The Paradox of the Prisoner in Logical Form[1]

by
A. N. Prior

The following is a simplified form of a paradox which has been circulating for a year or two[2]:

On a certain Saturday a judge sentenced a man to be hanged on Sunday or Monday at noon, stipulating at the same time that the man would not know the day of his hanging until the morning of the day itself. The condemned man argued that if he were hanged on Monday, he would be aware of the fact by noon on Sunday, and this would contravene the judge's stipulation. So the date of his hanging would have to be Sunday. Since, however, he had worked this out on Saturday, and so knew the date of his hanging the day before, the judge's stipulation was again contravened. The date, therefore, could not be Sunday either. The prisoner concluded that he would not be hanged at all. However, he was—on Sunday or Monday (it does not matter which), the announcement coming to him as a genuine surprise. The judge's sentence, and the accompanying stipulation, were thus both preserved; but what had been wrong with the prisoner's reasoning? [1][3]

Let us see what happens if we set out the prisoner's reasoning symbolically. Following Łukasiewicz [2][4], I shall use the symbols C and N for 'If' and 'Not', and will take over from the propositional calculus the following four laws (the principle of syllogism, two forms of the principle of transposition, and the principle of reduction ad absurdum):

1. CCpqCCqrCpr
2. CCpqCNqNp.
3. CCNpqCNqp.
4. CCpNpNp

I shall also use the following special symbols and forms:

---

[1] This article is edited by Lasse Burri Gram Hansen, Ulrik Sandborg-Petersen, and Peter Øhrstrøm. The original manuscript for this text from the Bodleian Library is undated. An earlier version of the text has been published in *Synthese* (2012) 188:411–416.

[2] It is unclear what version of the paradox Prior refers to. According to T.Y. Chow's paper "The surprise examination or unexpected hanging paradox" (American Mathematical Monthly, 105, 1998, 41–51) the first occurrence of the paradox was in 1943 as a surprise drill and the first version with an unexpected hanging was published by Quine in 1953. This was published in the January 1953 edition of *Mind*. In December 1953 a personal letter from Quine to Prior begins: "Mr. Bennett passed me your ingenious paradox". Related to the indication that this manuscript is written just before an emergence of a temporal logic, it is probably written in 1953 as a response to Quine's publication in *Mind*.

[3] [Prior's note 1:] In the original form of the paradox, the prisoner has seven days to think about, but his elimination of the remaining five depends on his elimination of the last two, and it is on this that we are in effect concentrating here.

[4] [Prior's note 2:] See, e.g., his Aristotles Syllogistic, Ch. IV.

's' for 'Sunday'
'm' for 'Monday'
'^t' for 'The day before t'
'Ht' for 'The prisoner will be hanged on the day t'
'Gtp' for 'The prisoner knows on the day t that p'.

'Sunday' is defined as 'The day before Monday', i.e. we have

Ds : s = ^m.

And we have the following special axioms:

5.	CNHsHm	('If the prisoner is not hanged on Sunday he will be on Monday').
6.	CHmNHs	('If the prisoner is hanged on Monday he will not be hanged on Sunday').
7.	CNHsGsNHs	('If the prisoner is not hanged on Sunday he will know on Sunday that he is not to be hanged on Sunday').
8.	CHt NG^tHt	('If the prisoner is hanged on the day t he will know on the day before t that he is to be hanged on the day t').

Here 5 expresses the judge's sentence; 6 expresses the fact that a man cannot be hanged twice over; 7 is a consequence of the fact that he is to be hanged at noon if at all on a given day; and 8 expresses the judge's stipulation. In deriving theorems we shall use the ordinary rules of substitution and detachment, together with these two special rules:

RGt: Cαβ →CGtαGtβ, i.e. if it is a law of the system that if α then β, then it is a law of the system that if the prisoner knows on the day t that α, he knows on the day t that β.

RG^s: α → G^sα, i.e. if it is a law of the system that α, then it is a law of the system that the prisoner knows on Saturday ('the day before Sunday') that α.

RG^s expresses the assumption that so long as the prisoner is alive he knows the conditions of his execution and the consequences of these conditions. From these premises, and with these rules, we may make the following deductions:

9.	CHmGsNHs.	1p/Hm, q/NHs, r/GsNHs = C6 – C7 – 9
10.	CGtNHsGtHm.	5[5] × RGt = 10
11.	CHmGsHm	1 p/Hm, q/GsNHs, r/GsHm = C9 – C10 t/s – 11
12.	CHmG^mHm	11 × Df.s = 12
13.	CG^tHtNHt	2[6] p/Ht, q/G^tHt = C8 – 13

[5] In the original document Prior points to the law 4—reductio ad absurdum—in deriving statement 10. Using reduction ad absurdum on the special rule *RG^s* however does not yield the result displayed in statement 10. Using the special axiom 5 instead does yield the result displayed in statement 10 and therefore we assume that this was the intended reference in the original text.

[6] In the original document Prior points to law 2 from the principle of transposition. However, using law 2 on the special axiom 8 does not yield the result displayed in statement 13. Using the law *CCpNqCqNp* derived from law 2 however does yield the result displayed in statement 13.We

14.	CHmNHm.	1p/Hm, q/G^mHm, r/NHm = C12 – C13 t/m – 14.
15.	NHm.	4p/Hm = C14 – 15.
16.	Hs.	3p/Hs, q/Hm = C5 – C15 16.
17.	G^sHs.	16 × RG^s = 17.
18.	NHs.	13 t/s = C17 – 18.

These conclusions, including 15 and 18 (in which the prisoner has a special interest), all certainly follow from the premises 1–8 by the rules assumed. The prisoner, however, chooses not to notice that the conclusions drawn comprise not only the pair 15 and 18 but also the pair 16 and 18, which directly contradict one another. In other words, the prisoner has certainly proved, from what he has assumed that he will not be hanged on Sunday (18), but he has just as certainly proved, also, that he will be hanged on Sunday (16). If the original informal account of his reasoning is re-examined, it will be found that this is so (note, in our first paragraph, the sentence, 'So the date of his hanging would have to be Sunday'), but when the thing is done informally it is possible to sweep past this and forget about it.

What follows from this is, of course, that the prisoner's initial assumptions are not mutually consistent, and at least one of them must be false. If we accept the story as told, the judge's stipulation that the prisoner would not know until the day that he was hanged that that was the day, turned out to be true; but the prisoner was mistaken in assuming that, if true, it was possible for him to know this to be true.[7] In relation to his knowledge this truth was rather like the Gödelian formula[8], stating in effect its own unprovability, in relation to the system in which it occurs. (If the Gödelian formula is true it is not provable, and if it is provable, not true, and the system not consistent). Or to take a simpler parallel, the prisoner's situation was not in the end so very different from what it would have been if the judge had said, "You will be hanged tomorrow, but you will not know until tomorrow whether you will be hanged or not". These two statements do not contain in themselves any inconsistency, and they might very well both turn out to be true, but they are inconsistent with the prisoner's knowing both to be true (i.e. with his being able to trust the judge). We could exhibit the last contradiction very simply as follows:

1.	Hs	
2.	NG^sHs	
3.	G^sHs.	1× RG^s = 3

Here 1 and 2 express the judge's statements, and RG^s the prisoner's knowledge of whatever is laid down, and the contradiction is between 2 and 3. But it is possible to eliminate peculiar rules like RG^s, and use only substitution and detachment, if we introduce the form "J ^sp" for "The judge asserts on Saturday that p", and express our premises as assertions that (i) the judge asserts that the prisoner will be hanged on Sunday, (ii) the judge asserts that the prisoner will not know this on

---

assume this is the intended reasoning by Prior who omitted to include this step in the text. This statement is used as premise 6 of the last proof on page 5.

[7] This is a very central passage as it underlines Priors preoccupation with knowledge and time.

[8] Prior finds that the paradox is very much like Gödel's incompleteness theorems. A similar remark has been made by Chow, who states that the investigations of this paradox over the years have "inspired an amazing variety of philosophical and mathematical investigations that have in turn uncovered links to Gödel's incompleteness theorems, game theory, and several other logical paradoxes".

Saturday, (iii) whatever the judge asserts is true, and (iv) whatever the judge asserts is known to be true by the prisoner. The contradiction then emerges as follows:

1. J ^sHs
2. J ^sNG^sHs
3. CJ^spp
4. CJ^spG^sp
5. NG^sHs 3 p/NG^sHs = C2–5
6. G^sHs 4 p/Hs = C1 – 6

We could also depart still further from our first procedure, and eliminate false assertions as well as false rules from our premises, if we content ourselves with proving the principle that if the judge asserts on Saturday both that the prisoner will be hanged on Sunday and that the prisoner will not know this on Saturday, then it cannot be that both whatever the judge asserts on Saturday is true, and whatever the judge asserts on Saturday is known to be true by the prisoner on Saturday. Using "Kpq" for "Both p and q" and "Πpϕp" for "For all p, ϕp", this assertion may be symbolised as

CK J^sHs J^sNG^sHsNKΠpC J ^sppΠpC J ^spG^sp

This is the principle which is contravened when we simultaneously assert the premises of the last proof. Its own proof is as follows:

1. CΠpϕpϕq
2. CCpqCCrsCKprKqs
3. CKCpqCrsCKrpKsq
4. CCqKpNpNq
5. CCpqCCqrCCrsCps
6. CCpNqCqNp
   2 p/ΠpC J ^spp, q/CJ^sNG^sHsNG^sHs, r/ΠpC J ^spG^sp,
   s/CJ^sHsG^sHs
   = C1 ϕ/CJ^s" q/NG^sHs – C1 ϕ/CJ^s'G^s', q/Hs – 7[9]
7. CKΠpC J ^sppΠpC J ^spG^spKCJ^sNG^sHsNG^sHsCJ^sHsG^sHs
   5 p/KΠpC J ^sppΠpC J ^spG^sp, q/KCJ^sNG^sHsNG^sHs
   -CJ^sHsG^sHs,
   r/ CK J^sHs J^sNGsHsKG^sHsNG^sHs, s/NK J^sHs J^sNGsHs
   = C7 – C3 p/J ^sNG^sHs, q/NG^sHs, r/J ^sHs, s/G^sHs
   – C4 p/G^sHs, q/K J^sHs J^sNG^sHs – 8
8. CK_pC J ^spp_pC J ^spG^spNK J^sHs J^sNG^sHs
   6 p/KΠpC J ^sppΠpC J ^spG^sp, q/K J^sHs J^sNG^sHs – 9
9. CK J^sHs J^sNG^sHsNKΠpC J ^sppΠpC J ^spG^sp

This is a little involved (though not as bad as it looks—the proof of 8 for example, simply amounts to this: by 7, the long substitute for p in 5 implies the substitute for q; by 3, this in turn implies the substitute for r ; and by 4, this in turn implies the substitute for s; hence, by 5, the substitute for p

[9] Here 1. is used, which means that we have to specify a function, ϕ. This is done twice in 6. Thus the primes ' are placeholders for the variables of the functions chosen.

implies the substitute for s, this implication being 8); but it is important to show that 9 is deducible, by substitution and detachment alone, from the premises 1–6. For these premises are entirely drawn from the propositional calculus and the theory of quantification; not one of them expresses any special properties of knowledge or time, and in fact our special symbols J , G, etc. do not occur in them, and the whole proof could be carried through with our special proposition Hs replaced by a propositional variable, capable of representing any proposition at all, and our special operators J's and G's replaced by variables δ and γ , capable of representing any operators whatever which form statements out of statements. That is, we could replace 9 above by

9'. CKδqδNγqNKΠpCδppΠpCδpγ p

which contains no symbols of fixed meaning but those of "if", "and", "all" and "not". So that nothing but substitution in the ordinary laws of propositional logic and quantification theory is required to prove that if the judge makes the assertions mentioned, then it cannot be that both whatever the judge asserts is true and whatever he asserts is known to be true by the prisoner. It would by similarly possible to prove, using propositional calculus and quantification theory alone, that if, in our original paradox, the judge lays down all that we suppose him to lay down, and we ascribe to him also whatever he admits to follow from what he says, and to the prisoner the knowledge of whatever he knows to follow from what he knows, it cannot be that both whatever the judge asserts is true and whatever he asserts is known to be true by the prisoner. Even to formulate this assertion would, however, be a formidable task, and its proof tedious; it is sufficient to have shown in our earlier manner that a contradiction arises if we violate the principle just sketched by simultaneously laying down our original 5–8, RGt and RG^s, and to have shown in the simpler case how this procedure can be replaced by a formal proof of the principle contravened.