DOXASTIC PARADOXES WITHOUT SELF-REFERENCE

Robert Charles Koons
Philosophy Department
University of Texas
Austin, TX 78712

Certain doxastic paradoxes (paradoxes analogous to the Paradox of the Liar but involving ideal belief instead of truth) demonstrate that some formal paradoxes cannot be avoided simply by limiting the expressiveness of one's formal language in order to exclude the very possibility of self-referential thoughts and beliefs. These non-self-referential paradoxes, moreover, should be of special interest to such investigators of rationality as game theorist, economists and cognitive psychologists, since they occur more frequently in the world beyond the Gödel-theorist's laboratory.

Both Richard Montague[1] and Richmond Thomason[2] have taken their discoveries of Liar-like paradoxes in certain epistemic and doxastic logics as compelling reason for representing such notions only in languages in which no pernicious self-reference is possible. This can be achieved by representing the relevant epistemic or doxastic notion by means of a sentential operator, rather than as a predicate of sentences (or of other entities with sentence-like structure).

Nicholas Asher and Hans Kamp,[3] and Donald Perlis,[4] have shown that this strategy (called the "intensionalist" approach) alone is not enough to block the construction of paradoxes. If the language contains a binary predicate representing the relation between sentences and the propositions they express (Asher and Kamp), or if it contains a substitution operator Sub(P,Q,A) which is provably equivalent to the result of substituting the wff Q for all but the last occurrence of wff A in wff P (Perlis), then doxastic paradoxes can be constructed in an intensionalist logic.

Nonetheless, the intensionalist can reasonably respond that banning such expressions is a small price to pay for the avoidance of inconsistency.

If, however, it can be shown that versions of the doxastic paradoxes exist which do not depend in any way upon pernicious self-reference, then the whole point of the intensionalist strategy will be undermined. The paradoxes will then have to be avoided or made innocuous in some other way, and they will no longer provide any reason for abandoning the syntactic or representational approach to the representation of the objects of belief. This is precisely the task I propose to take on in this paper.

I will construct below a version of Thomason's paradox of ideal or rationally justifiable belief by means of modal logic rather than by means of Gödel theory. In this version, the crucial expression, 'is rationally justifiable', will be a statement operator rather than a sentential predicate. Thus, the semantics of the resulting formal language can represent the objects of justification (the propositions) as sets of possible worlds (as in Kripke

[1] Montague [1966].
[2] Thomason [1980].
[3] Asher & Kamp [1986].
[4] Perlis [1987].

semantics for modal logics) rather than as sentences of the language itself. In such a modal logic, it is impossible to construct a self-referential statement which is provably equivalent with a statement saying that the original statement is not justifiable.

We can nonetheless generate a paradox if it is plausible that there is some epistemic situation and some sentence 'p' such that the proposition expressed by the biconditional '$(p \equiv \neg Jp)$' is justifiable in that situation and the proposition that the biconditional proposition is justifiable is also justifiable. Without Gödelian self-reference, we cannot claim that any such biconditional is provable in Peano arithmetic, but the paradoxicality of the doxastic paradoxes did not depend on that fact. It depended only on two facts: that the biconditional is justifiable, and that the claim that the biconditional is justifiable is also justifiable. If we can show that it is very plausible to think that in some situations these two conditions hold with respect to sentences which are not self-referential, then such situations will constitute doxastic paradoxes in intensional logic.

Thus, to construct the paradox of justifiable belief in modal operator logic, it suffices to show that in some situations and for some proposition p̲ the following two claims are true, where 'J' is a statement operator representing the rational justifiability of a statement in some specified "epistemic situation":

A1. $J(p \equiv \neg Jp)$
A2. $JJ(p \equiv \neg Jp)$

Given these two assumptions, we can derive a contradiction within an epistemic logic consisting of the following doxastic axiom schemata:

J1. $J\neg J\varphi \rightarrow \neg J\varphi$
J2. $J\varphi$, where φ is a logical axiom
J3. $J(\varphi \rightarrow \pi) \rightarrow (J\varphi \rightarrow J\pi)$
J4. $J\varphi$, where φ is an instance of J1-J3

A contradiction can be derived as follows:

1. $J(p \rightarrow \neg Jp)$ A1, J2, J3
2. $Jp \rightarrow J\neg Jp$ 1, J3
3. $J\neg Jp \rightarrow \neg Jp$ J1
4. $\neg Jp$ 2,3

5. $J\neg Jp$ A2, J4, J2, J3 [see lines 1-4][5]
6. $J(\neg Jp \rightarrow p)$ A1, J2, J3
7. Jp 5,6, J3

The schemata J1 through J4 are modifications of some of the schemata discussed by Montague and Thomason. They are substantially weaker than Montague's in that schema J1 is a special case of the analogue of Montague's schema (i), '$J\varphi \rightarrow \varphi$'. This corresponds to that fact that these schemata are meant to capture the properties of *justifiability of belief*, as opposed to *knowledge*. At the same time, I suggest that J1-J4 are a substantial improvement over the schemata discussed by Thomason as characterizing *ideal belief*. In particular, schema J1 is much more plausible as a principle of ideal or rational belief than are the principles of Thomason's which I omit: '$J\varphi \rightarrow JJ\varphi$' and '$J(J\varphi \rightarrow \varphi)$'.

In an article on the surprise quiz paradox, Doris Olin discussed the principle I call J2. She argued:

> It can never be reasonable to believe a proposition of the form '*p* and I am not now justified in believing *p*'. For if a person A is justified in believing a proposition, then he is not (epistemically) blameworthy for believing it. But if A is justified in believing that he is not justified in believing *p*, then he would be at fault in believing *p*. Hence, if A is justified in believing that he is not justified in believing *p*, then he is *not* justified in believing *p*.[6]

If one has overwhelmingly good reason for believing that acceptance of *p* is not ultimately justifiable in one's present epistemic situation, then that fact must undermine any reasons one has for accepting *p* itself. To believe that *p* is not ultimately justifiable in one's present epistemic situation is to believe that it is inconsistent or otherwise

[5] Given the presence of J2 and J3, schema J4 could be replaced by a necessitation rule: if φ follows (in the doxastic logic consisting of J1--J3) from a set of premises each member of which is justifiable, then infer '$J\varphi$'. Therefore, since '$\neg Jp$' follows from A1 in that logic (as shown by lines 1--4 above), and since A1 is justifiable (which is just what A2 says), this necessitation rule would allow us to infer line 5.

[6] Olin [1983].

not cotenable with data which is, by one's own lights, weightier than the data (if any) which supports or seems to support p. This realization should undermine one's confidence in any data supporting p.[7]

The other axiom schemata are equally unexceptionable. J2 and J3 simply ensure that the property of being rationally justifiable in a situation is closed under logical entailment. If you are persuaded by what Henry Kyburg has said against "conjunctivitis",[8] then read "Jφ" as saying that φ belongs to the corpus of subjectively certain propositions in the relevant situation. Even Kyburg admits that the conjunction of two subjectively certain propositions is itself subjectively certain.

Schema J4 guarantees that certain obviously true axioms of doxastic logic are rationally justifiable in the situation under consideration. There can be little doubt that if schemata J1 through J3 are rationally defensible, there must be a large and variegated class of epistemic situations in which every instance of these schemata are rationally justifiable.

It remains to be shown that there are situations in which assumptions A1 and A2 are intuitively true, for some proposition p. In order to demonstrate this, I will appeal to two epistemological principles:

> (I) When the evidence in some epistemic situation for every member of some consistent set S is stronger than the evidence for any statement inconsistent with S, then each proposition expressed by a member of S is justifiable in that situation.
>
> (II) There are epistemic situations in which statements of the following forms are mutually consistent and are each supported by evidence stronger than any evidence supporting any statement inconsistent with their conjunction:
>
> $p \equiv \neg Jp$
>
> $J(p \equiv \neg Jp)$

These two principles together imply A1 and A2, since principle II simply states that the two statements above meet all of the conditions of principle I for

[7] The occurrence of this principle establishes an interesting connection between the paradox of reflexive reasoning and both Moore's paradox and the surprise quiz or hangman paradox (See Olin, Ibid., and Wright and Sudbury [1977].

[8] Kyburg [1970].

justifiability. Thus, both '$p \equiv \neg Jp$' and '$J(p \equiv \neg Jp)$' are justifiable, which is exactly what A1 and A2 claim. I will first discuss the justification of principle II by constructing several scenarios exhibiting the relevant features.

PARADOXICAL SITUATIONS

The first scenario comes from a gedankenexperiment suggested by Gideon Schwartz.[9] Adam is playing a game of checkers for the stake of 100 dollars. Simultaneously, an ideal reasoner offers Adam 1000 dollars if Adam will behave irrationally during the game. For our purposes, we can define "behaving irrationally" as acting in such a way that it is not ultimately justifiable in one's epistemic situaion to think that one is acting optimally. (Perhaps this would be better described as "not acting rationally". If so, simply assume that Adam will receive 1000 dollars if and only if he does not act rationally.) If we assume that the ideal reasoner gives Adam 1000 dollars if and only if Adam does in fact act irrationally, so defined, then the description of the situation entails that Adam's manifestly acting so as to lose the checkers game is optimal if and only if it is not ultimately justifiable for him to think that his manifestly playing to lose the game is optimal.

There seems to be no reason why Adam could not be apprised of the situation. If he is, then he has maximal evidence in support of a proposition which could be represented by a sentence of the form '$(p \equiv \neg Jp)$', where 'p' represents the proposition that Adam's manifestly playing with the intention to lose is his optimal action, and where 'J' is relativized to Adam's epistemic situation (which is essentially the same as our own). If we assume that Adam has maximally strong evidence for the epistemic principle I above (perhaps it counts as maximally strong evidence for itself, if it is self-evident), then Adam, by reflecting on the fact that he has maximally strong evidence for the proposition expressed by '$(p \equiv \neg Jp)$' and that the biconditional is obviously consistent, can also come to have maximally strong evidence for the proposition: $J(p \equiv \neg Jp)$. Thus, the described situation is one of the sort required by principle II.

[9] In Gaifman [1983], pp. 150-151.

As another example, suppose 'J' is relativized to my actual epistemic situation. Let 'p' represent the proposition that I am "rationally humble" (that is, I would still be humble even if I believed everything which is rationally justifiable in my present situaton). Let us suppose that we understand the virtue of humility in such a way that, given my available data, I am rationally humble if and only if it is not rationally justifiable for me to accept that I am rationally humble. [I'm supposing that anyone who believes of himself that he possesses such an important virtue as humility lacks humility.] Thus, we have a true and well-supported claim of the form '$(\neg p \equiv Jp)$' and another scenario satisfying the conditions of principle II.

I will now turn to principle I:

> (I) When the evidence in some epistemic situation for every member of some consistent set S is stronger than the evidence for any statement inconsistent with S, then any proposition expressed by a member of S is rationally justifiable in that situation.

I think that this is a very plausible principle of epistemology. If I have very good evidence for a claim, and no evidence (or much weaker evidence) against it, then ideally I should accept it.

Nonetheless, it could be objected that I am simply making inconsistent demands on the notion of *ultimate justifiability*, since I am simultaneously claiming that schema J1 is also a plausible epistemological principle:

J1. $J\neg J\varphi \rightarrow \neg J\varphi$

Schema J1 seems to demand that exceptions be made to principle I: if I'm justified in accepting '$\neg J\varphi$', then I can't simultaneously be justified in accepting φ, no matter whether I have maximally strong evidence for both '$\neg J\varphi$' and φ, and despite the fact that the two are logically consistent with one another.

Principle I and schema J1, however, are consistent with one another if we suppose that it is impossible to have evidence simultaneously for both of two claims of the form φ and '$\neg J\varphi$' (where 'J' is relativized to one's own epistemic situation). Evidence for two claims so related is mutually antagonistic: evidence for the second undermines the evidential character of what would otherwise be evidence for the first. Anything that could really count as evidence for a claim of the form '$\neg J\varphi$' must be sufficient to undermine as evidence for φ anything available in that epistemic situation which would otherwise be overwhelming evidence for φ. Conversely, if there is clearly overwhelming evidence for φ available in the situation, relfection on that fact should constitute conclusive evidence against the claim that φ is not ultimately justifiable.

Finally, as an alternative to principles I and II, there is a third epistemological principle to which I can appeal, principle III:

> (III) If φ is justifiable in situation E and E' differs from E only in having more evidence for φ, then φ is justifiable in E'.

In the Schwartz "Adam" scenario, we assumed that Adam possesses a very weighty body of apparent evidence for the biconditional: playing to lose is optimal if and only if it is not justifiable in situation E to think that playing to lose is optimal [where 'E' is some self-referential description of Adam's epistemic situation]. Suppose that Nemo is in situation E*, which differs from E only in having slightly less evidence for this very same biconditional (that is, the one concerning situation E). Unlike situation E, situation E* is not self-referential. Consequently, we can derive no contradiction from the supposition that this biconditional is justifiable in E*. Since the weight of apparent evidence, by hypothesis, in E* favors the biconditional, we seem to be forced to admit that the biconditional is justifiable in E*. Then, by principle III, we are forced to admit that the biconditional is justifiable in E as well, leading to the paradox.

A PROBABILISTIC SOLUTION?

The doxastic paradoxes I have presented so far concern when it is rational to accept a proposition. It might be thought that the generation of the paradox depended on working with the black-and-white dichotomy of accepting/rejecting. One might hope that replacing this dichotomy with a scheme of degrees of belief (represented as conforming to the probability calculus) would dissolve the paradox, especially if we insist that all non-mathematical statements are always believed with a probability some finite distance both from one and from zero. In fact, a re-examination of the putatively paradoxical situations from this perspective does lead to a non-paradoxical solution, if self-reference via syntax is forbidden.

We can replace each of the principles used in generating the paradox of justifiability with the corresponding principles concerning rational probability instead of justifiable acceptability. The following two schemata are consequences of the probability calculus:

(B1) $[\, J\varphi = y \,\&\, J(\varphi \rightarrow \pi) = z \,] \rightarrow J\pi \geq y + z - 1$

(B2) $J\varphi + J\neg\varphi = 1$

"$J\varphi$" is a function-operator which, when applied to a statement φ, yields a real number between zero and one, inclusive, representing the rational probability of φ in the relevant epistemic situation.

We also need a principle expressing the relationship between second-order and first-order probabilities. I will occasionally refer to B3 as "Miller's principle", from an article by D. Miller.[10]

(B3) $J\varphi \geq x \cdot J(J\varphi \geq x)$

B3' is an equivalent formulation of Miller's principle:

(B3') $J\varphi \leq 1 - x \cdot J(J\varphi \leq 1 - x)$

We can replace'$\geq$' with '$>$' in B3, and '$\leq$' with '$<$' in B3', if $x < 1$ and > 0. In the case of the inequality '$J\varphi > 0$', we can appeal to the closely related principle B3*:

(B3*) If $J(J\varphi > 0) > 0$, then $J\varphi > 0$

The claim that B3 holds whenever the relevant probabilities are defined is simply the generalization of the schema J4: if φ is justified, then that φ is not justified is not justified. If we interpret 'π is justified' as 'the rational probability of $\neg\pi$ is zero', then J4 is simply an instance of B3*.

Van Fraassen has produced a Dutch Book argument in favor of the B3 principles.[11] The principle has also been endorsed by Haim Gaifman and Brian Skyrms.[12] I will briefly give the Dutch Book argument for principle B3*. Suppose that $J(J\varphi > 0) > 0$. Then the conditional probability $J(\ \varphi\ /\ (J\varphi > 0)\ )$ is defined. Suppose for contradiction that this conditional probability is equal to zero. Then the agent is vulnerable to a dutch book. He is willing to accept any bet against φ on the condition that $J\varphi > 0$, no matter how unfavorable the odds. If $J\varphi$ is equal to zero, then the agent can gain nothing from these conditional bets. If $J\varphi$ is greater than zero, then the agent must also be willing to accept some bet on φ. Since he also accepted a conditional bet against φ at worse odds, he is bound to suffer a net loss. Therefore, the rational agent sets the conditional probability at some level greater than zero. By definition of conditional probability, we have:

$$J(\varphi \,\&\, J\varphi > 0) \,/\, J(J\varphi > 0) > 0$$

[10] Miller [1966].

[11] Van Fraassen [1984].

[12] Gaifman [1986]; Skyrms [1986].

From this it clearly follows that $J(\varphi)$ is greater than zero, since if $J(\varphi)$ were zero, so would be $J(\varphi \mathbin{\&} J\varphi > 0)$.

Transposing Schwartz's "Adam" story into probabilistic terms, we must assign for Adam some rational probability to the two conditionals making up the biconditional: p if and only if $Jp = 0$ [where 'p' stands for 'manifestly trying to lose the game of checkers is optimal for Adam']. Since these are non-mathematical propositions, the solution I am now discussing insists that they be given a rational probability of one minus ε, for some finite, non-zero ε. We must also assign a rational probability for Adam of the statement expressed by '$J(Jp > 0 \rightarrow \neg p)$'. Once again, since this is a non-mathematical statement, it must have a rational probability of one minus δ, for some finite δ. We can now prove that Adam must give p a probability of between zero and δ plus ε. First, we can show by reductio that Adam must give p a probability greater than zero.

1. Assume $Jp = 0$	
2. $J\neg p = 1$	1, B2
3. $J(\neg p \rightarrow Jp > 0) = 1 - \varepsilon$	Assumption
4. $J(Jp > 0) > 1 - \varepsilon$	2, 3, B1
5. $Jp > 0$	4, B3*

Similarly, we can show that Jp is less than or equal to δ plus ε.

6. $J(3 \mathbin{\&} B1 \mathbin{\&} B2 \mathbin{\&} B3^{*}) = 1 - \delta$	Assumption
7. $J(Jp > 0) \geq 1 - \delta$	1-5, 6, B1
8. $J(Jp > 0 \rightarrow \neg p) = 1 - \varepsilon$	Assumption
9. $J\neg p \geq 1 - \delta - \varepsilon$	7, 8, B1
10. $Jp \leq \delta + \varepsilon$	9, B2

Note that if δ and ε were both equal to zero, the above argument would paradoxically show that Jp is both greater than and equal to zero. This proof also constitutes a paradox if δ and ε are both infinitesimals. Interpret '$Jp = 0$' to mean that Jp is infinitely close to 0, and interpret '$Jp > 0$' to mean that Jp is some finite distance from 0. All of the principles used, including B3*, remain clearly true under such an interpretation. We can prove both that Jp is infinitely close to 0 (that it is less than the sum of two infinitesimals) and that it is finite distance from 0.

This probabilistic analysis does provide one route of escape from the doxastic paradoxes. However, we must assess the price which has to be paid for it: we must assume that it is never rational to give any non-mathematical proposition a probability of exactly one or a probability infinitely close to one. I would argue that this assumption is unacceptable.

There is no reason to think that all empirical, non-mathematical propositions have rational probabilities a finite distance from one. David Lewis[13] and Ellery Eells[14] have argued to the contrary. Assuming that to give a proposition a probability of one is to be dogmatically committed to the unrevisability of that proposition, they have urged that it is always irrational to give an empirical proposition the status of unrevisability.

I would like to set aside the question of whether it is ever rational to treat an empirical proposition as unrevisable, since I would instead like to challenge the underlying assumption that having a probability of one entails being unrevisable. Lawrence Davis has also challenged this assumption:

> I have not firmly resolved never to change my mind about the proposition "Zeus will strike me dead unless I beg him to spare me within the next 30 seconds." I (think I) can even imagine evidence that would persuade me of its truth. Yet I simply do not consider it at all in deliberating about what to do. Nor is this a matter of assigning it a low priority. I assign it a zero probability. I have entertained the proposition (and so, now, have you), but I do not consider it at all in planning my actions (and nor will you, if you are rational).[15]

Such propositions should surely be given probabilities infinitely close to, even if not exactly identical with, zero.

The assumption that probabilistic certainty implies unrevisability is based, I think, on the assumption that all rational revision of probabilities is by conditionalization on the evidence. This implies that once a proposition acquires a probability of exactly one, its probability is no longer subject to rational revision, since the relevant

[13] David Lewis, "Causal Decision Theory", Australasian Journal of Philosophy 59(1981): 5-30.

[14] Ellery Eells, Rational Decision and Causality (Cambridge Univ. Press, 1982), 207-208.

[15] Lawrence Davis, "Is the Symmetry Argument Valid?", in Paradoxes of Rationality and Cooperation, ed. Richmond Campbell & Lanning Sowden (Vancouver, UBC Press, 1985), pp. 255-262; 257.

conditional probabilities are undefined. This problem can be met by insisting merely that empirically certain propositions be given probabilities some (possibly infinitesimal) distance from one. Even if revision by conditionalization is assumed, such propositions are rationally revisable. Yet, as we have seen, the mere possibility of such only-infinitesimally-dubitable propositions is sufficient to permit the generation of doxastic paradoxes.

Acknowledgements

I would like to thank Tyler Burge and Tony Martin for their help and encouragement.

References

N. Asher & H. Kamp, "The Knower's Paradox and Representational Theories of Attitudes", Theoretical Aspects of Reasoning about Knowledge, ed. J. Halpern. Los Angeles: 1986, Morgan Kaufman, 131-148.

L. Davis, "Is the Symmetry Argument Valid?", in Paradoxes of Rationality and Cooperation, ed. Richmond Campbell & Lanning Sowden (Vancouver, UBC Press, 1985), pp. 255-262; 257.

E. Eells, Rational Decision and Causality (Cambridge Univ. Press, 1982), 207-208.

H. Gaifman, "Infinity and Self-Applications, I", Erkenntnis 20 (1983): 131-155.

H. Gaifman, "A Theory of Higher Order Probabilities", Theoretical Aspects of Reasoning about Knowledge, ed. J. Y. Halpern (Morgan Kaufman, Los Altos, Calif., 1986), 275-292.

H. Kyburg, "Conjunctivitis", in Induction, Acceptance and Rational Belief, ed. M. Swain (Dordrecht, Reidel, 1970), pp. 55-82.

D. Lewis, "Causal Decision Theory", Australasian Journal of Philosophy 59(1981): 5-30.

D. Miller, "A Paradox of Information", British Journal for the Philosophy of Science 17 (1966).

R. Montague, "Syntactical treatments of modality, with corollaries on reflexion principles and finite axiomatizability," Acta Philosophica Fennica 16 (1963), 155-167.

D. Olin, "The Prediction Paradox Resolved", Philosophical Studies 44 (1983) 229.

D. Perlis, "Languages with self-reference II: Knowledge, belief and modality," 1987, Computer Science Dept., University of Maryland, College Park, Maryland, 1- 42.

B. Skyrms, "Higher Order Degrees of Belief", Prospects for Pragmatism, ed. D. H. Mellor, 1986.

R. Thomason, "A note on syntactical treatments of modality," Synthese 44 (1980), 391-395.

B. van Fraassen, "Belief and the Will", Journal of Philosophy 81(1984):231-256.

C. Wright and A. Sudbury, "The Paradox of the Unexpected Examination", Australasian Journal of Philosophy 55 (1977):41-58).