Intuitive consequences of the Revision Theory of Truth

MICHAEL KREMER

Roy T. Cook (2002) offers a challenge to Gupta and Belnap’s book *The Revision Theory of Truth* (1993). Cook provides an example which, he claims, shows that Belnap and Gupta’s theory fails to meet the desideratum that ‘a definition should be evaluated only by how well it captures the material aspects of a notion’ (Gupta and Belnap 1993: 277, cited by Cook 2002: 17). Cook’s example involves four sentences $S_1$ through $S_4$ for which he claims the following properties:

(a) ‘In a (classical) language containing $S_1$ through $S_4$ plus the relevant Tarski biconditionals … we can prove that sentences $S_1$ and $S_2$ are true and sentences $S_3$ and $S_4$ are false.’ (2002: 18)

(b) ‘This assignment of truth values to $S_1$ through $S_4$ makes sense, and it is the only assignment to do so.’ (2002: 18)

(c) On Gupta and Belnap’s approach, ‘although in some cases … the revision function settles down to the “correct” assignment, it does not do so in all cases’. (2002: 19)

Cook argues that on Gupta and Belnap’s theory ‘these sentences are pathological’ (by (c)) whereas ‘intuitively … there is no problem with the truth values of these four sentences – $S_1$ and $S_2$ are true and $S_3$ and $S_4$ are false’ (by (a) and (b)). He concludes that ‘the Revision Theory of Truth does not satisfy the condition of material adequacy that Gupta and Belnap make so much of’ (2002: 18).
In this note I will argue that, quite the contrary, Gupta and Belnap’s theory delivers exactly the right analysis of Cook’s example. I will focus on a simplified version of Cook’s example, briefly indicating in an appendix how to apply the morals of my discussion to his more complex version. My example consists of just two sentences:

(S1) At least one of $S_1$ and $S_2$ is false.
(S2) Both of $S_1$ and $S_2$ are false.

An argument parallel to (but shorter than) Cook’s shows that

(a’) In a classical language containing $S_1$ and $S_2$ plus the relevant Tarski biconditionals

$S_1$ is true IFF at least one of $S_1$ and $S_2$ is false.
$S_2$ is true IFF both of $S_1$ and $S_2$ are false.

We can prove that $S_1$ is true and $S_2$ is false.

Proof: Assume that $S_2$ is true. Then both of $S_1$ and $S_2$ are false. This implies that $S_2$ is false. Contradiction, so by reductio $S_2$ is false. It follows that $S_1$ is true.

Further we can maintain:

(b’) ‘this assignment … makes sense, and it is the only assignment to do so’.

On the other hand, again paralleling Cook, we can observe that the revision rule for $S_1$ and $S_2$ is this:

$\delta(\{S_1, S_2\}) = \phi; \delta(\{S_1\}) = \{S_1\}; \delta(\{S_2\}) = \{S_1\}; \delta(\phi) = \{S_1, S_2\}.$

We then find:

(c’) on either of the two middle assignments ($[S_1], [S_2]$) ‘the revision function settles down to the “correct” assignment’ but on the other two assignments it does not.

With this simplified version of Cook’s argument in hand, let us consider more carefully the proof offered for (a’). It is certainly the case that in a classical language containing the Tarski biconditionals for $S_1$ and $S_2$ the argument for (a’) goes through. But a lesson of the liar paradox is that the sort of reasoning used in that argument has to be handled carefully. After all, the same sort of reasoning leads to a contradiction in a language containing the liar paradox. More to the point, precisely the reasoning used in Cook’s argument leads to analogous difficulties. For, in any classical language containing $S_2$, where $S_1$ is any sentence whatsoever, assuming just the Tarski biconditional for $S_2$, one can prove that $S_2$ is false and $S_1$ is true.

Proof: Assume that $S_2$ is true. Then both of $S_1$ and $S_2$ are false. This implies that $S_2$ is false. Contradiction, so by reductio $S_2$ is false.
However, if $S_1$ were false, then $S_2$ would be true. It follows that $S_1$ is true.

In this fashion we could ‘prove’ that Al Gore is president after all, or that $2 + 2 = 5$.

Moreover, exactly the same sort of reasoning shows that in any classical language containing $S_1$, where $S_2$ is any sentence whatsoever, assuming just the Tarski biconditional for $S_1$, one can prove that $S_2$ is false and $S_1$ is true.

Proof: Assume that $S_1$ is false. Then both of $S_1$ and $S_2$ are true. This implies that $S_1$ is true. Contradiction, so by reductio $S_1$ is true. However, if $S_2$ were true, then $S_1$ would be false. It follows that $S_2$ is false.

Again, in this fashion we could ‘prove’ that George W. Bush is not President after all, or that it is not the case that $2 + 2 = 4$.

The lesson of the paradoxes, then, is that one has to be careful in applying the sort of reasoning involved in the argument for $(a')$. Examining closely the reasoning of the above proofs we see that the conclusions, that $S_2$ must be false, and that $S_1$ must be true, are over-hasty; reliance on the Tarski biconditionals in drawing conclusions about the truth and falsity of sentences has to be carefully circumscribed in the case of sentences like $S_1$ and $S_2$.

In particular, when evaluating $S_2$, we have to consider several cases, depending on the status of $S_1$. If $S_1$ were unproblematically true, we could legitimately conclude that $S_2$ is false. But if $S_1$ were unproblematically false, then $S_2$ would be paradoxical, and the conclusion that it is false would be no more legitimate than the conclusion that it is true, since each could be derived from the other (and each would lead to contradiction). Further $S_1$ might somehow be itself problematic. In particular, if $S_1$ were paradoxical, we would again find that $S_2$ is paradoxical.

Similarly, the evaluation of $S_1$ depends on the status of $S_2$. If $S_2$ were unproblematically false, $S_1$ would be true; but if $S_2$ were unproblematically true, $S_1$ would be paradoxical, and similarly, if $S_2$ were paradoxical, $S_1$ would be paradoxical as well.

What these reflections show is that the argument for $(a')$, if meant to establish that ‘intuitively there is no problem with the truth values of these two sentences – $S_1$ is true and $S_2$ is false’, is circular. In order to employ the Tarski biconditional for $S_2$ legitimately to show that $S_2$ is false, we must already know that $S_1$ is true. But, in order to establish that $S_1$ is true, we must already know that $S_2$ is false.

None of this detracts from the truth of $(b')$: there is only one assignment of truth values to $S_1$ and $S_2$ which coheres with the Tarski biconditionals. But the above considerations do detract from the further claim that this
Assignment of truth values is ‘correct’ and that ‘intuitively, there is no problem with the truth values’ of $S_1$ and $S_2$. Intuitively, there is a problem with these sentences; each taken by itself is pathological since each taken by itself seems to determine the truth value of the other, regardless of the content of the other. Yet the pathological features of each sentence are cancelled by the pathological features of the other, so that there is exactly one coherent assignment of truth values to the two; taken together they exhibit a form of circular dependence such that this assignment of truth values cannot be justified in a non-question-begging manner.

Now let us compare the results of our analysis of (a’) and (b’) with the behaviour of $S_1$ and $S_2$ under the revision rule. First, under the revision rule, there are initial hypotheses for the extension of ‘true’ which yield as stably true $S_1$ and $\sim S_2$. Moreover, there are no initial hypotheses for the extension of ‘true’ which yield as stably true any other combination of $S_1$, $S_2$, $\sim S_1$, and $\sim S_2$. Thus the Revision Theory accommodates the correctness of (b’).

Second, there are initial hypotheses for the extension of ‘true’ under which $S_1$ and $S_2$ exhibit paradoxical behaviour, flipping back and forth indefinitely between being evaluated as true and being evaluated as false. Neither $S_1$ nor $\sim S_2$ is ‘valid’ in Gupta and Belnap’s sense (1993: 123). Thus, these sentences are treated by the Revision Theory as not simply ‘in order’ but pathologically interdependent. At the same time, they are not treated as simply paradoxical.

Third, the precise form of the pathological nature of $S_1$ and $S_2$ is made clear by the Revision Theory, and can’t really be made clear in any other way. Above, I noted that in evaluating $S_1$ we would need to take account of the status of $S_2$, and that in evaluating $S_2$ we would need to take account of the status of $S_1$. I mentioned the possibilities for each of being unproblematically true, unproblematically false, and paradoxical. But we have seen that in fact neither sentence fits into any of these categories – each is pathological without being paradoxical. There is no way to state the particular pathology of either sentence without reference to the other, for the status of each is inextricably bound up with the status of the other. What we can say about either is only to be said in what we can say about both, and what we can say about both is just four things.

First, if we begin by assuming that $S_1$ is true and $S_2$ is false, we will find our initial hypothesis confirmed – this corresponds to the case in which we argue circularly that $S_1$ is true, given that $S_2$ is false, and that $S_2$ is false, given that $S_1$ is true.

Second, if we begin with the apparently opposite hypothesis that $S_1$ is false and $S_2$ is true, we will find that we are brought around again to the self-confirming hypothesis that $S_1$ is true and $S_2$ is false. In this case, as in
the first, we have assumed that one but not both of our sentences is false, and so at the metalevel, as it were, we have assumed that $S_1$ is true and $S_2$ is false after all. Note that if we were somehow able to take $S_1$ to be fixed to be false, we would end up with the conclusion that $S_2$ is paradoxical; and if we were somehow able to take $S_2$ to be fixed to be true, we would end up with the conclusion that $S_1$ is paradoxical. But this we cannot do; the particular pathology of $S_1$ prevents us from taking it to be fixed to be false, just as the particular pathology of $S_2$ prevents us from taking it to be fixed to be true; and for just this reason we cannot take either sentence to be paradoxical either.

Third, though, if we begin by assuming that both sentences are false, we will be led to conclude that both are true; and, fourth, if we begin by assuming that both are true, we will be led to conclude that both are false; and so these two hypotheses each generate an endless cycle. Thus, while not paradoxical, the two sentences together are capable of exhibiting paradox-like behaviour. In evaluating either sentence we have to recognize its pathology as correlative to the pathology of the other.

I conclude that the Revision Theory of truth, far from yielding counter-intuitive consequences in cases like Cook’s, has exactly the consequences intuition requires – at least, when intuition has been appropriately refined by a minimum of reflection on the phenomenon of paradox. That Cook misses this point is, I think, connected to a fundamental misunderstanding of the Revision Theory which is revealed when he writes that Gupta and Belnap’s notion of validity ‘is meant to correspond to our intuitive notion of truth’ (2002: 17). To the contrary, Gupta and Belnap’s contention is that the Revision Theory provides a logical framework for exhibiting the behaviour of our intuitive concept of truth, itself a circular concept whose meaning is given by a rule of revision, which allows the calculation of the extension of the concept once a hypothesis has been made as to its extension. For such circular concepts no universal rule of application, determining the extension of the predicate in all circumstances, can be provided (Gupta and Belnap 1993: 118–19). The notion of ‘validity’, on the other hand, is part of the logical framework they develop for discussing such circular concepts. Given a language which contains predicates expressing circular concepts, for example a language containing its own truth predicate, the notion of validity is a precisely definable non-circular concept with a simple rule of application. Gupta and Belnap offer no replacement or surrogate for our intuitive notion of truth; what they purport to offer is a description of our intuitive notion of truth, albeit a description which lives within an idealized theoretical model of ordinary discourse, and is consequently in some ways idealized as well. What we have seen in our discussion of Cook’s supposed counter-example is a success of their descriptive effort, not its failure.
Appendix

In this appendix I will briefly indicate how an analogous analysis of Cook’s own example would go. Cook considers the four sentences:

(S1) At least one of S1 through S4 is false.
(S2) At least two of S1 through S4 are false.
(S3) At least three of S1 through S4 are false.
(S4) At least four of S1 through S4 are false.

Cook then argues for (a)–(c) above.

Inspection of the Tarski biconditionals for S1–S4 and calculation shows the following. (1) From the Tarski biconditional for S1 alone, one can argue that S1 is true and hence that at least one of S2–S4 is false — regardless of the content of S2–S4. (2) From the Tarski biconditional for S2 alone, one can argue that it is not the case that exactly one of the other sentences is false — regardless of the content of those other sentences. (If exactly one of the other sentences is false, then S2 is true if and only if it is false.) (3) From the Tarski biconditional for S3 alone, one can argue that it is not the case that exactly two of the other sentences are false — regardless of the content of those other sentences. (4) From the Tarski biconditional for S4 alone, one can argue that S4 is false, and hence that at least one of the other sentences is true — regardless of the content of those other sentences. Thus each of (S1)–(S4) exhibits a kind of pathology. It is only the mutual cancelling out of these pathologies which accounts for the existence of a unique distribution of truth values coherent with the Tarski biconditionals for S1–S4.

Moreover, the argument Cook gives for (a) is circular if taken as a justification of this distribution of truth values. The first stage of his argument goes like this:

Assume that S4 is true. Then at least four of S1 through S4 are false. This implies that S4 is false. Contradiction, so by reductio S4 is false.

It follows that S1 is true. (2002: 18)

But, to use the Tarski biconditional for S4 as a ground for claiming that S4 is false, one must already have ruled out the case that all of the other sentences are unproblematically false (a case in which S4 is paradoxical) as well as various cases in which some or all of the other sentences are problematic, for example paradoxical. That is, the argument, meant as a justification of the claim that S4 is false, presupposes that S1 is true, and cannot be used to establish it.

Similarly, the second phase of Cook’s argument is circular:

Now assume that S3 is true. Then at least three of S1 through S4 are false. We have already shown that S1 is true, so S3 would have to be false. Contradiction, so again by reductio S3 is false. It follows that S2 is true. (2002: 18)
But, to use the Tarski biconditional for $S_3$ as a ground for claiming that $S_3$ is false, given the truth of $S_1$, one must already have ruled out the case that both of the remaining sentences are unproblematically false (a case in which $S_3$ is paradoxical) as well as the various cases in which one or the other of the remaining sentences are problematic, for example paradoxical. Since it is also given at this point in the argument that $S_4$ is false, what has to be ruled out is that $S_2$ is either unproblematically false or somehow problematic. Therefore, the argument, meant as a justification of the claim that $S_3$ is false, presupposes that $S_2$ is true, and cannot be used to establish it.

Once again, I would argue, the Revision Theory gets things exactly right, counting Cook’s four sentences as pathological but not paradoxical, and accounting for the existence of a unique distribution of truth values coherent with the Tarski biconditionals as a by-product of the ways in which the pathologies of these sentences are fitted to cancel one another out. Once again, the precise nature of the joint pathology of the four sentences is made clear by the revision rule, and cannot really be made clear in any other way. This fits Belnap and Gupta’s claim not to be offering a replacement of our ordinary intuitive concept of truth, but rather to be describing its behaviour, in pathological as well as normal cases, as a circular concept.

References