# The humble origins of Russell's paradox

*by J. Alberto Coffa*

ON SEVERAL OCCASIONS Russell pointed out that the discovery of his celebrated paradox concerning the class of all classes not belonging to themselves was intimately related to Cantor's proof that there is no greatest cardinal.[1] One of the earliest remarks to that effect occurs in *The Principles of Mathematics* where, referring to the universal class, the class of all classes and the class of all propositions, he notes that

> when we apply the reasoning of his [Cantor's] proof to the cases in question we find ourselves met by definite contradictions, of which the one discussed in Chapter X is an example. (P. 362)

And in a footnote he adds: "It was in this way that I discovered this contradiction".

Throughout his writings Russell left a number of hints concerning the sort of connection he had drawn between Cantor's proof and his own discovery. In fact, his suggestions are so specific that there would seem to be little room left for speculation concerning how the discovery took place.[2] The picture that emerges almost immediately from Russell's observations is the following. For reasons which are

[1] See, e.g., Bertrand Russell, *The Principles of Mathematics*, 2nd ed. (London: Allen & Unwin, 1937), §§100, 344-9; G. Frege, *Wissenschaftlicher Briefwechsel* (Hamburg: Felix Meiner Verlag, 1976), pp. 215-16; B. Russell, *Essays in Analysis*, ed. D. Lackey (New York: George Braziller, 1973), p. 139; B. Russell, *Introduction to Mathematical Philosophy* (New York: Simon and Schuster, 1971), p. 136; B. Russell, *My Philosophical Development* (New York: Simon and Schuster, 1959), pp. 75-6; *The Autobiography of Bertrand Russell*, I (New York: Bantam Books, 1968), 195.

[2] See, e.g., Ch. Thiel's "Einleitung des Herausgebers" in Frege, *Wissenschaftlicher Briefwechsel*, pp. 203 and 216 (footnote), and I. Grattan-Guinness, "How Bertrand Russell Discovered his Paradox", *Historia Mathematica*, 5 (1978), 127-37.

never made clear, Russell decided to analyze Cantor's argument applying it to "large" classes such as the universal class $V$, and the class of all classes. When, for example, we consider $V$, its power set and the correlation $f(x) = \{x\}$ if $x$ is not a class, $f(x) = x$ otherwise, then Cantor's diagonal class $D$ turns out to be the class of all classes not belonging to themselves. Moreover, since the element of $V$ which is taken to $D$ by $f$ is $D$ itself (i.e., since $f(D) = \mathrm{D}$), Cantor's reasoning invites us to raise the question whether $D$ belongs to $f(D)$ (i.e., to $D$) or not; and it establishes that it does precisely if it doesn't.[3]

The purpose of this note is to present recently uncovered information that complements and corrects our present understanding of Russell's discovery of his paradox. As it turns out, far from originating from his desire to apply the ideas in Cantor's theorem, a version of Russell's paradox first occurred in an argument that Russell had devised, late in 1900, in order to establish the invalidity of that theorem. An appeal to "large" classes such as $V$ or *Class* was, as we shall see, essential to Russell's attempted refutation. As he displayed the details of his counterexample Russell's paradox emerged, at first unrecognized, as the by-product of a project that the paradox itself would eventually undermine.

In a paper written in 1901 and largely devoted to a popular exposition and defence of Cantor's theory of the infinite, Russell expressed what appeared to be a minor reservation to Cantor's treatment:

> There is a greatest of all infinite numbers, which is the number of things altogether, of every sort and kind. It is obvious that there cannot be a greater number than this, because, if everything has been taken, there is nothing left to add. Cantor has a proof that there is no greatest number, and if this proof were valid, the contradictions of infinity would reappear in a sublimated form. But in this one point, the master has been guilty of a very subtle fallacy, which I hope to explain in some future work.[4]

Russell never published his main criticism of Cantor's proof—indeed, we have reason to think that by mid-1901 he had abandoned it; but materials in the Russell Archives allow us to reconstruct it quite accurately.

On December 8, 1900 Russell wrote to Louis Couturat:

> I have found a mistake in Cantor, who holds that there is no greatest cardinal number. But the number of classes is the greatest number. The best of Cantor's proofs of the opposite appears in Jahresb. d. deutschen Math. Ver'g., I, 1892, pp. 75-78. In essence it consists of showing that, if $u$ is a class whose number is $\alpha$, the number of classes contained in $u$ (which is $2^{\alpha}$) is greater than $\alpha$. But the proof presupposes that there are classes contained in $u$ which are not individuals [members] of $u$; but if $u =$ Class, that is false: every class of classes is a class.[5]

On January 17, 1901 Russell reiterated the point: there is a greatest cardinal, he wrote once again to Couturat, but from this

> no contradiction follows, since the proof that Cantor gives that
>
> $$\alpha \epsilon \mathrm{Nc.} \supset . \; 2^{\alpha} > \alpha$$
>
> presupposes that there is at least a class contained in a given class $u$ (whose number is $\alpha$) that is not itself an individual of $u$, i.e., that we have:
>
> $$\exists \mathrm{cls} \cap v \ni (v \subset u \,.\, v \sim \epsilon u).$$
>
> If we put $u =$ Cls, this would become false. Hence the proof does not hold.[6]

[3] See *Principles*, §349. Perhaps I should remind the reader of Russell's reformulation of Cantor's reasoning. We are invited to consider a one-one onto function (bijection) $f$ between an arbitrary class $A$ and its power set $PA$, and to concentrate on the diagonal class $D$ of elements $x$ in $A$ which do not belong to $f(x)$. $D$ must be in $PA$ so that, since $f$ is a bijection, for some $t$ in $A$, $f(t) = D$. But as we raise the question whether $t$ belongs to $D$ we find that it does if and only if it does not: a contradiction. Hence $t$ does not exist and $f$ cannot be a bijection.

[4] "Mathematics and the Metaphysicians", in *Mysticism and Logic* (London: Unwin Books, 1963), p. 69. The date of composition is given by Russell in his Introduction to *Mysticism and Logic*, p. 7. In 1917 Russell added the following footnote to the passage I quote in the text: "Cantor was not guilty of a fallacy on this point. His proof that there is no greatest number is valid. The solution of the puzzle is complicated and depends upon the theory of types, which is explained in *Principia Mathematica*, Vol. I (Camb. Univ. Press, 1910)."

[5] "J'ai découvert une erreur dans Cantor, qui soutient qu'il n'y a pas un nombre cardinal maximum. Or le nombre des classes est le nombre maximum. La meilleure des preuves du contraire que donne Cantor se trouve dans Jahresb. d. deutschen Math. Ver'g. I, 1892, pp. 75-78. Elle consiste au fond à montrer que, si $u$ est un classe dont le nombre est $\alpha$, le nombre des classes contenues dans $u$ (qui est $2^{\alpha}$) est plus grand que $\alpha$. Mais la preuve présuppose qu'il y a des classes contenues dans $u$ qui ne sont pas des individus d'$u$; or si $u =$ classe, ceci est faux: tout classe de classes est une classe."

[6] "Mais il n'en résulte aucune contradiction, puisque la preuve que donne Cantor que

$$\alpha \epsilon \mathrm{Nc.} \supset . \; 2^{\alpha} > \alpha$$

présuppose qu'il y ait au moins une classe contenue dans une classe donnée $u$ (dont le nombre est $\alpha$) qui n'est pas elle-même un individu de $u$, c'est à dire qu'on a

$$\exists \mathrm{cls} \cap v \ni (v \subset u \,.\, v \sim \epsilon u)$$

Si l'on met $u =$ Cls, ceci devient faux. Donc la preuve ne tient plus." [" $\ni$ " is Peano's sign for "such that".]

Russell's letters to Couturat solve the problem of deciding what Cantor's "subtle fallacy" was, but they create another one. For they do not contain so much as a hint of why Russell thought that Cantor's reasoning involved such an assumption. The situation is all the more intriguing since, on the face of it, there is nothing in Cantor's proof involving an assumption of the sort Russell thought he could uncover. Fortunately, what Russell had in mind is fully explained in another document in the Russell Archives, an early draft of what would eventually become Chapter 43 of *Principles*.[7]

The draft is concerned with the philosophy of the infinite, focusing on Cantor's ideas. Once again, when he comes to deal with Cantor's theorem Russell observes that the universal class must have the largest cardinal, and concludes that there must be an error in Cantor's proofs. "If these proofs be valid", he tells us, "there would seem to be still a contradiction. But perhaps we shall find that his proofs only apply to numbers of classes not containing all individuals ... " (folio 189). The draft continues with a detailed criticism of Cantor's first proof, which is preserved almost unchanged in §345 of the published version of *Principles* (pp. 363-4).[8] The discussion in the draft then proceeds essentially as in the printed version through the first paragraph of §347, which concludes with Russell's reformulation of the conclusion of Cantor's second proof: "the number of classes contained in any class exceeds the number of terms belonging to the class". At this point the draft and the printed version diverge drastically. The draft proceeds as follows:

> Now if *u* be the class of classes, this is plainly self-contradictory, for classes contained in *u* will be only classes of classes, whereas terms belonging to *u* will be all classes without restriction, so that the classes contained in *u* are a proper part of the class *u* itself. Hence there must be somewhere in Cantor's argument a concealed assumption not verified when *u* is the class of all classes. (f. 196)

[7] The last page of this draft is dated Nov. 24, 1900. This must be part of the first draft of *Principles* which Russell says he finished "on the last day of the nineteenth century" (*My Philosophical Development*, p. 73). The relevant file is 230.03050-F14. In a private communication the Archivist has offered compelling evidence that this "draft" manuscript is also the version that Russell sent to the Cambridge University Press as Chapter 43 of *Principles*. Consequently, the passages which we discuss below seem to have been part of the 900-page manuscript which Russell delivered to the printer in May 1902. This poses the problem of determining why Russell would have submitted a manuscript which contained an argument that he regarded (at least since October 1901; see fn. 14 below) as fallacious.

[8] §344 of *Principles*, which contains the passage quoted in the first paragraph of this note, does not occur in the manuscript that we are discussing.

Having established to his satisfaction that Nc‘$u <$ Nc‘$2^u$ is false when $u$ = *Class*,[9] Russell now turns to an examination of the second proof. Since $u$ = *Class* is a counterexample to the conclusion, what better way to identify the gap in the proof than to go through it, step by step, but bearing now in mind not an arbitrary class $u$ but the specific class *Class*? In order to apply Cantor's reasoning we need, of course, not only *Class* and its power set, *Class of classes*, but also a function (not necessarily a bijection[10]) to correlate them; and is there a simpler function from *Class* to *Class of classes* than $k(x) = \{x\}$, if $x$ is not a subclass of *Class*; $k(x) = x$ otherwise?[11] Now we can return to Russell's draft:

> The argument by which it is to be shown that the number of classes of classes exceeds the number of classes may be disproved in the following manner. We have $u$ = *Class*, so that "$x$ is a $u$" means "$x$ is a class." When $x$ is not a class of classes, let $k_x$ be the class of classes whose only member is $x$. When $x$ is a class of classes, let $k_x$ be $x$ itself. Then we define a class $u'$, in accordance with the above procedure [i.e., Cantor's diagonal method], as containing every $x$ which is not a member of its $k_x$, and no $x$ which is a member of its $k_x$. Thus when $x$ is not a class of classes, $x$ is not a $u'$; when $x$ is *class*, or *class of classes*, or *class of classes of classes*, or *etc.*, $x$ is not a $u'$;[12] but when $x$ is any other class of classes,[13] $x$ is a $u'$. Then Cantor infers that $u'$ is not identical with $k_x$ for any value of $x$. But $u'$ is a class of classes, and is therefore identical with $k_{u'}$. Hence Cantor's method has not given a new term, and has therefore failed to give the requisite proof that there are numbers greater than that of classes. In fact, the procedure is, in this case, impossible; for if

[9] *Class* is the class of all classes; *Class of classes* is the class of all classes of classes (i.e., the class of subclasses of *Class*); etc.

[10] Cantor's reasoning establishes that given a class $A$, its power set $PA$, and a *function* $f$ from $A$ into $PA$, the class $u = \{x | x \epsilon A \;\&\; x \not\epsilon f(x)\}$ cannot be $f(t)$ for any $t \epsilon A$. For, if there were such a $t$, one could show that $t \epsilon f(t)$ iff $t \not\epsilon f(t)$. Hence, any function from $A$ into $PA$ must leave some subclass of $A$ without a partner in $A$. Russell's counterexample is addressed to this claim.

[11] In "How Bertrand Russell Discovered his Paradox", Grattan-Guinness claims that had Russell used *Class* rather than $V$ in the development of his paradox (as in fact he did), he "could have simplified his reasoning by ... setting up $f$ [the correlation between *Class* and its power set] as the (one-one) identity correspondence between [*Class*] and its power-class" (p. 130). But Grattan-Guinness' correlation is not even a relation (let alone a 1-1 function) with domain identical with *Class* and range in its power-class (since there are elements in *Class* not contained in *Class*).

[12] The reason being that the class of all classes is a member of itself; that the class of all classes of classes is a member of itself; and so on.

[13] Russell seems to have overlooked sets such as the class of classes with more than two elements.

> we apply it to $u'$ itself, we find that $u'$ is a $k_{u'}$, and therefore not a $u'$; but from the definition, $u'$ should be a $u'$. In fact, when our original class consists of all possible combinations of all possible terms, the method, which assumes new combinations to be possible, necessarily fails, since, in this case, $u'$ itself is a $u$. Thus what Cantor has proved is, that any power other than that of all classes can be exceeded, but there is no contradiction in the fact that this power cannot be exceeded. The exact assumption in Cantor, which *class* fails to satisfy, is that if $u$ be the class whose power is to be exceeded, not all classes of $u$ are themselves terms of $u$. (f. 197-8)

Russell's remark to Couturat is now clear: Cantor presupposes that not all subclasses of a set belong to it because *under the correlation k drawn by Russell* each subclass of a certain set $u$ (namely, *Class*) is associated with itself. Under these circumstances, only if we assume that some subclass must always be absent from $u$ (in Russell's case, only if we assume that his strange diagonal set $u'$ is not in $u$) could the correlation $k$ fail to be a function from $u$ into its power set. Moreover, Cantor had attempted to show that, for any correlation (such as Russell's $k$) there are classes (such as Russell's $u'$) which $k$ cannot correlate with anything in $u$, and which could therefore at best be correlated with a "new" element, i.e., with an element from outside $u$. But, Russell notes, when $u$ is *Class* no new element is needed or, indeed, possible, since every subclass of $u$ is in $u$. In fact, for the correlation $k$ between *Class* and its power set it is even possible to identify the element of $u$ which takes $u'$ as its $k$-value: it is $u'$ itself. "Hence Cantor's method has not given a new term."

At this point, almost as an afterthought, the contradiction emerges. With no awareness of the damage that the point makes to his preceding considerations, Russell observes that Cantor's "procedure is, in this case, impossible." What he seems to mean is that the class $u'$ that he has come up with by applying the diagonal strategy has logically unacceptable features: by definition "$u'$ should be a $u'$", but at the same time it "is a $k_{u'}$, and therefore not a $u'$". It is not easy to decide on the basis of this short, cryptic sentence, what was the reasoning behind Russell's statement. Perhaps he was led to these conflicting conclusions by simply reproducing Cantor's proofs that $D \in f(D)$ and $D \notin f(D)$ (where $D$ is the diagonal class and $f$ the purported bijection); or he may have appealed to the fallacious assumption challenged in footnote 13, in order to conclude that "by definition" $u'$ should be a $u'$. Be that as it may, Russell's claim is correct. His $u'$ is, of course, the class of all classes of classes not belonging to themselves; it is therefore true that we can prove both $u' \in u'$ and $u' \notin u'$. But far from concluding that $u'$ does not exist, Russell's reasoning is predicated on the assumption that it does. If $u'$ did not exist Russell's objection to Cantor's second proof would vanish.

That it does, in fact, vanish, is something that Russell came to recognize sometime in 1901. His next remark to Couturat on this subject occurs in a letter dated October 2, 1901: "I thought that I could refute Cantor; now I see that he is irrefutable."[14]

When did Russell discover his paradox? In several places he indicates that it happened around May 1901. And yet, as we have seen, he had set on paper, late in 1900, an argument that was a Gestalt-switch away from the celebrated paradox.[15] Something must have dawned on Russell around May 1901, but we really don't know what. He must have realized at least that his afterthought was not only destructive of his objection to Cantor but itself a contradiction stemming from assumptions common to Cantor, himself and most everyone dealing with classes in a broadly Cantorean spirit. But the published record does not allow for any much more specific conclusions concerning what Russell came to see on that *dies mirabilis*. Was it that the class $u'$ as one does not exist (or subsist?), or that—as he wrote to Frege in his letter of 16 June 1902—a function cannot act "as the indeterminate element"; or was it something else? Under the influence of Russell's Platonizing account of the course of his thought we have come to look at the discovery of his paradox as a punctual event in which a Platonic realm of set-theoretic truth was suddenly revealed to him. But perhaps the impact of the paradox slowly emerged over a period of time, and perhaps tacit decisions played no less significant a role than ostensive discoveries. The answers to these questions lie, if anywhere, in some other corner of the Russell Archives.

*Department of History and Philosophy of Science*
*Indiana University*[16]

[14] "Je croyais pouvoir réfuter Cantor; maintenant je vois qu'il est irréfutable."

[15] Except for the trivial fact that, in the draft, the construction is applied to *Class* rather than to $V$. As noted above, when the procedure is applied to $V$ the diagonal class turns out to be the class of all classes not belonging to themselves.

[16] I should like to thank Indiana University for research support which made my trip to McMaster possible.