The Structure of the Paradoxes of Self-Reference

GRAHAM PRIEST

Introduction

When one meets the paradoxes of self-reference for the first time, one is struck by the fact that they all appear to be members of a single family, generated by a common underlying principle. Indeed, Russell, who inaugurated the study of these paradoxes this century, held that this is precisely what they are. It is notoriously difficult to pin down the common underlying principle, however. For example, on closer scrutiny it is not at all clear what the Burali-Forti and Liar Paradoxes have in common. Russell himself was unable to say what held the family of paradoxes together beyond some rather unsatisfactory remarks concerning vicious circles in *Principia Mathematica*. (See for example Gödel 1944.)

It is therefore unsurprising that the modern view of the paradoxes is to the effect that there are two distinct families here, which arise from different sources, and which are to be treated quite differently. (The view is now so orthodox that this claim needs no documentation.) Although one can find something like this view expressed in Peano (1906), the founder of the orthodoxy was Ramsey (1925).

The only point of this paper is to demonstrate that Russell was right and Ramsey was wrong. The paradoxes of self-reference do have a common underlying structure, which generates the contradiction involved, and I shall spell out exactly what that is. Moreover, as we shall see, Russell had been on the right track some five years earlier.

1. Ramsey's division

Though Ramsey's division of the family into two classes is wrong, it will be useful to have it before us. His exact words are not without historical interest, so let me simply quote him:

It is not sufficiently remarked, and the fact is entirely neglected in *Principia Mathematica*, that these contradictions [the paradoxes of self-reference] fall into two fundamental distinct groups, which we will call A and B. The best known are divided as follows:

A. (1) The class of all classes which are not members of themselves.

© Oxford University Press 1994

Mind, Vol. 103 . 409. January 1994

- (2) The relation between two relations when one does not have itself to the other.
- (3) Burali Forti's contradiction of the greatest ordinal.
- B. (4) "I am lying."
 - (5) The least integer not nameable in fewer than nineteen syllables.
 - (6) The least indefinable ordinal.
 - (7) Richards's contradiction.
 - (8) Weyl's contradiction about "heterologische".

The principle according to which I have divided them is of fundamental importance. Group A consists of contradictions which, were no provision made against them, would occur in a logical or mathematical system itself. They involve only logical or mathematical terms such as class and number, and show that there must be something wrong with our logic or mathematics. But the contradictions in Group B are not purely logical, and cannot be stated in logical terms alone; for they all contain some reference to thought, language, or symbolism, which are not formal but empirical terms. So they may be due not to faulty logic or mathematics, but to faulty ideas concerning thought and language. If so, they would not be relevant to mathematical logic, if by "logic" we mean a symbolic system, though of course they would be relevant to logic in the sense of the analysis to thought. [Footnote: these two meanings of "logic" are frequently confused...](1925, p. 171 of reprint)

In hindsight, the unsatisfactory nature of Ramsey's criterion is clear. For a start, if one wants to draw a fundamental distinction, this ought to be done in terms of the *structure* of the different paradoxes. Ramsey's distinction depends on the relatively superficial fact of what vocabulary is used in the paradoxes, and, in particular, whether this belongs to mathematics properly so called. But worse, this is a notoriously shifting boundary. Ramsey was, of course, writing before the heyday of meta-mathematics. Had he been writing ten years later, it would have been clear that a number of items of vocabulary occurring in paradoxes of Group B *do* belong to mathematics. In particular, both syntactical and semantical linguistic notions became quite integral parts of mathematics. Indeed, in a sense, the work of Gödel and Tarski showed how these notions could be reduced to other parts of mathematics (number theory and set theory, respectively).

Despite this, it behoves anyone who says that Ramsey is wrong to come good with the structure underlying paradoxes of self-reference. This we will turn to in a moment. But first it will also be convenient to divide Ramsey's Group B itself into two. In the first class, Bi, belong Ramsey's (5), (6) and (7); in the second, Bii, belong, (4) and (8). A rough criterion for the distinction is that paradoxes in Bi, but not Bii, make explicit use of the notion of definability. More fundamentally, paradoxes in Bii use diagonalisation, or some similar mechanism, to establish something of the form $\alpha \leftrightarrow \neg \alpha$, and hence α and $\neg \alpha$; whereas paradoxes in Bi give independent arguments for each of α and $\neg \alpha$.

2. Russell's Schema

To exhibit the structure underlying all these paradoxes, let us start with what I shall call *Russell's Schema*, an idea dating from some years before he lighted upon the Vicious Circle Principle. In his own words this is as follows:

Given a property φ and a function δ , such that, if φ belongs to all members of u, $\delta(u)$ always exists, has the property φ , and is not a member of u; then the supposition that there is a class w of all terms having property φ and that $\delta(w)$ exists leads to the conclusion that $\delta(x)$ both has and has not the property φ . (1905, p. 142 of reprint)

Less tersely, given a property φ , and function δ , consider the following conditions:

```
1) w = \{x: \varphi(x)\} exists
2) if x is a subset of w: a) \delta(x) \notin x
and b) \delta(x) \in w
```

For future reference I will call clause (2a) the *Transcendence Condition* and clause (2b) the *Closure Condition*.

Given (1) and (2) we have a contradiction. For when Transcendence and Closure are applied to w, an irresistible force meets an immovable object. The result: $\delta(w) \in w$ and $\delta(w) \in w$. Hence, any φ and δ that instantiate Russell's Schema generate a contradiction. Russell did not claim that this structure was behind all the paradoxes. He did, however, demonstrate that it was behind all paradoxes of Ramsey's Group A. To see how, it is sufficient to examine a couple of examples.

In Russell's Paradox, the property $\varphi(x)$ is " $x \notin x$ ", so that w is the Russell set $R=\{y; y\notin y\}$; and the function δ is simply the identity function, id. Suppose that $x\subseteq w$; then $x\in x\to x\notin x$. Hence $x\notin x$, demonstrating Transcendence; it follows that $x\in R$, which is Closure. The contradiction is that $R\in R$ and $R\notin R$.

Russell's Paradox is a stripped-down version of Cantor's Paradox (see Russell, 1905), so let us move on to Burali-Forti's. $\varphi(x)$ is "x is an ordinal", and so w is the set of all ordinals, On; $\delta(x)$ is the least ordinal greater than every member of x (abbreviation: $\log(x)$). Transcendence holds by definition, as does Closure. The contradiction is that $\log(\operatorname{On}) \in \operatorname{On}$ and $\log(\operatorname{On}) \notin \operatorname{On}$.

One further example: Mirimanoff's Paradox. In this, $\varphi(x)$ is "x is well founded", so that w is the cumulative hierarchy, V; $\delta(x)$ is just the power set of x, P(x). If $x \subseteq V$, Transcendence and Closure hold simply in virtue of x being well-founded. The contradiction in this case is that $V \in V$ and $V \notin V$, since $\delta(V)$ is just V itself. (Since all members of V are well-founded $P(V) \subseteq V$. But if $x \in V$ then every member of x is in V. Hence $V \subseteq P(V)$.)

28 Graham Priest

We can tabulate these observations as follows:

Paradox	$\delta(x)$	$\varphi(x)$	w
Russell	id(x)	x∉x	R
Burali-Forti	$\log(x)$	x is an ordinal	On
Mirimanoff	P(x)	x is well-founded	V

3. Definability paradoxes

Let us now move on to the paradoxes in Ramsey's Group Bi, the definability paradoxes, and recall that something is definable iff there is a (non-indexical) nounphrase that refers to it. Now consider König's Paradox as an example: there are indefinable ordinals, and the least indefinable ordinal has just been defined in that very phrase. At first glance, this would appear to fit into Russell's Schema very nicely. $\varphi(x)$ is the predicate "x is a definable ordinal", so that w is the set of definable ordinals, DOn; and $\delta(x)$ is $\operatorname{lon}(x)$ (the least ordinal not in x). The contradiction is that the least indefinable ordinal $\operatorname{lon}(\operatorname{DOn})$ both is and is not definable (that is, a member of DOn).

Unfortunately, things are not so simple. For whilst Transcendence is true, $lon(x) \notin x$, Closure need not be: even if every member of x is definable, there is no reason to suppose that x itself is definable. And if it is not, there is no reason to suppose that the least ordinal not in x is definable, i.e., that $lon(x) \in DOn$.

If, however, x is definable, then we do have that $lon(x) \in DOn$; and this suggests a small modification of Russell's conditions, as follows: For given properties φ and ψ , and (possibly partial) function δ :

- 1) $w = \{x : \varphi(x)\}$ exists and $\psi(w)$
- 2) if x is a subset of w such that $\psi(x)$:

a)
$$\delta(x) \notin x$$

and b)
$$\delta(x) \in w$$

If these conditions are satisfied we still have a contradiction. For since $\psi(w)$, we have both $\delta(w) \notin w$ and $\delta(w) \in w$. I will call this the *Qualified Russell's Schema*. The essential qualification is that applying δ to an arbitrary subset of w is not now guaranteed to transcend it; only if the set satisfies ψ will it do so. Thus δ works on a *sub-family* of the power set of w, P(w).

Bearing this in mind, it is clear that Russell's Schema proper is just a special case of the Qualified Russell Schema, where the sub-family in question is P(w) itself (i.e., where ψ is the universal property, $\lambda xx=x$). Thus, the Qualified Russell Schema still encompasses the paradoxes of Group A. But by choosing ψ appropriately, it also encompasses König's Paradox. We just take the property $\psi(x)$ to

be "x is definable". DOn is clearly definable, and as we have already observed, for definable x, Transcendence and Closure hold.

Unsurprisingly, all the other paradoxes in Group Bi fit the Qualified Russell Schema. Take Berry's Paradox. $\varphi(x)$ is "x is a natural number (= finite ordinal) definable in less than 19 words", so $w=\{n; n \text{ is a natural number definable in less than 19 words}\}$, DN_{19} ; $\psi(x)$ is "x is definable in less than 14 words"; and $\delta(x)$ is just lon(x). Clearly, $lon(x)\notin x$, and if x is definable in less than 14 words $lon(x)\in DN_{19}$. (Its definition is "the least ordinal not in" plus the definition of x.) Hence Transcendence and Closure are satisfied. The contradiction is that $lon(DN_{19})\in DN_{19}$ and $lon(DN_{19})\notin DN_{19}$.

One further example: Richard's Paradox. In this, $\varphi(x)$ is "x is a definable real number between 0 and 1", so that w is the set of such reals, DR; $\psi(x)$ is "x is definable" and $\delta(x)$ is diag(x), a real, defined by diagonalisation on x, in such a way as to ensure that $\delta(x)\notin x$. If x is definable, diag(x) is definable. Hence both Transcendence and Closure are satisfied. The contradiction is that diag $(DR)\in DR$ and diag $(DR)\notin DR$.

We can tabulate these observations as follows:

Paradox	$\delta(x)$	$\psi(x)$	$\varphi(x)$	w
König's	lon(x)	x is definable	x is a definable ordinal	DOn
Berry's	lon(x)	x is definable in less than 14 words	x is a natural number definable in less than 19 words	DN ₁₉
Richard's	diag(x)	x is definable	x is a definable real between 0 and 1	DR

We have seen that all the paradoxes of Groups A and Bi fit the Qualified Russell Schema. In fact, they all fit the Russell Schema itself. To see this it suffices to note that (assuming the Axiom of Choice) for any structure that fits the Qualified Schema there is another, identical in all relevant respects, that fits the regular Schema.

Let φ , ψ and δ be any quantities satisfying the Qualified Russell Schema. As usual, $w = \{x: \varphi(x)\}$. Let Θ be a choice function on P(w). We now define a new function, δ' , as follows. Let $x \subseteq w$.

$$\delta'(x) = \delta(x) \text{ if } \psi(x);$$

$$\delta'(x) = \Theta(w-x)$$
 otherwise

It is easily seen that φ and δ' satisfy the Russell Schema. The only point to note is that $\Theta(w-x)$ is undefined if w=x. But since $\psi(w)$, $\delta'(x)$ is simply $\delta(x)$ in this case.

For the same reason, the contradictions $\delta(w) \in w \land \delta(w) \notin w$ and $\delta'(w) \in w \land \delta'(w) \notin w$ are identical.

4. The Liar, etc.

We have seen that Russell's Schema proper underlies the paradoxes in Ramsey's Groups A and Bi. It remains to show that it underlies the paradoxes in Group Bii. The Liar Paradox is a paradigm of this Group. *Prima facie* this paradox appears to have nothing to do with the Russell Schema: the Schema concerns totalities which can be transcended by means of the application of a suitable operator; the Liar Paradox does not seem to involve totalities at all.

However, with a bit of care, it can be put into the appropriate form. In fact, for reasons that will become clear later, it is safer to show that it fits the Qualified Russell Schema, so let me show this. Assume for the moment that it is sentences (rather than propositions or some other semantic entities) that are true or false. Let $\varphi(x)$ be "x is true", so that w is the set of true sentences, Tr; let $\psi(x)$ be "x is definable". δ is a function, σ , defined by some suitable technique of diagonalisation so that if a is any definable set $\sigma(a) = \alpha$ where $\alpha = \langle \alpha \neq a \rangle$. The angle-bracket expression is a sentence expressing the fact that α is not in the set a. (Hence, α says "This sentence is not in a".) Note that a must be definable, or there would be no guarantee that there is such a sentence.

Now, if a is definable and $a\subseteq Tr$:

$$o(a) \in a \implies <\alpha \notin a > \in a$$

$$\Rightarrow <\alpha \notin a > \in Tr$$

$$\Rightarrow \alpha \notin a \qquad \text{(by the T-schema)}$$

$$\Rightarrow o(a) \notin a$$

Hence $o(a) \notin a$, and Transcendence is satisfied. Moreover, it follows that $o \notin a$, and hence by the T-schema $o \notin a$. Tr., i.e. $o(a) \in Tr$. Hence, Closure is satisfied too. The Liar is the sentence o(Tr) and the contradiction is that $o(Tr) \in Tr$ and $o(Tr) \notin Tr$.

Another paradox in this family is the Knower Paradox. This is the same as the Liar, except that $\varphi(x)$ is "x is known to be true", and so w is the set of known things, Kn. Transcendence is verified as before, since knowledge implies truth; Closure follows, since it has been established that $\alpha \notin a$, and so $\alpha \in Kn$. The paradox is that $\sigma(Kn) \in Kn$ and $\sigma(Kn) \notin Kn$.

A variant on these paradoxes is the liar (or knower) chain, where several sentences are inter-related. These fit the pattern, too. To see this, just consider the simplest paradox of this kind: the Liar pair. This is exactly the same as the Liar, except that σ is slightly more complicated. This time, $\sigma(a) = \alpha$, where $\alpha = \langle \beta \in \text{Tr} \rangle$ and $\beta = \langle \alpha \notin a \rangle$. To see that σ satisfies the appropriate conditions, suppose that $\alpha \subset \text{Tr}$ and that a is definable. Then:

```
\sigma(a) \in a \qquad \Rightarrow \alpha \in a
\Rightarrow \alpha \in \text{Tr}
\Rightarrow \langle \beta \in \text{Tr} \rangle \in \text{Tr}
\Rightarrow \beta \in \text{Tr} \qquad \text{(by the T-schema)}
\Rightarrow \langle \alpha \notin a \rangle \in \text{Tr}
\Rightarrow \alpha \notin a \qquad \text{(by the T-schema)}
\Rightarrow \sigma(a) \notin a
```

Hence, $\sigma(a) \not\equiv a$, which is Transcendence. To see that Closure is satisfied, note the following:

$$\sigma(a) \notin a$$
 $\Rightarrow \alpha \notin a$
 $\Rightarrow \langle \alpha \notin a \rangle \in \text{Tr}$ (by the T-schema)
 $\Rightarrow \beta \in \text{Tr}$
 $\Rightarrow \langle \beta \in \text{Tr} \rangle \in \text{Tr}$ (by the T-schema)
 $\Rightarrow \alpha \in \text{Tr}$
 $\Rightarrow \sigma(a) \in \text{Tr}$

The contradiction is that $\sigma(Tr) \in Tr$ and $\sigma(Tr) \notin Tr$.

A final example from this Group of paradoxes will suffice. This is the Heterological Paradox. $\varphi(x)$ is " $\neg x$ sat x" where x is the satisfaction predicate, so $w=\{y:\neg y \text{ sat } y\}$, Het; $\psi(x)$ is "x is definable", as before; if a is a definable set, $\delta(a)=\langle v\in a\rangle$, where v is any new variable, and so this is an open sentence. Now, suppose that $a\subseteq \text{Het}$ and a is definable. Then:

$$< v \in a > \in a$$
 $\Rightarrow < v \in a > \in Het$
 $\Rightarrow \neg < v \in a > sat < v \in a >$
 $\Rightarrow < v \in a > \notin a$ (by the Satisfaction Schema)

Hence $\langle v \in a \rangle \notin a$, i.e., Transcendence. So by the Satisfaction Schema $\neg \langle v \in a \rangle$ sat $\langle v \in a \rangle$, i.e., $\langle v \in a \rangle \in$ Het, Closure. The paradox is that $\langle v \in$ Het $\rangle \notin$ Het.

We can record these observations in the following table:

Paradox	$\delta(x)$	$\psi(x)$	$\varphi(x)$	w
Liar	α , where $\alpha = \langle \alpha \notin x \rangle$	x is definable	x is true	Tr
Knower	α , where $\alpha = \langle \alpha \notin x \rangle$	x is definable	x is known	Kn
Liar Chain	α , where $\alpha = <\beta = \in Tr >$ where $<\beta > = <\alpha \notin x >$	x is definable	x is true	Tr
Heterological	< <i>v</i> ∈ <i>x</i> >	x is definable	$\neg x \text{ sat } x$	Het

Finally, note that nothing is changed essentially if we take propositions or some other semantic entities (not sentences) to be truth bearers; we just take $<\alpha \in x>$ to be the appropriate proposition. Similarly, in the Heterological Paradox, we take $\langle v \in x \rangle$ to be the appropriate property. In fact, things are actually simplified by this. For presumably, for any set (not just a definable one), x, there is a proposition, α , such that $\alpha = \langle \alpha \notin x \rangle$ (the proposition that this proposition is not in x). Similarly, in the Heterological paradox, if x is any set there is a property of being in x. In this case, one no longer needs to assume that x is definable; and the paradoxes fit the regular Russell Schema directly.

5. The Principle of Uniform Solution

We have now seen that all the paradoxes of self-reference, whether from Group A, Bi or Bii, fit the Russell Schema. The structure that this describes is, therefore, the structure that generates all the paradoxes. Russell was right: there is a single family here. What one should make of this demands a larger discussion than is possible here; but I will at least indicate what seems to me to be the single most important conclusion.

If two paradoxes are of different kinds, it is reasonable to expect them to have different kinds of solutions; on the other hand, if two paradoxes are of the same kind, then it is reasonable to expect them to have the same kind of solution. Generalising, it is natural to expect all the paradoxes of a single family to have a single kind of solution. Any solution that can handle only some members of the family is bound to appear somewhat one-eyed, and as not having got to grips with the fundamental issue. Of course, this observation puts a lot of weight on the notion of kind; to convince ourselves that two paradoxes are of the same kind we must convince ourselves (a) that there is a certain structure that produces contradiction and (b) that this structure is common to the paradoxes. Still, once kindredship has been established in this way the point seems undeniable. Let us call it the Principle of Uniform Solution (PUS, sorry): same kind of paradox, same kind of solution.

Post Ramsey, logicians have accepted that there are essentially two distinct families of paradox and, with an intuitive appreciation of the PUS, have, generally speaking, been content with two different kinds of solution. Typically, solutions to paradoxes of Group A are of the Zermelo Fraenkel kind and deny the existence of the totality w (Clause 1 of the Russell Schema). There is much less agreement about solutions to Group B paradoxes, but, typically, such solutions have denied principles such as the T-schema, or cognate principles, involved in establishing Transcendence and Closure. (All this is documented in Chs. 1 and 2 of Priest 1987.)

Whether or not each kind of solution is satisfactory within its own group (neither is), neither kind of solution gets a grip on paradoxes from the other group. Clearly, semantic principles such as the T-schema do not occur in the paradoxes of Group A. Totalities do occur in paradoxes of Group B. However, in most cases, these are totalities of a size whose existence is guaranteed by Zermelo Fraenkel set theory. Typically, they are countable sets (e.g., the set of definable ordinals in König's Paradox), but they can even be finite! (as in Berry's). Hence, the PUS, in conjunction with the main result of this paper, is sufficient to sink virtually all orthodox solutions to the paradoxes.

It is worth noting that some logicians have found Ramsey's categorisation problematic on the ground that the schema $\exists x \forall y (Rxy \Leftrightarrow \neg Rxx)$ underlies paradoxes in both of Ramsey's classes; again with an intuitive appreciation of the PUS, they have proposed faulting this as a solution to the paradoxes. (See, e.g., Martin 1977.) However adequate such solutions are, it should be noted that, as a general approach, this fails because there are many paradoxes, such as Berry's, that are simply not of this form. (See e.g. Priest 1983.)

6. Curry paradoxes

Let me conclude with a word on a kind of paradox that is related to some of the paradoxes I have been discussing, and which also cuts across Ramsey's division: Curry paradoxes. Some of the paradoxes I have discussed proceed by establishing a sentence of the form $\alpha \Leftrightarrow \neg \alpha$. (All the paradoxes in Group Bii do this and some of the paradoxes in Group A, notably Russell's, but not the paradoxes in Bi or the other paradoxes in Group A.) For each paradox of this kind, we can form a new paradox by replacing $\neg \alpha$ uniformly with $\alpha \rightarrow \beta$, where β is an arbitrary formula, or, more simply, with $\alpha \rightarrow \bot$, where \bot is some logical constant entailing everything. Using the Absorption principle $(\alpha \rightarrow (\alpha \rightarrow \beta) \vdash \alpha \rightarrow \beta)$ we can then infer $\alpha \rightarrow \bot$, and hence α , and hence \bot . (See Priest 1987, 6.2.)

Do such paradoxes fit the general scheme given here? Yes and no, depending on what \rightarrow is. If it is a material conditional then, in most logics, $\alpha \rightarrow \bot$ is logically equivalent to $\neg \alpha$, and so the curried version of each paradox is essentially the same as the uncurried form. If, on the other hand, \rightarrow is a non-material conditional (e.g., a strict conditional), then $\alpha \rightarrow \bot$ and $\neg \alpha$ are quite different notions, at least prima facie. (Evaluating the truth of the first at a world first requires a consideration of what is happening at other worlds; evaluating the truth of the second does not.) In this case, the curried versions of the paradoxes belong to a quite different family. One may therefore expect them to be solved in a different way, most obviously by the rejection of Absorption.

Department of Philosophy The University of Queensland Brisbane Queensland 4072 Australia **GRAHAM PRIEST**

¹ A version of this paper was given at the meeting of the Australian Association for

REFERENCES

- Gödel, K. 1944: "Russell's Mathematical Logic" in *The Philosophy of Bertrand Russell*, ed. P. Schilpp. Tudor Publishing Co. Reprinted in P. Benacerraf and H. Putnam (eds.), *Philosophy of Mathematics*. Oxford: Blackwell, 1964.
- Martin, R.L. 1977: "On a Puzzling Classical Validity". *Philosophical Review*, 8, pp. 454-73.
- Peano, G. 1906: "Additione E". Rivista di Mathematica, 8, pp. 143-57.
- Priest, G. 1983: "The Logical Paradoxes and the Law of Excluded Middle". *Philosophical Quarterly*, 33, pp. 160-5.
- ——1987: In Contradiction. The Hague: Martinns Nijhoff.
- Ramsey, F.P. 1925: "The Foundations of Mathematics". Proceedings of the London Mathematics Society, 25, pp. 338-84; reprinted in D. H. Mellor, ed. Foundations: Essays in Philosophy, Logic, Mathematics and Economics. London: Routledge and Kegan Paul, 1978.
- Russell, B. 1905: "On Some Difficulties in the Theory of Transfinite Numbers and Order Types". *Proceedings of the London Mathematical Society*, (series 2) 4, pp. 29-53; reprinted in D. Lackey, ed. *Essays in Analysis*. London: Allen and Unwin, 1973.

Logic in Sydney, August 1991, under the title "All the Paradoxes of Self-reference Have a Common Underlying Structure Generating the Contradiction", and abstracted under that title in the *Journal of Symbolic Logic*, 57, 1992, pp. 1501. I am grateful to Greg Restall for his comments.