A Liar Paradox*

Richard G. Heck, Jr.

Brown University

It is widely supposed nowadays that, whatever the 'right' theory of truth may be, it needs to satisfy a principle sometimes known as 'transparency': Any sentence S must be replacable, *salva veritate*, with "S is true", and conversely.[1] But, whatever the merits of this principle, it is known to be incompatible with others we might also have wanted to affirm, for example, the law of excluded middle. For suppose there is a sentence Λ for which we have:

$$\Lambda \equiv \neg \mathrm{T}(\ulcorner \Lambda \urcorner) \tag{1}$$

Then we can argue as follows. Suppose Λ. By transparency, $\mathrm{T}(\ulcorner \Lambda \urcorner)$; so by (1), $\neg\Lambda$. If $\Lambda \vee \neg\Lambda$, however, then we have $\neg\Lambda$ either way, so $\neg\Lambda$. By (1) again, $\mathrm{T}(\ulcorner \Lambda \urcorner)$; hence Λ, by transparency. So $\Lambda \wedge \neg\Lambda$, a contradiction.

It is not obvious, however, that this argument cannot be resisted. We might try rejecting the use of proof by cases[2] or look carefully at how we are allowing ourselves to reason with the biconditional (1). But there is a better form of the argument. It begins, not with a sentence like Λ, but with a *term* λ for which we have:

$$\lambda = \ulcorner \neg \mathrm{T}(\lambda) \urcorner \tag{2}$$

Now we can argue as follows:

$\mathrm{T}(\lambda) \vee \neg \mathrm{T}(\lambda)$	Premise
$\mathrm{T}(\ulcorner \neg \mathrm{T}(\lambda) \urcorner) \vee \neg \mathrm{T}(\lambda)$	Identity

*Published in *Thought* 1 (2012), pp. 36–40.

[1] The theory of truth in Saul Kripke's "Outline of a Theory of Truth" (Kripke, 1975) was perhaps the first to satisfy this condition, which plays a crucial role in the more recent investigations of Hartry Field (2008).

[2] This could be replaced by *reductio*—we got $\neg\Lambda$ out of Λ, so we have $\neg\Lambda$—but then we don't need excluded middle.

$\neg\mathrm{T}(\lambda) \vee \neg\mathrm{T}(\lambda)$	Transparency
$\neg\mathrm{T}(\lambda)$	$p \vee p \vdash p$
$\mathrm{T}(\ulcorner\neg\mathrm{T}(\lambda)\urcorner)$	Transparency
$\mathrm{T}(\lambda)$	Identity
$\mathrm{T}(\lambda) \wedge \neg\mathrm{T}(\lambda)$	$\wedge+$

This argument uses very meagre logical resources. We are using Leibniz's Law, in a form allowing for the substitution of identicals; we are using the inference $p \vee p \vdash p$; and we are using conjunction introduction.[3] We are not assuming *anything* about negation, other than that $p \wedge \neg p$ is contradictory.

Of course, the argument depends crucially upon the existence of a term like λ. In the usual sort of setting, that is to say, where truth-theories are developed as extensions of arithmetical theories, the argument depends upon the availability of what is sometimes called the 'strong' form of the diagonal lemma. The strong form tells us that, for any formula $A(x)$, there is a term g_A such that we can prove:

$$g_A = \ulcorner A(g_A)\urcorner$$

and not just that there is a formula G_A for which we can prove:

$$G_A \equiv A(\ulcorner G_A\urcorner)$$

But, as I have argued elsewhere (Heck, 2007), the strong form, though less well-known, is what we need if we want to capture the structure of the informal reasoning that leads to the Liar paradox. One typically begins with the assumption that there is a self-referential sentence, the Liar, that says of itself that it is not true. The weaker form of the diagonal lemma does not give us such a sentence. It only gives us a formula Λ that is *provably equivalent* to a sentence that says of Λ that it is not true. Neither Λ nor $\neg\mathrm{T}(\ulcorner\Lambda\urcorner)$ refers to itself, and neither *says of itself* that it is not true. The strong form, on the other hand, does deliver a truly self-referential liar sentence. Since $\lambda = \ulcorner\neg\mathrm{T}(\lambda)\urcorner$, $\neg\mathrm{T}(\lambda)$ is

[3] We need conjunction introducion only to conclude that (1) and (2) imply a single sentence that is contradictory. If we say that a set of formulae is inconsistent if it implies both some sentence and its negation, we do not need it.

a sentence that really does refer to itself and really does say of itself that it is not true.[4]

It is very difficult to see, therefore, how any theory of truth that licenses transparency—or even the one direction of it, allowing use to replace S with $\mathbf{T}(\ulcorner S \urcorner)$, which is all we used—can validate excluded middle. But all theories known to me that are committed to transparency reject bivalence, so rejecting excluded middle isn't a great cost; indeed, without bivalence, excluded middle has no special plausibility.

The law of non-contradiction is a different matter, however. Abandoning bivalence, by itself, does not motivate a rejection of non-contradiction: The claim that no sentence is both true and false does not appear, *prima facie*, to be incompatible with the claim that not every sentence is either true nor false, and there are plenty of logics that deny bivalence but endorse non-contradiction. But we can reason much as we just did to show that transparency is incompatible with the law of non-contradiction:

$\neg[\mathbf{T}(\lambda) \wedge \neg\mathbf{T}(\lambda)]$	Premise
$\neg[\mathbf{T}(\ulcorner \neg\mathbf{T}(\lambda) \urcorner) \wedge \neg\mathbf{T}(\lambda)]$	Identity
$\neg[\neg\mathbf{T}(\lambda) \wedge \neg\mathbf{T}(\lambda)]$	Transparency
$\neg\neg\mathbf{T}(\lambda)$	$\neg(p \vee p) \vdash \neg p$
$\neg\mathbf{T}(\ulcorner \neg\mathbf{T}(\lambda) \urcorner)$	Transparency
$\neg\mathbf{T}(\lambda)$	Identity
$\neg\mathbf{T}(\lambda) \wedge \neg\neg\mathbf{T}(\lambda)$	$\wedge+$

This argument uses the same resources as the previous one, except that, instead of $p \vee p \vdash p$ it uses: $\neg(p \vee p) \vdash \neg p$. (Indeed, the steps of the argument are almost the same.) It therefore seems safe to say that no theory that licenses transparency—or, again, even the one direction of it—can validate the law of non-contradiction.[5]

[4] Moreover, the strong form of the diagonal lemma is typically what is needed when bivalence is not being assumed, as here. In Kripke's theory, for example, the biconditional delivered by the weaker version is paradoxical: It does not have a truth-value in any fixed point, since neither side has a truth-value in any fixed point. The strong form is naturally available in primitive recursive arithmetic and other arithmetical theories with a rich stock of functional expressions. It can also be made available, through trickery, in theories formulated in the more familiar language $\{0, S, +, \times\}$ (Heck, 2007, §3.3).

[5] Note how much the use of the strong diagonal lemma improves the situation here. If we have only (1), then we will of course still be able to extract a contradiction from $\neg(\Lambda \wedge \neg\Lambda)$, but we will need substantial logical resources to do so. For example, we might

Whether gaining transparency is worth abandoning the law of non-contradiction is not an issue I can hope to resolve here.[6] But, for what it's worth, I don't myself regard transparency as non-negotiable—it is of a piece with a sort of deflationism I find problematic (Heck, 2004)—and some theories of truth reject it. For example, it follows from the foregoing that supervaluational accounts of truth, such as the one defended by Vann McGee (1990), must always be incompatible with transparency. As I said, however, maybe giving up transparency is worth saving the law of non-contradiction. Still, there are closely related principles about truth that such views must reject—in fact, that *any* decent theory of truth must reject.

Consider the following schemata:

$$\neg(S \wedge \mathrm{T}(\ulcorner \neg S \urcorner)) \tag{3}$$

$$\neg(\neg S \wedge \neg\mathrm{T}(\ulcorner \neg S \urcorner)) \tag{4}$$

These obviously follow from non-contradiction, given transparency, but they seem intuitively compelling to me, even absent transparency. It cannot be *both* that snow is white *and* that "snow is not white" is true; it cannot be *both* that grass is not green *and* that "grass is not green" is not true.

Now consider the instance of (3) where S is replaced by $\mathrm{T}(\lambda)$:

$$\neg(\mathrm{T}(\lambda) \wedge \mathrm{T}(\ulcorner \neg\mathrm{T}(\lambda)\urcorner))) \tag{5}$$

By identity, $\neg(\mathrm{T}(\lambda) \wedge \mathrm{T}(\lambda))$, so $\neg\mathrm{T}(\lambda)$. So (5) implies $\neg\mathrm{T}(\lambda)$. Taking S to be $\mathrm{T}(\lambda)$ in (4) gives us:

$$\neg(\neg\mathrm{T}(\lambda) \wedge \neg\mathrm{T}(\ulcorner \neg\mathrm{T}(\lambda)\urcorner)) \tag{6}$$

By identity, $\neg(\neg\mathrm{T}(\lambda) \wedge \neg\mathrm{T}(\lambda))$, so $\neg\neg\mathrm{T}(\lambda)$. Thus, (6) entails $\neg\neg\mathrm{T}(\lambda)$. So (5) and (6) together entail a contradiction: $\neg\mathrm{T}(\lambda) \wedge \neg\neg\mathrm{T}(\lambda)$. This argument depends only upon Leibniz's Law, conjunction introduction, and the

appeal to the DeMorgan laws and double negation elimination to derive $\Lambda \vee \neg\Lambda$. But it seems clear we will need a lot more than we did with excluded middle.

[6] It would nowadays be a common move to insist that we should not assert that the Liar is not both true and false, but only reject the claim that it is both true and false. But then we have also to reject the claim that the Liar is *not* both true and false, since that claim leads to paradox. This is not in itself contradictory, but it does point to the fact that rejection is a very weak attitude (Shapiro, 2004), if it is intelligible at all.

inference: $\neg(p \wedge p) \vdash \neg p$.[7] So it is hard to see how *any* theory of truth can validate (5) and (6).

But I am tempted to draw an even broader lesson. The paradox I have just described is, in a sense, just another semantic paradox. But it is no less worthy of the name 'the Liar paradox' than any other paradox deriving from truth-theoretic principles. In particular, there is no good sense in which the Liar paradox must begin or depend upon the so-called T-scheme

$$S \equiv \mathbf{T}(\ulcorner S \urcorner)$$

or its modern replacements, be these the T-rules—allowing the inference from S to $\mathbf{T}(\ulcorner S \urcorner)$, etc—or the principle of transparency. It has been known for a long time now that many different sorts of truth-theoretic principles, all of them *prima facie* quite plausible, can give rise to paradox (Friedman and Sheard, 1987, 1988). What makes the paradox just described especially interesting is how weak the logical resources needed to generate it are. Moreover, (3) and (4) strike me as *more* compelling than the T-scheme, and they are logically weaker. Even given classical logic, which of course we have not been assuming, they together imply only a restricted form of the T-scheme: $\neg S \equiv \mathbf{T}(\ulcorner \neg S \urcorner)$. Intuitionistically, the best one can do is: $\neg S \equiv \neg\neg\mathbf{T}(\ulcorner \neg S \urcorner)$.

I would suggest, therefore, that the version of the Liar paradox that begins with (3) and (4) is the strongest yet formulated, both in the sense that it relies upon the fewer logical resources than any other and in the sense that the principles with which it begins are, intutively, weaker than those involved in other versions. But however that may be, this form of the Liar shows, it seems to me, that there can be no consistent resolution of the semantic paradoxes that does not involve abandoning truth-theoretic principles that should be every bit as dear to our hearts as the T-scheme once was. And that leads me, anyway, to be tempted to conclude that there can be no truly satisfying, consistent resolution of the Liar paradox.[8]

[7] No doubt, the inference $\neg(p \wedge p) \vdash \neg p$ would be proven by *reductio* in many presentations. But it is available in minimal logic, where *reductio* is not. We can think of negation in minimal logic as defined thus: $\neg A$ abbreviates: $A \rightarrow \Omega$, for some arbitrary (and possibly consistent) formula Ω. In this case, the inference becomes $p \wedge p \rightarrow \Omega \vdash p \rightarrow \Omega$, which requires no more than $p \rightarrow p \wedge p$ and the transitivity of the conditional, which are of course available in minimal logic.

[8] Thanks to J. C. Beall, Hartry Field, Michael Glanzberg, Øystein Linnebo, Michael Lynch, Graham Priest, Joshua Schechter, and Lionel Shapiro for discussions of this material, and to an anonymous referee whose suggestions led to substantial changes in

References

Field, H. (2008). *Saving Truth From Paradox*. Oxford University Press.

Friedman, H. and Sheard, M. (1987). 'An axiomatic approach to self-referential truth', *Annals of Pure and Applied Logic* 33: 1–21.

—— (1988). 'The disjunction and existence properties for axiomatic systems of truth', *Annals of Pure and Applied Logic* 40: 1–10.

Heck, R. G. (2004). 'Truth and disquotation', *Synthese* 142: 317–52.

—— (2007). 'Self-reference and the languages of arithmetic', *Philosophia Mathematica* 15: 1–29.

Kripke, S. (1975). 'Outline of a theory of truth', *Journal of Philosophy* 72: 690–716.

McGee, V. (1990). *Truth, Vagueness, and Paradox: An Essay on the Logic of Truth*, Indianapolis, ed. Hackett.

Shapiro, S. (2004). 'Simple truth, contradiction, and consistency', in G. Priest, *et al*. (eds.), *The Law of Non-Contradiction*. Oxford, Clarendon Press, 336–54.

the way the paper was organize.