sec. 43), simply declare there to be no such things as conceptual properties. That's certainly a possible view (for which there may be some motivation independent of the paradoxes), but it is not ultimately of much help in the present context since the apparently analogous heterologicality paradox, which involves only linguistic expressions, would still remain. Another possible reaction is that our common-sense conception of conceptual properties should be replaced with a hierarchical conception; though that would be a very radical alteration of the common-sense conception. Fortunately there are other possibilities.

4. SOLUTION ROUTES TO RUSSELL'S PARADOX FOR PROPERTIES

We've seen that one possible reaction to Russell's paradox for (conceptual) properties is

Non-existence Solution: There is no such thing as "the Russell property", i.e. "the (conceptual) property of not instantiating itself". (Put loosely: There is no (conceptual) property corresponding to the predicate 'doesn't instantiate itself'. I take this as just an informal way of putting what's in the first sentence. Note that talk of "correspondence" doesn't appear in the official formulation.)

The idea is to decompose the schema (INST) into two components: a correct component

($INST_w$) *If* there is a property of being F, *then* it is instantiated by all and only the things that are F;

and a component that needs restriction

(COMP) There is a property of being F.

According to the Non-existence Solution, ($INST_w$) holds for all intelligible predicates F, but (COMP) fails for some of them, e.g. 'doesn't instantiate itself'. There will need to be many other failures of (COMP) too, if other paradoxes like Russell's are to be handled along similar lines. A serious theory that incorporates the Non-Existence Solution will have to tell us for which F's (COMP) holds and for which ones it fails.

I've suggested, following Gödel, that the Non-Existence solution isn't very attractive (for anyone who doesn't just reject conceptual properties out of hand).

[4] At least, I will consider analogous theories of truth-of, and will suggest a general method for getting from any theory of truth-of to a corresponding theory of properties.

Conceptual properties aren't like sets, and it would violate their *raison d'etre* to suppose that for an intelligible predicate like 'doesn't instantiate itself', there is no corresponding property of *not instantiating itself*.

This is doubtless too quick, for one possibility would be to grant that every *intelligible* predicate has a corresponding property, but to deny that 'doesn't instantiate itself' is intelligible. On the most obvious version of this, we maintain the original (INST) (and hence (COMP)), for any intelligible predicate F. But there's something very odd about holding that 'doesn't instantiate itself' isn't intelligible. It can be brought out by asking "What part of 'doesn't instantiate itself' don't you understand?" It seems that if you accept (INST) (restricted to intelligible predicates), then you must regard the predicate 'instantiates' as intelligible, since it is used in the formulation of (INST)! Presumably 'not' and 'itself' are intelligible too, so how can 'doesn't instantiate itself' fail to be intelligible? 'Intelligible' would have to be given a very special reading for this to make any sense. Of course we can always give it the special reading 'expresses a property', but (COMP) restricted to predicates that are "intelligible" in *this* sense becomes totally vacuous. We would in effect just be granting that (COMP) and the original (INST) fail for some predicates that are intelligible by ordinary standards. The pretense that we've *explained* the failures of (COMP) as due to the unintelligibility of the predicates is exposed as a fraud, and we still need an account of which predicates it is for which (COMP) fails.

There is a more interesting version of the view that 'instantiates' is not an intelligible predicate: the **Stratification View**. According to it, we should stop using the predicate 'instantiates', and instead use a hierarchy of more restrictive predicates 'instantiates$_\alpha$', where the subscript is a notation for an ordinal number. This is a radical view: since (INST) and (INST_w) involve the "unintelligible" 'instantiates', the view requires that we abandon these principles and replace them by principles that employ only the subscripted predicates. I will eventually be considering this view (or a linguistic analog of it: see note 4), but I would like to defer it for the time being.

Let us for now put the Stratification View aside, and assume that 'instantiates' is a perfectly intelligible predicate (as are predicates built up from it using ordinary logical devices). And let us also put aside the Non-Existence Solution: let us assume that there is a property *doesn't instantiate itself* (the Russell property, which I'll call R). Then (INST_w) allows us to argue to the classical contradiction

(*) R instantiates R if and only if R does not instantiate R.

When faced with this, one possibility would be to turn this argument around to argue against (INST_w). More specifically, either (i) R instantiates R or (ii) R doesn't instantiate R. Since R is the property of not instantiating itself, case (i) has it that there is an object (viz. R) which instantiates itself and yet instantiates the property of *not* instantiating itself. This conclusion violates (INST_w). It also seems highly surprising. Case (ii) is equally surprising: there is an object (viz. R)

which doesn't instantiate itself, but yet doesn't instantiate the property of not instantiating itself. If something isn't a dog, it's hard to see how it can fail to instantiate the property of not being a dog; similarly, if something doesn't instantiate itself, it's hard to see how it can fail to instantiate the property of not instantiating itself. So (i) and (ii) are each counterintuitive; and yet, it follows from classical logic (together with our property-existence assumption) that one of these violations of ($INST_w$) must occur.

So if we put aside the Non-Existence Solution and its variants, classical logic would seem to commit us to one of two solutions; these are

Overspill Solution: For some predicates F, we have: there are objects that aren't F but have the property of being F;

and

Underspill Solution: For some predicates F, we have: there are objects that are F but don't have the property of being F.

Just as there are really many versions of the Non-Existence Solution (differing over *which* predicates have no corresponding properties), so there are many versions of the Underspill and Overspill Solutions: theories of property-instantiation can vary as to how much underspill or overspill there is. At the extreme of Underspill, we could have that no property is ever instantiated by anything: there are red things, and a property of redness, but the red things don't instantiate that property! Similarly, at the extreme of Overspill, we could have that every property is instantiated by everything: the property of redness is instantiated by things that aren't red! Of course no one would go to either extreme, but to get an Underspill or Overspill solution to the paradox one needs to specify precisely how far toward the extreme one is to go.

Actually Underspill and Overspill aren't mutually exclusive: a given theory could have some predicates that underspill and some that overspill. (In fact, if one supposes that there is underspill for one predicate, it is at least somewhat natural to suppose that there is overspill for its negation, and vice versa.)[5] Indeed, a given theory can have a single predicate that both underspills and overspills: it could postulate both an object c_1 that is F but doesn't have the property of being F, and another object c_2 that has the property of being F without being F. But assuming classical logic, it seems we must have either Underspill or Overspill, if there is a property of not instantiating itself.

So unless we follow the Stratified View in abandoning the notion of instantiation, then classical logic commits us to at least one of these three options: either there is no such property as "the property of not instantiating itself"; or there is one, but it's instantiated by some things that "shouldn't" instantiate it, i.e. some

[5] It would also be possible to combine property non-existence for some predicates with underspill and/or overspill for others, but I doubt that that route would have much appeal.

things that do instantiate themselves; or there is one, but it isn't instantiated by some things that "should" instantiate it, i.e. things that don't instantiate themselves. None of these seem terribly attractive. Even in the Stratified View, on which there is a variety of restricted instantiation predicates, these three options are the only ones available for any one of them.[6]

Might the problem be the assumption that classical logic applies to the instantiation predicate? That is, might there be a generalization of classical logic that takes the classical rules to be appropriate for dealing with "ordinary" predicates (such as those of standard mathematics and physics) but which allows only weaker rules when dealing with certain "extraordinary" predicates like 'instantiates'? (One obvious rationale for counting this predicate as extraordinary is that it exhibits a kind of circularity: e.g. if U is the predicate 'instantiates itself', then to find out if U instantiates itself it seems *prima facie* that you must first find out whether it instantiates itself.) The most natural way to try to implement this idea of weakening the logic is to argue that (*) isn't really contradictory. In that case, we might be able to simultaneously accept the "naive theory" that there is a property corresponding to every predicate and that such properties satisfy (INST). But is there any plausibility to the claim that (*) isn't really contradictory?

Here is a simple argument that claims of the form $B \leftrightarrow \neg B$ are always contradictory, even for sentences B that contain extraordinary predicates. (Here '$\leftrightarrow$' abbreviates 'if and only if' and '$\neg$' abbreviates 'not'.) For later reference I'll call it the *Central Argument from Equivalence to Contradiction*.

Step One: The claim $B \leftrightarrow \neg B$ and the claim B together imply the claim $B \wedge \neg B$ (where '$\wedge$' means 'and').

This step seems obvious: the two premises $B \leftrightarrow \neg B$ and B together certainly imply $\neg B$, and the second premise by itself trivially implies B, so the two together imply the conjunction $B \wedge \neg B$.

Step Two: The claim $B \leftrightarrow \neg B$ and the claim $\neg B$ together imply the claim $B \wedge \neg B$.

This is similar to Step One: $B \leftrightarrow \neg B$ and $\neg B$ certainly imply B, and the second alone implies $\neg B$, so they imply the conjunction $B \wedge \neg B$.

Step Three: The claim $B \leftrightarrow \neg B$ and the claim $B \vee \neg B$ together imply the claim $B \wedge \neg B$ (where '$\vee$' means 'or').

This follows from Steps One and Two, by the rule of Reasoning by Cases: if assumptions Γ plus A_1 imply C, and those same assumptions Γ plus A_2 imply the same conclusion C, then Γ plus $A_1 \vee A_2$ imply C. And the rule of Reasoning by Cases would seem to be pretty central to the meaning of 'or'.

[6] Stratified theories typically adopt the Underspill Option for each instantiation predicate, and argue that the underspill at any one level is made up for at later levels. We will consider this in detail in Chapter 14 (though in the context of truth rather than instantiation).

Step Four: The claim $B \leftrightarrow \neg B$ itself implies $B \wedge \neg B$.

Step Four follows from Step Three, if we assume that $B \vee \neg B$ is a logical truth. That assumption is famously called the law of excluded middle (LEM). It is not only famous, it is famously controversial. There are some (e.g. mathematical intuitionists like the Dutch mathematicians Brouwer and Heyting) who deny its general applicability even within mathematics; indeed some intuitionists (e.g. Michael Dummett) would deny its applicability to any statement that is not "in principle verifiable". But one needn't fall victim to "Dutch LEM disease" to suspect that there might be something suspicious about Step Four: questioning Step Four needn't involve questioning excluded middle in mathematics, physics, etc. but only questioning it in connection with certain applications of "circular" predicates like 'instantiates'.[7]

Without Step Four, it isn't immediately obvious why $B \leftrightarrow \neg B$ should be regarded as contradictory. This suggests the possibility of a fourth solution route:

Paracomplete Solutions: Excluded middle is not generally valid for sentences involving 'instantiates'. In particular, the assumption that the Russell property *R either instantiates itself or doesn't* should be rejected. Indeed, the reasoning of the "paradox" shows that that assumption leads to contradiction.

For this to be an interesting option it must preclude the need for the restrictions on (INST) and (COMP); that is, an interesting Paracomplete Theory must be one in which Naive Property Theory is consistent (where Naive Property Theory is the theory that for every predicate, there is a corresponding property that satisfies the "Instantiation Schema" (INST)). It is far from evident that an interesting Paracomplete Theory meeting this requirement is possible.

Indeed, a precondition of its possibility is that the logic vitiate *all* arguments from $B \leftrightarrow \neg B$ to contradiction, not just the Central Argument; and some logics without excluded middle, such as intuitionist logic, invalidate the Central Argument for the contradictoriness of $B \leftrightarrow \neg B$ while leaving other arguments intact. The most obvious route from $B \leftrightarrow \neg B$ to $B \wedge \neg B$ within intuitionism comes from the intuitionist *reductio* rule, which says that if $\neg B$ follows from Γ and B together then it follows from Γ alone. Although intuitionists have reasons of their own for accepting this rule, the most obvious arguments for accepting it assume excluded middle. For instance: $\neg B$ certainly follows from Γ and $\neg B$ together, so if it also follows from Γ and B together then it must follow from Γ and $B \vee \neg B$ together, and hence *assuming excluded middle* it follows from Γ alone.[8]

[7] In Chapter 5 I will entertain the possibility of a further bar on excluded middle, to prevent its application to sentences that are crucially vague. But even this further restriction (on which I take no firm stand) would presumably leave the application to mathematics untouched, and wouldn't affect the application to physics in a disturbing way.

[8] An alternative route from $B \leftrightarrow \neg B$ to $B \wedge \neg B$ within intuitionism comes from the so-called "law of non-contradiction", $\neg(B \wedge \neg B)$. Given one of the deMorgan laws, this is equivalent to

There is a class of logics without excluded middle ("deMorgan logics") that don't contain the *reductio* rule and that are in many respects far more natural than intuitionist logic: for instance, intuitionist logic restricts the inference from $\neg\neg A$ to A, and also restricts one of the deMorgan laws (viz., the inference from $\neg(A \wedge B)$ to $\neg A \vee \neg B$), whereas deMorgan logics maintain all the deMorgan laws plus the equivalence of $\neg\neg A$ to A. And in deMorgan logics without excluded middle, $B \leftrightarrow \neg B$ is not contradictory.

A logic in which $B \leftrightarrow \neg B$ is not contradictory is *necessary* for a paracomplete solution, but far from *sufficient*: it remains to be shown that it is possible to consistently maintain the Naive Property Theory ((COMP) and (INST)) in such a logic. Indeed, we really want Naive Property Theory to include a bit more than (COMP) together with the Instantiation Schema (INST); we want it to include an intersubstitutivity claim, according to which

(i) *o* instantiates the property of being *F*

is fully equivalent to

(ii) *Fo*

in the sense that the claims (i) and (ii) can be intersubstituted even in embedded contexts (so long as these contexts are "transparent", that is, don't involve quotation marks, intentional operators, or the like). For instance, we want a principle that guarantees such things as

[It is not the case that *o* instantiates the property of being *F*] if and only if [it is not the case that *Fo*];

[If *o* instantiates the property of being *F* then *B*] if and only if [if *Fo* then *B*];

and so forth. Don't these follow from (INST)? They do in classical logic, but we're considering weakening classical logic. Still, I don't myself think a paracomplete solution would be very satisfactory if it required such a weak logic that these no longer followed.

So a good paracomplete solution would involve showing that (COMP) and (INST) can be maintained in a *reasonably strong* paracomplete logic that allows the derivation of the intersubstitutivity of (i) with (ii). Such a solution is in fact possible, as I will demonstrate in due course. In fact, it is a paracomplete solution

$\neg B \vee \neg\neg B$, which is a special case of excluded middle—a special case that the intuitionist rejects along with the general principle. But the intuitionist rejects that deMorgan law, and accepts $\neg(B \wedge \neg B)$. Given this latter and the intersubstitutivity of equivalents, $B \leftrightarrow \neg B$ implies $\neg(B \wedge B)$ and hence implies $\neg B$, from which it is easily seen to imply $B \wedge \neg B$.

In the paracomplete solutions I'll consider, $\neg(B \wedge \neg B)$ is equivalent to $B \vee \neg B$, so is not generally accepted, and so this derivation of a contradiction from $B \leftrightarrow \neg B$ is blocked. (As we'll see, calling $\neg(B \wedge \neg B)$ the "law of non-contradiction" is quite misleading. One can reject all contradictions without accepting all instances of that "law". And theorists who accept both B and $\neg B$ will almost certainly accept $\neg(B \wedge \neg B)$ as well as accepting $B \wedge \neg B$. So accepting the "law" is neither necessary nor sufficient for rejecting contradictions.)

within a reasonably strong deMorgan logic that I will eventually be advocating in this book.

The types of solution sketched above do not exhaust the possibilities. For instance, I've noted that one can block the Central Argument at Step Three instead of Step Four:

Semi-classical (*aka* **Weakly Classical**)[9] **Solutions:** Reasoning by Cases fails for some arguments involving 'instantiates' (but excluded middle is valid). In particular, the claim that R instantiates itself implies a contradiction, and the claim that it doesn't instantiate itself also implies a contradiction. But the claim that it either instantiates itself or doesn't is perfectly consistent, and indeed true! (We must reject Yogi Berra's advice "If you come to a fork in the road, take it.")

This semi-classical route doesn't really block the Russell paradox (not to mention other paradoxes of Naive Property Theory): it blocks the Central Argument for the contradictoriness of $B \leftrightarrow \neg B$ but leaves other arguments intact. It turns out that just as no theory based on full classical logic can consistently accept the Naive Theory of Properties, no semi-classical theory can either (where by a semi-classical theory I mean one that keeps pretty much all of classical logic except Reasoning by Cases and other meta-rules based on it; I'll be more precise in later chapters). But there are ways to develop consistent semi-classical solutions in which the limitations on (INST) are a bit less drastic than on the fully classical solutions. I will devote several chapters to such semi-classical solutions.

The paracomplete approach is not the only one that can be combined with acceptance of the Naive Theory of Properties. Another such approach accepts the conclusion of Step Four: that is, certain explicit contradictions (sentences of form $B \wedge \neg B$) are deemed acceptable. This would be intolerable if, as in classical logic, contradictions were taken to imply everything, but the idea behind this option is to weaken classical logic to a "paraconsistent" logic in which this is not so. So we have

Paraconsistent Dialetheic Solutions: R both instantiates itself and doesn't instantiate itself (but this contradiction is "contained" so as not to imply e.g. that the earth is flat).

I'm not myself a fan of such solutions, but I do think they merit discussion, and they will receive it in this book.

There is a still more radical kind of approach that I shall not consider: it involves revising the structural rules of logic, such as the transitivity of implication, or the principle that the question of whether $A_1, \ldots, A_n$ imply B depends only on B and the formulas in $\{A_1, \ldots, A_n\}$ (that is, not on the ordering of the A_i or on whether any of the A_i occur more than once on the list). I haven't seen sufficient reason to explore this kind of approach (which I find very hard to get my head

[9] There is a terminological issue here, of whether to count these solutions as classical. My preference would be not to, but proponents of such solutions tend to regard them as classical; so it seems less prejudicial to use the term 'weakly classical', and this is the policy I have adopted in the main text. But in the rest of the Introduction I have decided to give in to my prejudices.

around), since I believe we can do quite well without it. Throughout the book I will take the standard structural rules for granted.[10]

5. GRELLING AGAIN

I noted at the end of Section 3 that one reaction to Russell's paradox for properties might be: "So much the worse for properties! This just shows that there's no reason to believe in such obscure entities!"[11] Indeed, even one who believed there to be a place for *natural* properties could say that it is *conceptual* properties that lead to paradox, and that there is no reason to believe in *them*. In short, the view would be that if there is an intelligible notion of property at all, it must be one for which the Non-Existence Solution is right: there is simply no such thing as the alleged Russell property.

Rather than taking a stand on this issue, let's go back to the case of predicates, whose existence is much clearer. Because of the evident structural analogy between Russell's paradox for properties and the heterologicality paradox, it is clear that most of the solutions for the former have analogs for the latter. In particular we have the following:

Underspill Solutions (predicate version): For some predicates F, we have: 'F' isn't true of some objects that are F.

Overspill Solutions (predicate version): For some predicates F, we have: 'F' is true of some objects that aren't F.

Paracomplete Solutions (predicate version): Excluded middle fails for some sentences involving 'true of'. In particular, the assumption that 'is not true of itself' is either true of itself or not true of itself is fallacious.[12]

Paraconsistent Dialetheic Solutions (predicate version): 'Is not true of itself' both is true of itself and is not true of itself.

[10] I will also take for granted the usual introduction and elimination rules for conjunction. This rules out of consideration "non-adjunctive logics", which can allow the acceptance of contradictory pairs $\{A, \neg A\}$ without going on to accept contradictions in the sense officially defined (single sentences of form $A \wedge \neg A$). Since I assume the conjunction rules, I'll often speak of contradictory pairs as contradictions.

[11] Quine regarded properties as obscure not on grounds of their being "platonic entities" (after all, sets in the standard mathematical sense are "platonic" also), but on the grounds of their being non-extensional: distinct properties can be instantiated by exactly the same things. But note that *this non-extensionality played no role in Russell's paradox*, and anyone who thinks that classical logic must carry over unchanged even when dealing with "circular concepts" must reject a naive theory even of extensional analogs of conceptual properties ("Fregean extensions").

[12] The use of the term 'paracomplete' for solutions that restrict excluded middle was suggested to me by JC Beall, and seems the appropriate dual for 'paraconsistent', which is standardly taken as the term for theories that reject the rule of explosion (explained in the Appendix to Chapter 3). But a warning is in order: the term 'paracomplete' is sometimes used in a very different sense, for "gap theories" on which 'is not true of itself' is neither true of itself nor false of itself. These theories do not restrict excluded middle, but instead posit Underspill; I will discuss this in more detail in Section 3.5.

Semi-classical Solutions (predicate version): It's absurd to suppose that 'is not true of itself' is true of itself, and equally absurd to suppose that it isn't true of itself. Nonetheless, either it is true of itself or it isn't true of itself.

My preference is going to be for a Paracomplete solution without Underspill or Overspill: one which not only validates the schema (TO), but involves a strong enough logic so that this entails the following

Intersubstitutivity Principle: If C and D are alike except that (in some transparent context) one has the formula $A(x)$ where the other has the formula $\langle A \rangle$ *is true of* x, then C implies D and *vice versa*.

(A theory with (TO) and Intersubstitutivity will be called a Naive Theory of truth-of.)

What about a Non-Existence Solution to the heterologicality paradox? The analog of the Non-Existence Solution to Russell's Paradox would seem to be that there's no such thing as the expression 'is not true of itself'. But while this is minimally consistent, it just seems ludicrous: you can see the expression before your eyes, inside quotation marks at the end of the previous sentence. And not only is it an expression, it clearly satisfies the syntactic criteria for being a predicate of English. The schema (TO) from Section 1

(TO) 'F' is true of all and only the things that are F

was supposed to apply to all predicates of English. If we insist on applying full classical logic (including reasoning by cases) to the predicate 'true of', it seems that the Underspill and Overspill options are the only ones that could possibly be maintained.

Are there ways to get around this? One idea would be to insist

(I) that the Schema (TO) should be restricted not just to predicates in the syntactic sense, but to *meaningful* (or *intelligible*) predicates,

and to argue

(II) that 'is not true of itself' isn't a meaningful (or intelligible) predicate.

Then we could with some stretching call this a Non-Existence Solution: what doesn't exist is a *meaningful* predicate.[13]

But in any ordinary sense of 'meaningful', the claim (II) just seems obviously false. (Indeed, in any ordinary sense of 'meaningful', it's hard to see how 'is not true of itself' can fail to be meaningful unless 'true of' fails to be meaningful.) Evidently it must be some special technical sense of 'meaningful' that is intended.

[13] I do not mean to include here views on which 'is not true of itself' is *ambiguous* (due to an ambiguity in 'true of'). I will consider such stratified views in later chapters; they are not naturally thought of as "Non-Existence Solutions" to the heterologicality paradox, since they are much more reminiscent of type-theoretic solutions to the set-theoretic paradoxes than of theories like Zermelo-Fraenkel that deny set existence.

The most obvious possibility is that a 'meaningful predicate' in the technical sense is a predicate that expresses a property. If we adopt this line we must rely on a theory of properties that avoids the Russell paradox by the Non-Existence Solution to that. (If only we had an attractive theory of properties of this sort!)

Moreover, the claim in (I) that (TO) holds for "meaningful" predicates (or better, predicates that express properties) doesn't settle the application of 'true of' to "meaningless" predicates (predicates that don't express properties). We can of course settle this by a separate stipulation.

One possible stipulation is that a predicate that doesn't express a property is never true of anything. In that case, 'is not true of itself' isn't true of anything, so it isn't true of itself; but this supposition is just a version of the Underspill solution. (The predicate in question isn't true of itself, but 'isn't true of itself' isn't true of it!) In other words, to handle this possibility we don't need a separate "Non-Existence Solution"; it is already covered in our classification as an "Underspill Solution".

A second possible stipulation (with less *prima facie* appeal) is that a predicate that doesn't express a property is true of everything. That is easily seen, by a parallel argument, to be one version of an Overspill Solution.

A third possible stipulation is that predicates that don't express properties are true of some things and not others, but in a rather unsystematic way; such a view is likely to be committed to both Overspill and Underspill, but again it is already covered in the classification.

Of course, an advocate of the route laid out in (I) and (II) above might well say that the question of whether "meaningless" predicates are true of things is itself "meaningless", at least in some cases: in particular, it is "meaningless" to ask whether the "meaningless" predicate 'is not true of itself' is true of itself. And (especially if we forget for the moment that it is a special technical sense of meaningfulness that is being employed!) this might be taken to imply that we shouldn't assert that it is true of itself, or assert that it isn't, or assert the disjunction of the two. But in that case too, the view is covered in our classification: it is a kind of paracomplete view, one which withholds assent to excluded middle for (some or all) sentences containing predicates that are "meaningless" in the technical sense. Again, there's no need for a separate "Non-Existence Solution" to cover this case.[14]

In short, to say that the predicate 'is not true of itself' is in some sense meaningless is thoroughly unhelpful: it simply disguises the true nature of the solution being proposed.

[14] If on the other hand the view asserts that though it is "meaningless" in the technical sense to ask whether 'is not true of itself' is true of itself, still (since it's meaningful in the ordinary sense) we ought to assert

'Isn't true of itself' is either true of itself or not true of itself,

then we will have either Underspill or Overspill or some semi-classical or dialetheic option, despite the attribution of "meaninglessness".

Admittedly, saying that the heterologicality predicate doesn't express a property isn't as unnatural for some solutions to the heterologicality paradox as it is for others. The solutions most naturally described in this way are (1) ones according to which predicates that don't express properties aren't true of anything, and (2) ones according to which atomic sentences containing predicates that don't express properties should be neither asserted nor denied. But even in these cases, the description of the theory as a Non-Existence solution hides crucial details.

Solutions of type (1) are disguised Underspill solutions to the heterologicality paradox; but in order to tell *how much* underspill they posit, we must first find out what they tell us about the extent of the failure of predicates to express properties. Of course, in order to disguise an Underspill solution to the heterologicality paradox in this way, we must adopt a Non-Existence solution to the Russell paradox for properties. That is not my preferred approach to the property paradox. But putting that aside, it seems methodologically preferable to state the solution to the heterologicality paradox without invoking issues of property-existence, for the simple reason that we have clearer intuitions about what predicates are true of than about what properties exist.

Solutions of type (2) are disguised forms of Paracomplete solutions to the heterologicality paradox; but in order to tell *where excluded middle needs restricting* we need to know which predicates fail to express properties.[15] The remarks under (1) then carry over: we'd do better just to present the solution directly as a Paracomplete solution to the heterologicality paradox. For in the first place, it's easier to evaluate theories of truth-of than theories of property-existence. And in the second place, developing the solution directly as a solution to the heterologicality paradox means that we don't need to rely on a Non-Existence solution to the Russell paradox for properties. We might, for instance, want to adopt the same type of solution to the Russell paradox as to the heterologicality paradox: e.g. a Paracomplete theory for both.[16]

I will have more to say about these issues as we proceed.

6. CHANGE OF LOGIC AND CHANGE OF MEANING

There are many other paradoxes closely related to the ones I've discussed. The most well-known is the Liar paradox, which is very similar to the heterologicality

[15] Actually some such solutions are disguised semi-classical solutions (see previous footnote); in this case, it is *the extent to which reasoning by cases needs restricting* that depends on the degree to which there's a failure of predicates to express properties.

[16] Moreover, if one thinks that the predicates for which excluded middle is suspect are exactly those that don't express properties, this is likely to prejudice the type of Paracomplete solution one entertains. e.g. one might think that if a predicate doesn't express a property, then no disjunction of that predicate with another could express a property either; but that would force a rather unattractive Paracomplete solution with a very weak logic.

paradox but in some respects simpler to discuss. It will be my starting point. But there are many others, really an infinite variety: these include the rather troublesome Curry paradox (also called Löb's paradox or Henkin's paradox), and the even more troublesome *iterated* Curry paradoxes. There are also paradoxes of denotation and definability. It is important that we not just deal with each paradox piecemeal, but provide a systematic account that can be shown to handle them all. I will be surveying a wide range of options for such an account, and discussing their advantages and disadvantages.

A theme of this book will be that we ought to seriously consider restricting classical logic to deal with all these paradoxes. In particular, we should seriously consider restricting the law of excluded middle (though not in the way intuitionists propose). I say 'restricting' rather than 'abandoning', because there is a wide range of circumstances in which classical logic works fine. Indeed, I take excluded middle to be *clearly* suspect only for certain sentences that have a kind of "inherent circularity" because they contain predicates like 'true'; and most sentences with those predicates can be argued to satisfy excluded middle too. The idea is *not* that we need two distinct logics, classical logic and a paracomplete logic (one without excluded middle). On the contrary, the idea is that we can take the paracomplete logic to be our single all-purpose logic. But we can recognize the truth of all instances of excluded middle in certain domains (e.g. those that don't contain 'true', 'true of', 'instantiates', or other suspect terms). We may not want to count these instances of excluded middle as *logical* truths (though you can if you like—this is a matter for verbal stipulation), but we can recognize them as truths, and truths with something of the high degree of certitude that many logical truths have. So we can reason from them, and reasoning from them will look just like classical reasoning in the domains in question.

Weakening classical logic (whether by restricting excluded middle or in some other way) is not something to be done lightly. There are some obvious advantages to keeping to classical logic even for "circular" predicates: advantages of simplicity, familiarity, and so on. Choosing to forgo these advantages has its costs. But I will argue (primarily in Part II) that the *disadvantages* of keeping classical logic for "circular" predicates are also very great, so that the undoubted cost of weakening the logic is worth bearing.

Perhaps there are some who think that this cost-benefit analysis is inappropriate, that the very idea of tinkering with classical logic is irrational on its face since classical logic is obviously superior. The word 'logicism' would be a natural name for this attitude—in analogy to 'sexism', 'racism', 'species-ism' and so forth. Unfortunately it's already taken, so let's call the view 'Logical Dogmatism'.

One possible defense of such Dogmatism is that if logic is not held fixed then anything goes. As an anonymous referee put it to me, "We didn't weaken the logic as a way of hiding the defects in Ptolemaic astronomy or old quantum theory; why should we modify the logic to hide the blemishes in the naive theory of truth?" The answer to this, I think, is that there is no known way (and little prospect of finding a way) to save either Ptolemaic astronomy or the old quantum theory by a change of logic, and little benefit to so doing since we have far more satisfactory alternatives. The proposal that we save the naive theory of truth by a change of logic is not the cheap non-solution that the objection envisages: it is something that must be earned by working out the details of the logic and of the theory based on it. Once we've worked out such a theory, we must weigh it against competing theories that keep classical logic but restrict the naive principles of truth, using the usual (somewhat vague and subjective) criteria for theory choice. With Ptolemaic astronomy or the old quantum theory, there is no serious prospect for such a theory being worked out that survives such a competition. The reader may think there is little prospect in the case of the theory of truth either, but I invite him to withhold judgement until he has read the book.

A second common defense of Logical Dogmatism is based on the idea that "change of logic requires change of meaning". To this I say, first, that the paradoxes force a change in the basic laws *either* of logic in a narrow sense *or* of the logic of truth, instantiation, etc.; or if you like, it forces a change in opinion about those laws. If change of (opinion about) the basic laws of '$\neg$' and '$\rightarrow$' counts as change of meaning, why doesn't change of (opinion about) the basic laws of truth and the basic laws of instantiation? And as we'll see, adhering to the principles of classical logic requires a *huge* change in standard principles about truth and instantiation. The upshot of this is that there is at least as good a case that the classical truth theorist is "changing the meaning of 'true' " as that the defender of the Intersubstitutivity Principle who restricts excluded middle is "changing the meaning of 'not' " (or of 'or').

But second, why make a fetish about whether these things involve a change of meaning? As Putnam 1968 taught us, there is a distinction to be made between change of meaning and *mere* change of meaning. The switch from Euclidean geometry to generalized (variable curvature) Riemannian geometry involved revision of enough basic principles about straight lines that it may be somewhat natural to say that 'straight line' took on a different meaning. But if so, an abandonment of the old meaning and the invention of a new one was required to get a decent physical theory that is observationally adequate: for no reasonably simple observationally adequate theory allows for the existence of "straight lines in the Euclidean sense". We needn't of course have carried over the old term 'straight line' from Euclidean geometry to Riemannian, but there is enough continuity of doctrine to make it natural to do so. This is certainly not a *mere* change of meaning, i.e. a relabelling of terms without alteration of basic theory. The situation with truth is similar: here the "old theory", involving both classical logic and the naive theory of truth, is simply inconsistent. Indeed it's trivial: it implies everything, e.g. that the Earth is flat. If you don't want to be committed to the view that the Earth is flat you need a theory that differs from the naive theory in basic principles, either principles about truth or principles about logical matters more narrowly conceived. If giving up those basic principles involves a "change of meaning", so be it: for then the "old meanings" aren't really coherent, and they *need* changing. This is certainly no *mere* change of meaning, i.e. no mere relabelling.

Any resolution of the paradoxes will involve giving up (or at least restricting) some very firmly held principles: either principles of a narrowly logical sort, or principles central to the ordinary use of truth and instantiation predicates, or both. The principles to be given up are ones to which the average person simply can't conceive of alternatives. That's why the paradoxes are *paradoxes*. In this situation, I think we should be skeptical that asking whether the attempted resolution of the paradoxes "changes the meaning" of the terms involved is a clear question (whether these be ordinary logical terms like 'not' and 'if ... then' or terms like 'true' and 'instantiates'). And I'm even more skeptical that it's a useful question.

The question is *clear* only to the extent that we know how to divide up such firmly held principles into those that are "meaning constitutive" or "analytic" and those which aren't, and this is notoriously difficult. If forced to answer the question in the case of the paradoxes, perhaps the best thing to say is that *all* the principles leading to paradox are "meaning constituting"—logical principles like excluded middle and disjunctive syllogism as well as principles like (TO) and (INST). In that case, the ordinary meanings are jointly incoherent, and a "change of meaning" is required to restore coherence. (Of course, we shouldn't "change meanings" gratuitously, and there are many reasons for trying to keep the new uses of our terms broadly consonant with the old uses.) Perhaps there are subtle techniques of linguistic analysis that would enable us to discover that certain of these principles involved in the paradoxes aren't really constitutive of the "meanings" that English speakers attach to their words in the way that the rest of the principles are, but I am skeptical.

But these linguistic questions, even if deemed clear, strike me as of very little interest. The paradoxes show that there's something wrong with firmly held patterns of reasoning, whether or not these patterns are meaning-constituting. What's of interest is to figure out how best to modify this reasoning: to find a new way of reasoning that we can convince ourselves is intuitively acceptable, and which avoids plainly unacceptable conclusions. Conceivably there's more than one reasonable way to do it, with no non-arbitrary way to choose between them, but we'll be in a better position to judge this after we've investigated the possible solutions. In any case, there is no obvious reason in advance of such an investigation that would rule out the possibility that the best resolution(s) will involve a revision of some general logical principles rather than a revision of the principles of truth and instantiation.

I'm certainly not suggesting that anyone adhering to classical logic is a Dogmatist: one may well think that the benefits of keeping classical logic outweigh the costs. One of the major goals of the book is to provide the materials for an informed decision on this issue.

7. SOME CONNECTIONS TO OTHER ISSUES

I take the resolution of the paradoxes to be a matter that should be of quite general interest, as it has a bearing on many other topics. Obviously it is crucial for our investigations into the nature of truth that we have a consistent theory of truth,[17] and likewise for properties. Moreover, what our consistent theory looks like will affect our answers to general philosophical issues about truth and properties. For instance, in recent years there has been much discussion of what purposes the notion of truth serves. One common answer (particularly among advocates of "deflationary" or "minimal" theories of truth) is that 'true' is primarily a logical device that enables us to make generalizations we would otherwise be unable to make, or unable to make succinctly. According to this viewpoint, even apparently "deep" claims about truth turn out to result from the generalizing function. For instance, the "deep" claim

> Truth is distinct from long-run verifiability

is basically just a generalization of

> It is possible that either Thales' maternal grandmother ate cabbage on the day she died but there is no way to verify this, or there are now eight neutrinos in the interior of the sun that are arranged precisely in a regular octagon but there is no way to verify this, or ... .

For this view of truth to work, we require a quite strong equivalence between claims of form "'*p*' is true" and the corresponding claims "*p*"—or so I will argue in Chapter 13. If this is right, then the question of whether this strong equivalence can be maintained in general (without, e.g., analogs of "Underspill" or "Overspill") will turn out to be crucial to these "deflationary" or "minimalist" views.

The paradoxes are also of relevance to metaphysical issues about "indefinite extensibility" of our concepts and/or our ontology. For instance, it is often thought that our concept of truth is indefinitely extensible: there is a simple recipe whereby given any precise theory of truth, we can automatically construct a more extensive one. The general tenor of this book is opposed to that idea. In particular, I will argue against the common view that the semantics of any language can only be given in a broader meta-language; my claim will be that there are languages that are sufficiently powerful to serve as their own meta-languages. This is not to claim that there could be such a thing as a "universal" language capable of expressing any possible concept; only that there are languages comprehensive enough to be "self-sufficient" in that they do not contain within themselves the requirement that they be expanded.

[17] Or at least a non-trivial one, i.e. one that doesn't imply everything.

The topics of this book also bear on other issues within mathematical philosophy. One such issue that I will discuss is the nature of logical consequence: some widespread views about this are simply undermined by the paradoxes, or so I will argue. I will also discuss a related issue about Gödel's second incompleteness theorem. The theorem says, very roughly, that no good mathematical theory can prove its own consistency. It is natural to ask: "Why not? Can't it prove that its axioms are true, and that its rules of inference preserve truth? And if so, can't we argue by induction that its theorems are true, and infer from this that it is consistent?" It turns out that different theories of truth have very different things to say about where this reasoning goes wrong. (For many theories of truth, the diagnosis they must give points to highly counterintuitive features of those theories.) This is another topic to which I'll pay considerable attention.

Finally, the topics of this book may well relate to other cases where there seems intuitively to be "no determinate fact of the matter" about the answer to some question. For there is *some* reason to think that the "circularity" of predicates like 'instantiates' and 'true' leads to trouble *because it gives rise to an indeterminacy in the application of these predicates*; if so, it would be natural to extend our account of circular predicates to indeterminacy more generally. One plausible example where indeterminacy may arise is in the application of vague predicates: it is plausible that in some such cases "there is no determinate fact of the matter" as to whether the predicate applies. So there may well be a connection between the truth-theoretic paradoxes and issues of vagueness. (Such a connection has been posited in McGee 1989 and 1991 and Tappenden 1993, though it is by no means generally accepted.) This suggestion raises huge issues, and I won't take a strong stand on it in this book. But I am, on the whole, sympathetic to the suggestion, and in Chapters 5 and 9 I explore some parallels and draw some tentative morals.