WOLFGANG LENZEN

Doxastic Logic and the Burge-Buridan-Paradox

I

Elsewhere[1] I have tried to show that Lemmon's modal system DE4 (cf. [2], section 4) represents a fairly reasonable logic of (maximally strong) belief, provided, of course, that the alethic modal operator □ is replaced by the doxastic operator B_a, to be read as '*a* (strongly) believes that'. In particular, the following counterparts of Lemmon's axioms D, 4, and E have been argued as being sound principles of doxastic logic:

(1) $B_ap \rightarrow \neg B_a\neg p$
(2) $B_ap \rightarrow B_aB_ap$[2]
(3) $\neg B_ap \rightarrow B_a\neg B_ap$.

Tyler Burg's recent reconstruction of some paradoxes from Buridans's *Sophisms on Meaning and Truth*, [1], would appear to refute DE4 as an acceptable logic of belief, since it proves the set {(1), (2), (3)} to be incompatible with the subsequent sentence

(*) $q = \neg B_aq$.

The purpose of this note is to discuss Burge-Buridan's interesting puzzles in order to understand whether they really challenge the adequacy of DE4 as a system of doxastic logic.

II

The inconsistency which might disturb a doxastic logician can be established as follows. Truth-functionally, either B_ap or $\neg B_ap$ must hold for arbitrary propositions p. Consider the above proposition q and assume that $\neg B_aq$; by means of principle (3), one obtains $B_a\neg B_aq$ which appears to be 'the same' (or 'equivalent') to B_aq in view of 'equation' (*); this conclusion refutes our initial assumption by *reductio ad absurdum*, and we may infer that B_aq. But, by (2) and (1), one also obtains B_aB_aq and $\neg B_a\neg B_aq$, the latter being 'equivalent' to $\neg B_aq$, again on account of (*); hence paradoxically *reductio ad absurdum* also proves B_aq to be false.

Now, this informal derivation cannot be reconstructed within any standard system of modal propositional logic (MPC), since it makes essential use of a deduction rule of the type

(**) $p = q \vdash B_ap \rightarrow B_aq$.

But the identity sign does not belong to the usual stock of logical constants in MPC; so neither (*) nor (**) are well-formed formulae of MPC.

To be sure, one might introduce a symbol $=_{df}$ as definitory identity (or equality) into the language of MPC. This would have the result that $p =_{df} q$ warrants the substitutivity of p for q both within and without modal contexts; but, clearly, any reasonable theory of definition will exclude the circular (*) from the field of admissible definitions.

Still, where ↔ denotes material equivalence, the expression

(4) $q \leftrightarrow \neg B_aq$

is not only a well-formed formula of MPC but also may quite well be true. But (4) in conjunction with the DE4-principles (1)−(3) fails to give rise to a paradox. The crucial

[1] Cf. [4] and [5], Section 3.1.
[2] Here and in (3) it is presupposed that the personal indices are interpreted as standard names. Otherwise *a*'s believing that *a* believes that *p* would not be equivalent to *a*'s believing that he himself believes that *p*. Cf. on this point [3], Section 4.1 and 5.5.

inferences from $B_a\neg B_a q$ to $B_a q$ and from $\neg B_a\neg B_a q$ to $\neg B_a q$ in the above informal deduction evidently presuppose that (4) not merely is true, but also is believed by *a* to be true. Once it is observed that doxastic DE4 contains

(5) $B_a(p\rightarrow q) \rightarrow (B_a p\rightarrow B_a q)$

as an additional axiom and also contains the analogue of the Rule of Necessitation, i.e.

(6) $p \vdash B_a p$

as a deduction rule, the reader may easily verify that {(1)-(6)} is perfectly consistent while {(1)-(3), (5), (6)} is incompatible with B_a(4). That is, sentence

(7) $B_a(q \leftrightarrow \neg B_a q)$

entails a contradiction within DE4.

The only conclusion to be drawn from this inconsistency, however, is that the negation of (7) is a doxastic theorem. So, what the Burge-Buridan-'paradox' tells us in the first instance is that it is doxastically impossible to believe, with respect to any proposition *q*, that (*q* if and only if one does not believe that *q*). Hence we may not speak of a genuine paradox here unless it can further be argued that (7) represents a doxastically possible situation.

III

In the second part of his paper, Burge tries to establish just such a possibility by means of the following two examples (the order of which has been inverted for systematic reasons):

> Surely a Cretan, not realizing that he is a Cretan in context *C*, could say or believe that everything said or believed by a Cretan in context *C* is not true (where all other statements or beliefs in such a context are in fact not true).

Secondly,

> ...suppose that Plato thinks that his friend in room 13 is considering the view that the forms are a figment of an overactive imagination. Plato then thinks to himself: I (Plato) do not subscribe to the thought being considered in room 13, Unfortunately, Plato has erred. He himself is in room 13, not his friend. Applying principles like [(1) and (2)] to this case leads to paradox. ([1], p. 30).

For a closer analysis of the former example, let Burge's Cretan be denoted by '*c*' and let '*C*' stand for the property of being 'a Cretan in context *C*'. Furthermore, to get a grasp of the rather involved situation, we will have to enrich our language of doxastic DE4 by introducing quantifiers, (*x*), (*Ex*), to range over persons and by expanding DE4 to a second order logic containing $(\forall\alpha)$, $(\exists\alpha)$ as quantifiers over propositions (α being a propositional variable). Simplifying Burge's assumptions about *c*'s belief, by cancelling the words 'say or' and 'said or' from the quoted text, and by neglecting the parenthesized clause, we thus obtain the following formalization

(8) $B_c((x)(\forall\alpha)(Cx \,\&\, B_x\alpha \rightarrow \neg\alpha))$.

Letting (9) be the formula

(9) $(x)(\forall\alpha)(Cx \,\&\, B_x\alpha \rightarrow \neg\alpha)$,

(8) may be abbreviated to B_c(9) which, admittedly, constitutes a consistent belief.

Now, within second order logic, (9) clearly entails

(10) $(x)(Cx \,\&\, B_x(9) \rightarrow \neg(9))$,

and, by means of a suitable predicate-logical principle, (10) further entails

(11) $Cc \,\&\, B_c(9) \rightarrow \neg(9)$.

Hence, in view of (5) and (6), *c* must believe the logical consequence (11) of his belief (9):

(12) $B_c(Cc \ \& \ B_c(9) \rightarrow \neg(9))$.

So if, in addition to *c*'s being in situation *C*, *c* would also know or believe *this*, i.e. if B_cCc, then we could further conclude that

(13) $B_c(B_c(9) \rightarrow \neg(9))$.

This formula, although not of the strict form (7), would indeed suffice for deriving a contradiction. However, it is just *c*'s 'not realizing that he is a Cretan in context *C*', $\neg B_cCc$, which rendered his former belief in (9), i.e. (8), consistent and which equally blocks the derivation of the inconsistent (13). Thus, Burge's doxastic variant of the Liar paradox fails to establish the possibility of (7).

IV

Let us now turn to the even more complicated case of Plato's beliefs. First consider the following variant of Burge's example: Plato thinks that his friend in room 13 is considering the view that the forms are a figment of an overactive imagination. Plato then thinks to himself: I (Plato) do not subscribe to *whatever thought is* being considered in room 13. Unfortunately, Plato has erred. He himself is in room 13, not his friend.

Letting 'Plato' and 'the philosopher in room 13' be abbreviated as '*d*' and '*e*', respectively, the core of Plato's considerations may then be formalized as

(14) $B_d(\forall\alpha)(B_e\alpha \rightarrow \neg\alpha)$.

The content of this belief,

(15) $(\forall\alpha)(B_e\alpha \rightarrow \neg\alpha)$,

entails, in view of the assumed identity $d = e$,

(16) $(\forall\alpha)(B_d\alpha \rightarrow \neg\alpha)$.

Applying the second order principle of universal instantiation, we obtain

(17) $\mathrm{B}_d(15) \rightarrow \neg(15)$.

Since (17) appears to be a logical consequence of (15), Plato might seem to have to believe that (17), because he is assumed to believe that (15):

(18) $B_d(B_d(15) \rightarrow \neg(15))$.

This immediate analogue of (13) really would lead to paradox, as Burge claimed. However, (17) is not a logical consequence of (15) alone, but only of the conjunction of (15) plus the identity statement '$d = e$'. Hence (18) could be legitimately derived from (14) (via (5) and (6)) only, if we had in addition that $B_d(d = e)$. But it is just Plato's lack of knowledge as to who the philosopher in room 13 is, which made his original belief consistent and which, again, blocks the derivation of inconsistencies. Thus, our first reconstruction of Burge's Plato-case does not pose any problem even for a second order logic of belief.

To do greater justice to the peculiarities of Burge's presentation of this example, one might be tempted to further strengthen the underlying formal language by admitting definite descriptions, $(\iota\alpha)$, for ('thoughts' formulated as) propositions, such that, e.g., $(\iota\alpha)B_d\alpha$ stands for the informal expression 'the proposition believed by Plato (at a certain time *t*)'. Just like the introduction of ordinary description terms $(\iota x)\phi x$ within predicate logic, this introduction presupposes that some sort of identity relation be defined for propositions. Unfortunately, we are not told what kind of laws should govern this relation to be denoted, say, by '≡', although it would seem to have been Burge's task to state some such laws. But perhaps we may attempt to derive inconsistencies without specifying such principles.

Plato's crucial belief may be taken to be, that the proposition believed by the philosopher in room 13 is false; formally, this might be rendered as

(19) $B_d\neg(\iota\alpha)B_e\alpha$.

Since Plato himself is e, the proposition believed by e might be held to be 'identical' with the proposition believed by Plato:

(20) $(\iota\alpha)B_e\alpha \equiv (\iota\alpha)B_d\alpha$.

Assuming that (20) guarantees the substitutivity of the two 'identicals' in belief-contexts, we might then infer

(21) $B_d\neg(\iota\alpha)B_d\alpha$.

Application of (2) and (1) thus yields $B_dB_d\neg(\iota\alpha)B_d\alpha$ and also

(22) $B_d\neg B_d(\iota\alpha)B_d\alpha$.

But it might be argued that the content of Plato's belief (22) is just what he believes (at that moment); hence we obtain

(23) $(\iota\alpha)B_d\alpha \equiv \neg B_d(\iota\alpha)B_d\alpha$,

which has the desired form (*) and which seems to confirm Burge's claim. However, the very same assumptions in support of (23) may also be used to establish a simpler paradox as follows. Since, according to (21), Plato is taken to believe that what he (then) believes is false, it is this very belief in his (d's) erroneous belief that he (then) believes. Thus we could equally argue from (21) that

(24) $(\iota\alpha)B_d\alpha \equiv \neg(\iota\alpha)B_d\alpha$.

Since this flat contradiction has been arrived at without the help of doxastic principles from {(1)-(3), (5), (6)}, it ought to be evident that the 'paradox' about Plato's belief(s) is not a paradox of doxastic logic, but rather one arising from the unclear notion of propositional 'identity' alone.

From an intuitive point of view, (21) does not follow from (19) at all. Although, loosely speaking, $d = e$ entails that the proposition believed by e be 'the same' as the proposition believed by d, this sort of 'sameness' is weak. It cannot support the inference from 'Plato believes that what e believes is false' to 'Plato believes that what d believes is false'.

To conclude, Burge-Buridan's paradoxes, interesting as they may be in their own domain, are not really relevant to doxastic logic.

Bibliography

[1] Burge, T., 'Buridan and epistemic paradox'. *Philosophical Studies* 34 (1978), pp. 21 - 35.

[2] Lemmon, E. J., *An Introduction to Modal Logic*, Oxford, 1977.

[3] Lenzen, W., *Recent Work in Epistemic Logic*, Amsterdam, 1978 (*Acta Philosophica Fennica* 30:1).

[4] Lenzen, W., 'On DE4 − The logic of conviction'. To appear in: *Studia Logica*.

[5] Lenzen, W., *Glauben, Wissen und Wahrscheinlichkeit*, Springer, Vienna (Library of Exact Philosophy), 1980.