Reaching transparent truth

Pablo Cobreros, Paul Égré, David Ripley, Robert van Rooij

Abstract

This paper presents and defends a way to add a transparent truth predicate to classical logic, such that $T\langle A\rangle$ and A are everywhere intersubstitutable, where all T-biconditionals hold, and where truth can be made compositional. A key feature of our framework, called STT (for Strict-Tolerant Truth), is that it supports a nontransitive relation of consequence. At the same time, it can be seen that the only failures of transitivity STT allows for arise in paradoxical cases.

1 Introduction

A *transparent* truth predicate T is one that, paired with some quotation device $\langle\ \rangle$, allows, for any wff A, for the claim $T\langle A\rangle$ to be substituted for A or vice versa, in all extensional contexts in all arguments without change in validity. This paper presents and defends a way to add a transparent truth predicate to classical logic, a way that builds on our earlier work on vagueness in [Cobreros et al., 2011b, Cobreros et al., 2011a]. A number of other authors have sought a transparent truth predicate, and reached it by weakening classical logic in various ways. The key advantage of our approach, from which a number of other advantages will follow, lies in its keeping to classical logic.

In §2, we present some of the usual reasons for desiring a transparent truth predicate. If you think transparency is a misguided desideratum, nothing in this section will convince you otherwise. However, we think many philosophers who would otherwise be interested in a transparent truth predicate have turned away from it because of the importance they assign to preserving classical logic. Since this paper will show that the two are compatible, we want to take the opportunity to briefly rehearse the reasons for wanting a transparent truth predicate, as well as to call attention to a few other key desiderata. §3 introduces our target logic, which we will call STT, and elaborates on its relation to T-free classical logic. §4 outlines a theory of paradoxical sentences based on STT. §5 considers the advantages of our approach, comparing it to a number of other approaches in the literature. Finally, §6 concludes.

2 Some desiderata for truth

Theories of the truth predicate differ on whether the latter should be seen as a 'thick' and structured concept, or whether it rather is to be viewed as a 'thin' and simple concept. The view we wish to investigate in this paper belongs to the latter family. On that view, the reason why truth should be transparent is related to the function of the truth predicate in natural language, namely to allow expressive generalizations ([Quine, 1970, Field, 2008, Beall, 2009]).

Truth is a generalization device insofar as it allows us to report that the conjunction of a set of sentences, or their disjunction, holds, without having to enumerate all sentences in the set, and even without having to know what sentences are in the set. For instance, if I accept the sentence (1) 'one of the things John said was true', and if it turns out that John said three things, then I must accept that the condition expressed by the disjunction of the three sentences said by John holds. For instance, if it turns out that (2) John said: 'Mary is 30 years old; Mary has a blue car; Mary works in a bank', I must accept (3) 'either Mary is 30 years old or Mary has a blue car, or Mary works in a bank'.

This is so because the last sentence is exactly equivalent to (4) 'either 'Mary is 30 years old' is true, or 'Mary has a blue car' is true, or 'Mary works in a bank' is true'. Thus, the equivalence between A and $T\langle A\rangle$ is what gets us from (1) and (2) to (3) via (4): as Quine famously put it, truth behaves as a *disquotation device* in the transition from (4) to (3). Conversely, it behaves as a device of *semantic ascent* in the transition from (3) to (4): assuming (2) and (3), in particular, we can only infer the generalization expressed in (1) via (4).

Theories of transparent truth postulate that the intersubstitutibility of A with $T\langle A\rangle$ captures this double function of the concept of truth in natural language, that is semantic ascent and disquotation. Although all theories of transparent truth to date agree on this requirement, they still differ on two further aspects of its articulation. The first concerns Tarski's T-biconditionals: $A \leftrightarrow T\langle A\rangle$. While Tarski's schema internalizes the very idea of transparency in the object-language by means of a conditional, the theory of [Kripke, 1975], for example, which is a theory of transparent truth, does not have the wherewithal to make it valid (because, in fact, it does not make conditionals of the form $A \rightarrow A$ valid in the first place). A second aspect concerns the interplay of the truth predicate with the logical vocabulary. On top of transparency, another natural requirement on truth is compositionality. Suppose John actually uttered (5) 'Mary has a blue car or Mary works in a bank'. By transparency, this sentence is true iff Mary has a blue car or Mary works in a bank, that is, again by transparency, iff 'Mary has a blue car' is true or 'Mary works in a bank' is true. More generally, a theory of transparent truth can be said to be compositional if it has the wherewithal to express generalizations such as 'for any sentences A and B, their disjunction is true iff A is true or B is true'. Again, however, this desideratum is not necessarily entailed by transparency, because it implies internalizing the effect of transparency within the theory.

In this paper, our aim is to propose a theory of transparent truth that can

be made to satisfy these two extra requirements on the truth predicate. We will offer a theory where truth is fully transparent, and in which the T-schema holds; and we will show that it can be extended to capture the compositional behavior of the truth predicate. (For this purpose, we will, as is customary, appeal to arithmetic coding to handle the syntactic functions and quantification over sentences that appear in the compositional principles.) The main challenge for such a project is posed by the paradoxes, and we will show how our approach handles them.

3 Trivalence and STT

A number of approaches to maintaining transparent truth have been tried in response to the well-known paradoxes that inevitably arise. Many of these (eg [Priest, 2006b, Kremer, 1988, Beall, 2009, Field, 2008]) are based in some way on the work in [Kripke, 1975], and our approach is no different. As such, this section first briefly reviews the so-called 'Kripke construction' and its upshots in §3.1, before proceeding to present our logical framework in §3.2. (Although we here present our logic model-theoretically, it is susceptible of a proof-theoretic treatment as well; see [Ripley, 2011a] for a three-sided sequent calculus, or [Ripley, 2011b] for a more traditional two-sided sequent calculus.)

3.1 Kripke-Kleene models

The Kripke construction starts from a classical model for a base language $\mathcal{L}$ without any truth predicate, and provides a way to generate a model for the language $\mathcal{L}^+$ that adds a transparent truth predicate T to $\mathcal{L}$. For our purposes here, the details of the construction are irrelevant, and we won't present them; what's important are the models it yields, and their relation to the base-language models. (For details of the construction, see [Kripke, 1975].)

Kripke's base-language models are three-valued models for $\mathcal{L}$ using the set $\{1, \frac{1}{2}, 0\}$ of values, with Kleene's strong valuation schema.[1] According to this schema, negation maps 1 to 0, 0 to 1, and $\frac{1}{2}$ onto itself; conjunction $\wedge$ is defined as the minimum of the values of the conjuncts, and universal quantification $\forall$ as the minimum of values over all assignments that differ at most on the value they assign to the variable bound by the quantifier [Kleene, 1952]. We can define disjunction $\vee$, material conditional $\supset$, material biconditional $\equiv$ and an existential quantifier $\exists$ as usual. We also include constants $\top$ and $\bot$, which are required on every model to take values 1 and 0 respectively.

The problems with truth mentioned above become acute only when the language in question has some way of talking about itself. For the bulk of this paper, we do this on the cheap, supposing that $\mathcal{L}$ includes a quote-name-forming

[1] Actually, Kripke considers the case where the value $\frac{1}{2}$ is unused for anything in the base language; these are then classical models. As he points out (his fn. 20), this restriction plays no role, and we drop it here.

operator $\langle\ \rangle$ such that $\langle A\rangle$ is always a name of A, for any wff A of $\mathcal{L}^+$.[2] (In §3.4, we will be concerned to discuss a full theory of syntax, and will there temporarily manage self-reference via Gödel coding.)

The models generated by the construction are also strong Kleene models, with the additional feature that the value assigned to an atomic sentence $T\langle A\rangle$ is always the same as the value assigned to A itself. Call any model with these features a *KK model* (for 'Kleene-Kripke').[3]

The models produced by this construction have two main features that make them interesting for our purposes: they are *conservative* and they are *transparent*. Conservativeness first. For any model M of $\mathcal{L}$, the model M^+ of $\mathcal{L}^+$ produced by this construction agrees with M in its interpretations on the entire language $\mathcal{L}$. This includes cases in which M interprets $\mathcal{L}$ fully classically; in these cases, so too will M^+. All the usual paradoxical sentences can be formulated, due to the presence of $\langle\ \rangle$. For example, we might have a sentence λ that is $\neg T\langle\lambda\rangle$. This is no impediment to the construction.[4] Moreover, M can be very rich indeed; its predicates and terms can be interpreted in any way whatsoever. Since M^+ agrees with M on $\mathcal{L}$, the addition of T can be seen to have no effect on the T-free fragment of the language.

The resulting models are also transparent: they assign A and $T\langle A\rangle$ the same value, for every A. If we use KK models to define a notion of consequence, that notion of consequence will feature transparent truth, for the simple reason that no KK model can assign a formula A a different value from $T\langle A\rangle$. So long as all connectives are value-functional, and validity itself depends only on values taken by formulas on KK models, this result will hold.

There is an important question left to be answered, though: how are we to define a notion of consequence on KK models? We can understand logical consequence as usual, as absence of countermodel. The question then amounts to: what is a countermodel to an argument? Classically, a countermodel to an argument from premises Γ to conclusions Δ is a model that assigns 1 to every member of Γ and 0 to every member of Δ. There are multiple ways to extend this notion to three-valued KK models.

Some of these ways result in relatively familiar logics. One way, resulting in the logic we'll call K3T (for K3 with transparent truth), is to take a countermodel to be a model that assigns 1 to every member of Γ and *some value less than 1* to every member of Δ. Another way, resulting in the logic we'll call LPT (for LP with transparent truth), is to take a countermodel to be a model that assigns *some value greater than 0* to every member of Γ and 0 to every member

[2] To define a naïve satisfaction predicate, we should allow as well that the model features some scheme for encoding finite sequences of members of the domain into members of the domain. We won't worry more about satisfaction here, as the relevant features of the Kripke construction are already present with truth. Satisfaction poses no additional problems.

[3] KK models are thus [Kripke, 1975]'s fixed points. Every fixed point—minimal, maximal, intrinsic, and otherwise—is a KK model. Every KK model is a Kripkean fixed point as well, so long as we remember not to impose Kripke's restriction to classical base models.

[4] Our presentation of $\mathcal{L}$ and $\mathcal{L}^+$ does not guarantee that there will be such a sentence; but neither does it guarantee that there will not be. We assume, for our purposes in this paper, that there will be a liar sentence, a Curry sentence, and any other sort of paradoxical sentence.

of Δ. A third way, resulting in the logic we'll call S3T (for S3 with transparent truth), is to take a countermodel to be a model on which the minimum value assigned to the Γs is greater than the maximum value assigned to the Δs. (An argument is S3T valid, then, iff it is both K3T valid and LPT valid.) Note that all three of these definitions become equivalent to each other, and to the usual classical definition, if we restrict ourselves to two-valued classical models.

These logics (particularly K3T and LPT) are familiar in the literature on transparent truth, but they are not much advocated for. The main reason is their relative weakness. All three are considerably weaker than classical logic, but, more importantly, they lose many intuitively plausible and useful inference forms. For example, K3T does not validate excluded middle ($\vDash A \vee \neg A$) or, equivalently, identity ($\vDash A \supset A$), LPT does not validate material modus ponens ($A, A \supset B \vDash B$), and S3T validates none of these. As a result, most authors who work with variations on these logics (such as [Field, 2008, Priest, 2006b, Beall, 2009]) vary them by adding extra connectives that recover some of the strength these systems give up.[5]

Here, though, we will consider a different way of using KK models to define a usably strong logic. We will add no extra connectives, staying fully within the usual classical logical vocabulary. Instead, we will define validity differently.

3.2 The logic STT

The definition we consider stays very close to the familiar classical definition. We say a model is an ST countermodel to an argument from premises Γ to conclusions Δ iff the model assigns 1 to every member of Γ and 0 to every member of Δ. The logic STT (for ST with transparent truth) is the logic that results from this definition over KK models.[6]

It is immediate that STT is stronger than both K3T and LPT: any STT countermodel is automatically both a K3T and an LPT countermodel, but there are K3T and LPT countermodels that are not STT countermodels.

In fact, STT is a strong logic indeed. First, consider its T-free fragment, ST. ST is exactly classical logic. That is, an argument from premises Γ to conclusions Δ is ST valid iff it is classically valid. For proof, see [Ripley, 2011a]; the rough idea is this. Any two-valued classical counterexample to an argument immediately provides an ST counterexample, since (by Kripke's result) any two-valued classical model can be extended to a KK model. Similarly, any ST counterexample can be used to provide a classical counterexample: we build a

[5] On the other hand, defenders of S3T, as far as we can see, do not take this route. This is odd, since S3T is weaker than either K3T or LPT, and so if anything needs even more help than they do. It might be explained by the lack of well-developed theories of truth based on S3T; [Kremer, 1988] and [Halbach and Horsten, 2006] both explore the logic, but neither spends much time defending it.

[6] We have considered (in [Cobreros et al., 2011b]) a similar approach to providing a logic for vagueness. There, our models were (implicitly) four-valued, but again, we took an ST countermodel to be a model assigning 1 (the top value) to all the premises and 0 (the bottom value) to all the conclusions. Related ideas are also explored in [Nait-Abdallah, 1995], among other places.

two-valued model for the T-free language by assigning to atomic wffs value 1 where the ST countermodel assigns value 1, 0 where it assigns 0, and 1 or 0 (it doesn't matter which) where it assigns $\frac{1}{2}$. It can be shown that this always results in a classical counterexample to the argument.

So ST is classical logic. This means that STT conservatively extends classical logic: the only difference between classical logic and ST comes in arguments that involve T. On its own, this might still leave us worried about STT's strength: STT preserves all classically-valid inferences in the T-free language, but what does it have to say about the full language? The conservative extension result assures us that $A \vee \neg A$, for example, is valid when A includes no T. But what about when A does include a T?

This is a sensible worry. But, as it turns out, STT preserves all classical inferences: if $\Gamma \models^{CL} \Delta$, then $\Gamma^\star \models^{STT} \Delta^\star$, for any uniform substitution * on the full language. (For proof, see [Ripley, 2011a].) This ensures that arguments valid in the base language retain their validity in the full (T-involving) language. Thus, STT adds to classical logic in a benign way; it does not affect validity in the T-free vocabulary, and it allows T-free validities to extend to the full vocabulary.

Since STT is defined on KK models, it includes a fully transparent truth predicate. So STT is a logic with some interesting features; it is a conservative extension of classical logic with a transparent truth predicate, which allows classical reasoning to be used over the full language. This also shows that STT includes the unrestricted T-schema; since $\models^{CL} A \equiv A$, by the above results we have $\models^{STT} A \equiv A$, and thus by transparency $\models^{STT} A \equiv T\langle A\rangle$. STT shows that we can use KK models to define a logic for transparent truth that does not suffer from the excessive weakness of K3T, LPT, and S3T, without adding any extra connectives or other vocabulary.

Despite its considerable affinities with classical logic, however, STT holds some surprises. First among these is that it is *nontransitive*. There are wffs A, B, and C such that $A \models^{STT} B$ and $B \models^{STT} C$, but $A \not\models^{STT} C$. For example, consider a liar sentence λ equivalent to $\neg T\langle\lambda\rangle$. This sentence must take value $\frac{1}{2}$ on every KK model; it can receive no other value compatible with the constraints on $\neg$ and T. Since ST requires countermodels to go from 1 to 0, there is no ST countermodel to the argument from p to λ; thus, $p \models^{STT} \lambda$. Similarly, there is no ST countermodel to the argument from λ to q; $\lambda \models^{STT} q$. Nevertheless, it is easy to find an ST countermodel to the argument from p to q; just assign 1 to p and 0 to q. Therefore, $p \not\models^{STT} q$. STT consequence is not transitive.

This nontransitivity, though, is quite limited. In fact, it is restricted to cases where paradoxical sentences rear their heads, in a quite particular way. Let *generalized transitivity* be the move from $\Gamma \models^{STT} A, \Delta$ and $\Gamma, A \models^{STT} \Delta$ to $\Gamma \models^{STT} \Delta$ (in a sequent-calculus presentation, generalized transitivity amounts to the rule of cut). We know that generalized transitivity cannot hold in general; the counterexample above shows that. But it will hold in very many cases. In order to get a counterexample, we need $\Gamma \not\models^{STT} \Delta$: there must be some KK model on which every member of Γ takes value 1 and every member of Δ takes value 0. Call the set of all such models $\mathfrak{M}$; we know $\mathfrak{M}$ is nonempty. Now, if

A takes value 1 on any model in $\mathfrak{M}$, then $\Gamma, A \not\models^{STT} \Delta$, so we do not have a counterexample to generalized transitivity; similarly, if A takes value 0 on any model in $\mathfrak{M}$, then $\Gamma \not\models^{STT} A, \Delta$, so we again do not have a counterexample. It follows that, in any counterexample to generalized transitivity, A must take value $\frac{1}{2}$ on every model in $\mathfrak{M}$; that is, there must be no way to assign A value 1 or 0 while the Γs all get value 1 and the Δs all get value 0. It is quick to verify that this is a sufficient condition for counterexample as well.

So we have a counterexample to generalized transitivity—$\Gamma \models^{STT} A, \Delta$ and $\Gamma, A \models^{STT} \Delta$ but $\Gamma \not\models^{STT} \Delta$—iff: there is some KK model that assigns 1 to everything in Γ and 0 to everything in Δ, and every such model assigns $\frac{1}{2}$ to A. These are just the sentences that [Kripke, 1975] calls *paradoxical*: those that, given certain assumptions (here embodied by Γ and Δ), cannot receive value 1 or 0. So there is a fully precise sense in which the *only* failures of transitivity STT allows for arise in paradoxical cases. Outside the realm of what Kripke calls the paradoxical, transitivity is perfectly safe.[7] (If there were some reason besides paradoxicality for a sentence to exhibit this sort of feature, then Kripke's explication of paradoxicality would be mistaken.)

3.3 Metainferences

Transitivity (and its generalized relative) are familiar *metainferences*: they are principles under which a consequence relation might (or might not) be closed. As STT shows, it's entirely possible for a logic to retain all classically valid arguments across its full vocabulary while still failing certain classical metainferences. This immediately leads to two questions about STT, one technical and one philosophical. First, just how many familiar metainferences does STT fail? Second, how classical can STT be if it fails metainferences like transitivity? Here, we answer each question in turn.

Generalized transitivity is not the only familiar metainference failed by STT; there are two more related failures. First, one cannot conclude from $\Gamma \models^{STT} A \supset B, \Delta$ and $\Gamma \models^{STT} A, \Delta$ that $\Gamma \models^{STT} B, \Delta$. For example, consider the liar sentence λ, discussed above. As is quick to verify, we have $\models^{STT} \lambda$ and $\models^{STT} \lambda \supset p$, but $\not\models^{STT} p$. (Despite this, modus ponens itself, as a classically-valid argument, is still valid: $A, A \supset B \models^{STT} B$.) We do not think this is an *additional* cost, over and above the loss of transitivity; [Negri and von Plato, 2001, p. 19] show that this metainference is equivalent to generalized transitivity, given certain assumptions (which hold for STT).

A bit of care is also called for around the metainference of reductio. (Since double-negation rules hold without restriction in STT, there is no difference between "intuitionist" and "classical" forms—the care required is different.) In one familiar form, reductio moves from $\Gamma, A \vdash \neg A, \Delta$ to $\Gamma \vdash \neg A, \Delta$; this form holds for STT. In another familiar form, it moves from $\Gamma, A \vdash \bot, \Delta$ to $\Gamma \vdash \neg A, \Delta$; this form also holds for STT. In a third form, though, reductio moves from $\Gamma, A \vdash B \wedge \neg B, \Delta$ to $\Gamma \vdash \neg A, \Delta$, and this form fails for STT. (For

[7] Thanks to Sam Butchart and Graham Priest for discussion here.

example, $p \vDash^{STT} \lambda \wedge \neg\lambda$, but $\nvDash^{STT} \neg p$.)

It is less apparent this is related to the loss of transitivity, but in fact it is. In the presence of transitivity, one can conclude from $\Gamma, A \vdash B \wedge \neg B, \Delta$ and $B \wedge \neg B \vdash \neg A$ that $\Gamma, A \vdash \neg A, \Delta$, or from $\Gamma, A \vdash B \wedge \neg B, \Delta$ and $B \wedge \neg B \vdash \bot$ that $\Gamma, A \vdash \bot, \Delta$; one is then in a position to apply a form of reductio that holds in STT. Without transitivity, though, there is no guarantee that one can get to $\Gamma, A \vdash \neg A, \Delta$ or $\Gamma, A \vdash \bot, \Delta$, and thus no guarantee that reductio can apply.

As far as loss of familiar and important metainferences goes, that's about it. (Of course new "failures" of unfamiliar and unimportant metainferences can be generated ad infinitum by quick tweaks on the above.) Just to reassure, all the following metainferences hold in STT (for proofs, see [Ripley, 2011a]):[8]

Monotonicity: If $\Gamma \vDash^{STT} \Delta$, then $\Gamma, \Gamma' \vDash^{STT} \Delta, \Delta'$.

Structural contraction: If $\Gamma, A, A \vDash^{STT} \Delta$, then $\Gamma, A \vDash^{STT} \Delta$; and if $\Gamma \vDash^{STT} A, A, \Delta$, then $\Gamma \vDash^{STT} A, \Delta$.

Proof by cases: If $\Gamma, A \vDash^{STT} \Delta$ and $\Gamma, B \vDash^{STT} \Delta$, then $\Gamma, A \vee B \vDash^{STT} \Delta$.

Classical deduction theorem: $\Gamma, A \vDash^{STT} B, \Delta$ iff $\Gamma \vDash^{STT} A \supset B, \Delta$.

Conjoining premises, disjoining conclusions: $\Gamma \vDash^{STT} \Delta$ iff $\Gamma' \vDash^{STT} \Delta'$, where Γ' comes from Γ by possibly conjoining some of its members, and Δ' comes from Δ by possibly disjoining some of its members.

It's worth noting that many other approaches to truth do not retain all these metainferences. For example, supervaluationist approaches based on [Kripke, 1975], as discussed in [Field, 2008, Hyde, 1997], give up proof by cases and disjoining conclusions, the nonclassical approaches in [Beall, 2009, Field, 2008] give up the deduction theorem (in fact, they even give up the much weaker version of the deduction theorem without side premises or conclusions), and the contraction-free approach recommended in [Zardini, 2011] gives up not just structural contraction, but proof by cases as well. Even the classical theory FS described in [Friedman and Sheard, 1987] gives up the deduction theorem. What's more, these failures are not incidental to these approaches; with the metainferences imposed the approaches simply do not work. That is, they trivialize, yielding the result that $\Gamma \vDash \Delta$ for any Γ, Δ. (For further discussion of these theories, see §5.)

This is enough to give a sense of the situation with familiar metainferences in STT. The question remains: is it appropriate to call STT classical, given that it fails some metainferences that hold of T-free classical logic? This is in some sense a purely terminological question, but there is a philosophical core to it. We often think of logics as involving both valid arguments and metainferences; by losing metainferences, it seems we weaken our logic. Even if STT keeps all classically-valid arguments, if it loses some metainferences, then it might seem

[8] Sequent calculi are a way to present a logic almost entirely through metainferences, and [Ripley, 2011b] shows that STT retains all the rules of usual (cut-free) classical sequent calculi as well.

to have weakened some aspect of classical logic, and this could be enough to put it in with other nonclassical approaches to paradox.

Even if this claim were right, it would not be too much trouble; it's not a bad crowd to be lumped in with. Nonclassical approaches to paradox include some of the subtlest, most valuable, and most plausible approaches. However, the claim is not right: one does not weaken a logic simply by losing a metainference.

We will explore this first in a specific case and then in some generality. First, the specifics. Consider the propositional modal logics S4 and S5. It is clear, we take it, that S5 is a strengthening of S4; indeed, if S5 is not a strengthening of S4, then we have no idea what use the notion of strengthening might be put to. Nonetheless, S5 fails some metainferences that S4 obeys. For example, consider the metainference: If $\vdash \Diamond p \supset \Box\Diamond p$, then $\vdash \bot$. S4's consequence relation is closed under this rule, since $\nvdash_{S4} \Diamond p \supset \Box\Diamond p$. However, S5's consequence relation is not, since $\vdash_{S5} \Diamond p \supset \Box\Diamond p$ but $\nvdash_{S5} \bot$.

This is not a coincidence; facts like this hold under *very* minimal conditions. Let the *universal* consequence relation be the relation $\vdash_U$ that holds between *every* possible combination of premises and conclusions, and suppose we have two consequence relations $\vdash_1$ and $\vdash_2$ such that $\vdash_1 \subset \vdash_2 \subset \vdash_U$ (note that these are *strict* inclusions). Then $\vdash_2$ fails some metainferences that $\vdash_1$ satisfies.

Here's why: let Γ, Δ fall in the difference between $\vdash_2$ and $\vdash_1$; that is, choose Γ, Δ so that $\Gamma \vdash_2 \Delta$ but $\Gamma \nvdash_1 \Delta$. (By the strict inclusion of $\vdash_1$ in $\vdash_2$, there will be some such.) Similarly, let Γ', Δ' fall in the difference between $\vdash_U$ and $\vdash_2$. Then $\vdash_1$ satisfies, but $\vdash_2$ does not, the metainference: if $\Gamma \vdash \Delta$, then $\Gamma' \vdash \Delta'$. We want to stress that these are *very* minimal conditions indeed; they arise just about every time a logic is extended at all. It thus makes no sense to think of losing a metainference as weakening a logic—it's impossible to avoid, so long as we don't adopt the universal consequence relation.[9]

In other words, if STT gives up something important about T-free classical logic, it is not because it fails some metainferences that hold for T-free classical logic; any way at all of extending classical logic (short of moving to the universal consequence relation) does that. It must rather be because there is something important about the *particular* metainferences in question. In the case of STT, we reckon the focus should rest on (generalized) transitivity.

Again we must be careful to set terminological questions aside (although it is interesting to notice how vague the concept of classical logic turns out to be). Even if one uses the word 'classical' so as to exclude STT on the grounds of its nontransitivity, it cannot be denied that STT allows its users to recognize that *every* classically-valid argument is valid. We take this to be the main advantage conferred by sticking to classical logic, and so STT is classical enough for us.

[9]The S4/S5 example above fits this mold; so too does the following example. Classical predicate logic fails some metainferences that hold in classical propositional logic; for example, if $\forall x Px \vdash Pa$, then $p \vdash q$. It would be a serious abuse of terminology to hold that classical predicate logic is not classical for this reason. (To be able to make a direct comparison, we assume that both logics share the same language; then classical propositional logic simply treats things like $\forall x Px$ as atoms.)

3.4 Coding, induction, and compositionality

This far, we've been working with a simple quote-name approach, on which $\langle A \rangle$ names the wff A, and there's nothing more to it. However, an ideal theory of truth should include more than this: we want a full theory of syntax. In this subsection, we'll discuss how to achieve this within STT. We use Peano arithmetic and Gödel coding to get the job done; for details, see eg [Boolos, 1995]. We'll write $\ulcorner A \urcorner$ for the code of a piece of vocabulary A. We use a predicate $\texttt{sent}(x)$ true of all and only the codes of sentences, a predicate $\texttt{var}(x)$ true of all and only the codes of variables, and functions $\dot{\neg}$, $\dot{\wedge}$, $\dot{\forall}$, and $\dot{T}$ such that for any formulas A, B, and variable x: $\dot{\neg}\ulcorner A \urcorner = \ulcorner \neg A \urcorner$, $\ulcorner A \urcorner \dot{\wedge} \ulcorner B \urcorner = \ulcorner A \wedge B \urcorner$, $\dot{\forall}\ulcorner v \urcorner \ulcorner A \urcorner = \ulcorner \forall v A \urcorner$, and $\dot{T}\ulcorner A \urcorner = \ulcorner T \ulcorner A \urcorner \urcorner$. Such predicates and functions are definable from the vocabulary of PA. (Corresponding functions for $\vee$, $\supset$, $\equiv$, and $\exists$ can also be defined, and will work the same, mutatis mutandis. For this subsection only, we forget all about quote-names.)

In this framework, we can express the so-called 'compositional principles': principles like $\forall x \forall y(\texttt{sent}(x \dot{\wedge} y) \supset (T(x \dot{\wedge} y) \equiv (Tx \wedge Ty)))$. These seem to express important claims about truth: in this case, that a conjunction of any two sentences is true iff the sentences themselves are both true. Each connective and quantifier gives rise to a compositional principle. The others, in the present vocabulary, are $\forall x(\texttt{sent}(x) \supset (T \dot{\neg} x \equiv \neg Tx))$ and $\forall x \forall y(\texttt{sent}(\dot{\forall} xy) \supset (T \dot{\forall} xy \equiv \forall t(y(t/x))))$, where if $y = \ulcorner A \urcorner$ and $x = \ulcorner v \urcorner$, $y(t/x)$ is the code of the formula that results from substituting t for v everywhere in A.

Starting from the standard classical model M of (T-free) PA, we can again use Kripke's result to show that there are models extending M with a truth predicate T such that for any formula A, $T\ulcorner A \urcorner$ gets the same value on M that A does. Call these models *KKP models* (for 'Kleene-Kripke-Peano'), and define a new notion $\models^{STT}_{PA}$ of consequence analogously to $\models^{STT}$, but restricted to KKP models.

Clearly, every theorem of T-free PA will receive value 1 in every KKP model. But with T in the language, there are new instances of PA's induction axiom schema formulable. Not all of these can take value 1, but they all do take value greater than 0 on every KKP model.[10] Thus, every instance I of the induction schema, even extended to those instances involving T, is such that $\models^{STT}_{PA} I$; they are all theorems.

Moreover, the compositionality principles alluded to above are also theorems of $\models^{STT}_{PA}$. It is shown in [Halbach, 2011] that the system there named PKF is sound over KKP models. PKF includes the turnstile versions of the compositionality principles; for example, it includes $\texttt{sent}(x \dot{\wedge} y), T(x \dot{\wedge} y) \vdash Tx \wedge Ty$. It can be shown that 1) if these principles hold in PKF, then they hold in $\mathrm{STT}_{\mathrm{PA}}$, and 2) if these principles hold in turnstile form in $\mathrm{STT}_{\mathrm{PA}}$, then they hold in quantified theorem form as well (due to $\mathrm{STT}_{\mathrm{PA}}$'s obeying a deduction theorem and allowing for the sequent metainference introducing $\forall$ on the right). As a

[10] The instances are all of the form $(A(0) \wedge \forall x(A(x) \supset A(x+1))) \supset \forall x A(x)$. The only way for this sentence to get value 0 on a KKP model M is for $A(0) \wedge \forall x(A(x) \supset A(x+1))$ to get value 1 and $\forall x A(x)$ to get value 0. This cannot happen, given the constraints on $\supset$ and $\forall$.

result, STT, when restricted to fixed points over the standard model of PA, allowing it to express its own syntax, automatically captures the compositional principles that some other theories of truth struggle with.

For the remainder of the paper, we return to the quote-name approach, for simplicity; but we will sometimes recall these nice features of the system including arithmetic.

4 Paradoxes

4.1 Paradoxical arguments

If every inference form valid in classical logic is STT-valid as well, and STT supports a transparent truth predicate, then where does the liar argument go wrong? Here's one version of the argument, as a proof by cases, where λ is the liar sentence $\neg T\langle\lambda\rangle$:

$$\dfrac{\dfrac{\text{LEM}\ \dfrac{}{T\langle\lambda\rangle \vee \neg T\langle\lambda\rangle}\top \qquad \dfrac{\dfrac{\dfrac{\dfrac{[T\langle\lambda\rangle]^1}{\lambda}\,T\text{E}}{\neg T\langle\lambda\rangle}\,\text{Def. }\lambda}{T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle}\wedge\text{I} \qquad \dfrac{\dfrac{\dfrac{[\neg T\langle\lambda\rangle]^1}{\lambda}\,\text{Def. }\lambda}{T\langle\lambda\rangle}\,T\text{I}}{T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle}\wedge\text{I}}{\vee\text{E, }1}}{\dfrac{T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle}{\bot}\,\text{Explosion}}$$

If indeed $\top \vDash^{STT} \bot$, something has gone very wrong: this would tell us that every model such that $1 = 1$ is such that $0 > 0$; in other words, it would tell us that there are no models, and so no countermodels, so $\Gamma \vDash^{STT} \Delta$ for every Γ, Δ. We know, since STT conservatively extends classical logic, that this is not the case, but how is it avoided?

Every step in the above proof except the T-steps is STT-valid, and the T-steps are STT-valid as well. (After all, $A \vDash^{STT} A$, so transparency guarantees that $A \vDash^{STT} T\langle A\rangle$ and $T\langle A\rangle \vDash^{STT} A$.) So every step is STT-valid. It's the attempt to chain them together that's gone wrong.

Let's say that a model *strictly* satisfies a sentence when it assigns value 1 to that sentence, and that it *tolerantly* satisfies a sentence when it assigns some value greater than 0 to that sentence. Remember, $\vDash^{STT}$ imposes weaker conditions on its conclusions than on its premises: when all the premises of an STT-valid argument are strictly satisfied, some of its conclusions must be tolerantly satisfied.

For most of the above argument, that feature doesn't particularly matter, but there are two steps where it does: the step from $\top$ to $T\langle\lambda\rangle \vee \neg T\langle\lambda\rangle$, and the step from $T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle$ to $\bot$. The excluded middle step only guarantees that $T\langle\lambda\rangle \vee \neg T\langle\lambda\rangle$ is *tolerantly* satisfied, not strictly. From here, we can (in fact)

safely conclude that $T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle$ is tolerantly satisfied.[11] But explosion needs *strictly* satisfied premises to do its work: it is STT-valid, but not LPT-valid. So although explosion is valid, its premises are not satisfied *enough* for that validity to matter.

A similar approach works for the Curry paradox, a sentence κ equivalent to $T\langle\kappa\rangle \supset \bot$. Consider the following proof:

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{[T\langle\kappa\rangle]^1}{\kappa}\,T\text{E}}{T\langle\kappa\rangle \supset \bot}\,\text{Def. }\kappa}{\bot}\,\supset\text{E}}{T\langle\kappa\rangle \supset \bot}\,\supset\text{I, 1}}{\kappa}\,\text{Def. }\kappa}{T\langle\kappa\rangle}\,T\text{I}}{\bot}\,\supset\text{E}$$

Again, every step is STT-valid, but the proof seems to show that $\vDash^{STT} \bot$. We know, since STT conservatively extends classical logic, that this is not the case, so the trouble must have again come from linking the steps together. Here, the trick is pulled between the $\supset$-intro and the final modus ponens. By deriving $\bot$ from $T\langle\kappa\rangle$, we can validly conclude $T\langle\kappa\rangle \supset \bot$, but this is guaranteed to hold only tolerantly. Modus ponens, on the other hand, requires strictly satisfied premises; like explosion, it is STT-valid but not LPT-valid.[12]

Since STT is a conservative extension of classical logic, we know that there is no way an as-yet-undiscovered paradox will trivialize it. All formulable paradoxes[13] will have treatments like the liar and Curry above; somewhere in the derivation of the troublesome conclusion, if every individual step is valid, there will be an illicit use of transitivity. Something will be demonstrated only tolerantly and drawn on as though it held strictly.[14]

4.2 The status of paradoxical sentences

So much for logical consequence. A natural next question, though, is what *status* paradoxical sentences have on our view. Consider again the liar λ. It is both a theorem ($\vDash^{STT} \lambda$) and refutable ($\lambda \vDash^{STT}$). Similarly, the claim that it's true is both a theorem and refutable, as is the claim that it's false. What do we say about such sentences, then?

[11] This is so because all the inferences that take place between $T\langle\lambda\rangle \vee \neg T\langle\lambda\rangle$ and $T\langle\lambda\rangle \wedge \neg T\langle\lambda\rangle$ are LPT-valid. LPT-valid inferences can always safely be chained on to the *end* of STT-valid inferences.

[12] Interestingly, it requires only *one* strictly satisfied premise, and it doesn't matter which; either the conditional or the antecedent will do. But if both are only tolerantly satisfied, no dice.

[13] An example of an (as-yet-) unformulable paradox: we include no treatment here of definite descriptions, and so cannot formulate Berry's paradox. We will treat this (and others) in future work.

[14] For example, in the Dean/Nixon case explored in [Kripke, 1975], if the circumstances are such as to render the case paradoxical, it will emerge that *both* Dean's and Nixon's utterances can be demonstrated to hold only tolerantly.

Here, we see two options that directly present themselves. Rather than argue for one in particular, we will briefly present them both, without much in the way of evaluation. Which is a better choice, or whether there is some third choice better than both, are issues we leave for future work.

The first approach works at the level of *pragmatics*. On this approach, what can be said about paradoxical sentences depends on how the saying is being done. As in [Ripley, 2011b], we distinguish two forms of assertion, strict and tolerant. Strictly, the liar and other paradoxical sentences cannot be asserted; tolerantly, they can. The same goes for their negations. Since the truth predicate is fully intersubstitutable, if we speak strictly we do not claim either that these sentences are true or that they are not true; if we speak tolerantly, we happily claim both.

It is natural to see the values in a model theory as intimately tied to (idealized) assertibility; this is so whether one thinks that assertibility is prior to semantic value or vice versa (or neither). More familiar approaches to three-valued models invoke a notion of "designated value"; this amounts to imposing a two-way division over the top: either value-1 sentences are assertible and others are not, or else value-0 sentences are not assertible and others are. But there is no way to understand an STT-based approach in terms of designated values, and we do not impose this two-way division.[15]

Instead, we can see a direct connection between model-theoretic value and assertibility. A sentence is either strictly assertible, tolerantly but not strictly assertible, or not assertible at all. We do not allow for sentences that are strictly but not tolerantly assertible; strict assertion, on this picture, is a (strictly) stronger speech act than tolerant assertion. Paradoxical sentences reveal the difference between strict and tolerant assertion: they are tolerantly but not strictly assertible.

The other approach works at the level of *meaning*. Rather than supposing that there are two distinct speech acts of assertion, this approach supposes that each sentence has two distinct meanings (or two distinct aspects of its meaning, if you like) that can be asserted: its strict meaning and its tolerant meaning. Understanding meanings as dividing the space of models in two, we can understand a sentence's strict meaning as one drawing a division between those models on which the sentence takes value 1 and those on which it takes some value less than 1, and we can understand a sentence's tolerant meaning as one drawing a division between those models on which the sentence takes some value greater than 0 and those on which it takes value 0.

This is the approach we explored for vague language in [Cobreros et al., 2011b]. Again, strict and tolerant are related by strength: every sentence's strict meaning is at least as strong as its tolerant meaning. Paradoxical sentences, on this picture, reveal the difference between strict and tolerant meaning; they are those sentences whose tolerant meanings are true but whose strict meanings are not.[16]

[15] As [Dunn and Hardegree, 2001] show, every logic based on designated values in the usual way is transitive.

[16] If we like, we can call sentences whose tolerant meanings are true "tolerantly true" and

Unlike the pragmatic approach, this approach must immediately grapple with apparent revenge problems in the present context. For example, the sentence 'This sentence's strict meaning is not true' would seem to function as a liar. We are not so worried about this possibility. One can try to argue as follows: "If its strict meaning is true, then its strict meaning is not true (since that's what it says); so its strict meaning is not true. But then what its strict meaning says is the case, so its strict meaning is also true. Its strict meaning, then, is both true and not true. But then everything follows." This reasoning, though, assumes transitivity throughout, and we've given a theory on which transitivity cannot be assumed, particularly in reasoning involving truth. What the reasoning shows is that, even when an appropriate treatment of strict and tolerant meaning is brought into the language itself, there can still be failures of transitivity due to paradoxes.

As far as we can see, then, there are at least two ways to understand the status paradoxical sentences have on an STT-based theory like the one we've advanced here. Both ways take paradoxical sentences to fall in between strict and tolerant, but one way takes the distinction between strict and tolerant to be a pragmatic distinction, and the other to be a distinction in meaning. On the second approach, revenge troubles might seem to loom, but they, just like the original paradoxes, depend on transitivity, which we expect to fail when paradoxes are around.

5 Comparisons

This section serves to locate STT as a formal approach to truth by comparing it and contrasting it to some of its relatives in the literature. One key difference between STT and most other approaches is clear: transitivity. Almost all existing approaches to truth are based on transitive logics (but see §5.4), while STT, quite crucially, is not. The other main distinction is STT's combination of transparency and classicality; no other theory combines these features.

5.1 FS

The first relative of STT we should look to is FS, or the Friedman-Sheard theory of truth. (This theory is presented in [Friedman and Sheard, 1987] and discussed in eg [Halbach, 2011, Ch. 14].) It is typically presented axiomatically, by adding a variety of axioms to Peano Arithmetic (PA), along with a pair of rules:

sentences whose strict meanings are true "strictly true", but one should not assume particular truth-table-based accounts of these predicates. For instance, it cannot be that 'A is strictly true' takes value 1 iff A takes value 1, and takes value 0 otherwise. This would impose inconsistent requirements on our models, due to the existence of a sentence claiming its own strict untruth. Note that similar restrictions must be required by any approach based on Kripke's construction, and can be understood in a number of different ways (as in [Priest, 2006a, Field, 2008]).

$$\text{Nec:}\quad \frac{A}{T\langle A\rangle} \qquad\qquad \text{Co-nec:}\quad \frac{T\langle A\rangle}{A}$$

Crucially, FS includes neither $A \supset T\langle A\rangle$ nor $T\langle A\rangle \supset A$ as theorems, and neither can be added, on pain of triviality; it thus does not validate the T-schema in either direction, one major difference with STT. (The same goes for many other classically-minded theories of truth, including those in [Gupta and Belnap, 1993, Maudlin, 2004].) Since $A \supset A$ is valid in the FS theory, these cases provide counterexamples to transparency as well, another difference with STT.

FS is usually considered to be a theory of truth within classical logic. We think this is right, but want to call attention to what is involved. First, every classically valid argument remains valid in FS, and this feature extends to arguments involving truth vocabulary; these are features FS shares with STT. Another feature FS shares with STT is the failure of familiar and useful metainferences. For STT, transitivity goes; for FS, it is the deduction theorem that must fail. With a deduction theorem, we could derive $A \supset T\langle A\rangle$ from the rule Nec, or $T\langle A\rangle \supset A$ from the rule Co-nec, and either would immediately trivialize the system. The sense in which FS is classical is thus a sense that allows for failure of familiar and useful metainferential properties like the deduction theorem.

FS and STT are thus equally examples of classical approaches to truth that achieve some measure of control over paradoxes by allowing for the failure of certain metainferences. The existence of STT undermines any attempt to defend FS's failure to support the T-schema and transparency by insisting that no classical approach can support these principles. STT does support these principles, and, as above, it is as classical as FS. In addition, STT, like FS, includes the compositional principles for truth, if we restrict our attention to fixed points over the standard model of PA, as we pointed out in §3.4.

The final difference between FS and STT that we'll mention here: FS is ω-inconsistent, and can have no standard models. STT, on the other hand, is shown to have standard models by the Kripke construction.[17]

5.2 Extra-arrow theories

One subfamily of STT's nonclassical relatives includes the logics of [Priest, 2006b, Brady, 2006, Field, 2008], and [Beall, 2009]. While these logics differ from each other in various ways, their differences from STT are more uniform; here, we'll discuss them together, paying more attention to their common features than to what differentiates them.

[17] $\mathrm{STT}_{\mathrm{PA}}$, which contains the compositional principles, PA, and a transparent truth predicate, more than satisfies the conditions for the "negative result" in [McGee, 1985], showing that any system meeting weaker conditions than these must be ω-inconsistent. (It is this result that shows FS to be ω-inconsistent.) Nonetheless, the result does not apply here, as McGee's argument depends on assuming transitivity.

Like STT and unlike FS, most of these logics support full transparency.[18] All four logics include, in addition to the defined conditional $\supset$, a new conditional $\rightarrow$, and they all validate the T-schema, at least in the form $A \leftrightarrow T\langle A\rangle$. Priest's and Beall's systems in addition validate the T-schema in $\equiv$ form; Brady's and Field's do not.

Unlike FS, these theories of truth involve genuinely nonclassical logics; Priest's and Beall's logics are extensions of LPT, Field's is an extension of K3T, and Brady's, as a relevant logic, is an extension of the logic FDE (see [Anderson and Belnap, 1975] or [Priest, 2008] for details of FDE). The most apparent nonclassicalities involve negation; none of the logics validates both excluded middle ($A \vdash B \vee \neg B$) and explosion ($A \wedge \neg A \vdash B$), and Brady's validates neither. The situation around reductio is also delicate. While Priest's and Beall's logics support reductio in two of the above-discussed forms—allowing passage from $\Gamma, A \vdash \neg A, \Delta$ or from $\Gamma, A \vdash \bot, \Delta$ to $\Gamma \vdash \neg A, \Delta$—none of these logics support reductio in a different form—none allow passage from $\Gamma, A \vdash B \wedge \neg B, \Delta$ to $\Gamma \vdash \neg A, \Delta$. (The usual equivalence between these forms depends inter alia on explosion, which neither Priest's nor Beall's logic validates.) In contrast, STT supports both excluded middle and explosion, as well as the first two forms of reductio. As we mentioned in §3.3, it also does not support the third form of reductio—there, STT matches these logics, albeit for different reasons.

The two conditionals in these logics ($\supset$ and $\rightarrow$) approximate the classical $\supset$ in different ways. Because of the failures of excluded middle and explosion, none of these logics includes both of $\supset$-identity ($\vdash A \supset A$) and $\supset$-modus ponens ($A, A \supset B \vdash B$). This is the usual reason for adding $\rightarrow$; all four logics validate both $\rightarrow$-identity and $\rightarrow$-modus ponens. A difference in the other direction between the conditionals occurs over the rule of (conditional, rather than structural) contraction: for all four logics, $A \supset (A \supset B) \vdash A \supset B$, but $A \rightarrow (A \rightarrow B) \nvdash A \rightarrow B$. In fact, adding this last validity to any of the logics would trivialize it immediately. The same goes for the arrow form of $\rightarrow$-modus ponens ($\vdash (A \wedge (A \rightarrow B)) \rightarrow B$); this too cannot be added to any of these logics. As a result, none of them can enjoy a deduction theorem for $\rightarrow$. In addition, none of them enjoys both directions of the deduction theorem for $\supset$ (even in the weak form: $A \vdash B$ iff $\vdash A \supset B$); Priest, Brady, and Beall all fail the right-to-left direction, while Brady and Field fail the left-to-right direction.

By contrast, STT's single conditional $\supset$ validates all the principles discussed here: identity, modus ponens, arrow form modus ponens, contraction, and a full deduction theorem (even in the strong form: $\Gamma, A \vdash B, \Delta$ iff $\Gamma \vdash A \supset B, \Delta$). So these theories, while (at least potentially) sharing STT's transparency, share little of its classicality. A number of important inferences and metainferences around negation and the conditional are lost.

When it comes to offering a theory of paradoxical sentences, however, there is more affinity between STT and these extra-arrow theories. Consider the liar

[18] Priest is the only exception, for philosophical rather than technical reasons; transparency can be added to Priest's system without triviality.

sentence λ. Priest and Beall offer theories on which both λ and $\neg\lambda$ are to be asserted, and neither is to be denied. If assertion is understood tolerantly and denial strictly, this is our approach as well. Dually, Field offers a theory on which both λ and $\neg\lambda$ are to be denied, and neither is to be asserted. If assertion is understood strictly and denial tolerantly, this is our approach as well.[19]

5.3 Contraction-free

Recently, [Zardini, 2011] has advanced a theory of transparent truth based on restricting the structural rules of contraction (the rules that allow one to move from $\Gamma, A, A \vdash \Delta$ to $\Gamma, A \vdash \Delta$, and from $\Gamma \vdash A, A, \Delta$ to $\Gamma \vdash A, \Delta$), and [Beall and Murzi, 2011] has also offered some arguments in favor of such a view.

Amongst nonclassical approaches, this is probably the closest to STT. Zardini's logic $\mathbf{IKT}^\omega$, for example, retains a deduction theorem, excluded middle, explosion, and weakened forms of reductio. In addition, both $\mathbf{IKT}^\omega$ and STT have as a theorem every instance of the claim that modus ponens is truth-preserving: $\vdash (T\langle A \supset B\rangle \wedge T\langle A\rangle) \supset T\langle B\rangle$.[20]

There are some notable differences, however. First, $\mathbf{IKT}^\omega$ is weaker than classical logic, even on some very basic arguments: for example, $A \not\models^{\mathbf{IKT}^\omega} A \wedge A$, and $A \vee A \not\models^{\mathbf{IKT}^\omega} A$. This is crucial; adding these principles would trivialize the logic. A number of familiar metainferences also fall by the wayside; for example, both reductio and proof by cases hold only in a weakened form, since the full forms of these metainferences would bring enough contraction into the system to trivialize it. Although the loss of classical principles is perhaps less drastic than in the case of many other nonclassical systems, it is still very much a part of Zardini's approach.

Second, while $\mathbf{IKT}^\omega$ is known to be nontrivial, its relation to models of PA has not yet been explored. This leaves in question the status of the compositional principles mentioned in §5.1. STT, by building on the well-explored Kripke construction, can provide these principles.

5.4 Nontransitive

Finally, we mention the relation between STT and the nontransitive system advanced in [Weir, 2005] to address paradoxes of truth. As with the contraction-free systems, this system comes quite close to classical logic. In fact, we think it's the closest to classical of the nonclassical systems we consider here. However, it still exhibits some nonclassical, and we think odd, behavior.

[19] There is also a real connection "under the hood". All four extra-arrow logics are proved nontrivial by their authors via a model construction whose prototype is the construction in [Brady, 1989]; this construction involves a transfinite series of what are essentially Kripke fixed-point constructions. The Kripke construction, in all cases, handles T completely, as it does for us in §3.1; the transfinite series is only necessary to handle the extra arrow.

[20] $\mathrm{STT}_{\mathrm{PA}}$ also includes as a theorem the quantified version of this principle: $\forall x \forall y(\mathtt{form}(x) \wedge \mathtt{form}(y) \supset ((T(x \dot{\supset} y) \wedge Tx) \supset Ty))$. $\mathbf{IKT}^\omega$'s relation to arithmetic, and its take on this quantified form of the principle, is still unknown.

A number of crucial arguments, such as modus ponens, hold in Weir's logic only under restricted conditions. In addition, the theoremhood cannot be defined in the usual way (as consequences of the empty set of premises); rather, Weir says, "The notion of theoremhood... has to be: ϕ is a theorem iff for some A, B, we have that $A \to A, B \to B \vdash \phi$ is provable" (246). (Here, $\to$ is a special conditional in Weir's logic, not $\supset$.)

If one is willing, with Weir, to give up transitivity in the pursuit of truth, STT shows that there is no need to make these further modifications. It's possible, as we've shown here, to give up transitivity within classical logic, and thus retain unrestricted modus ponens, the usual notion of theoremhood, and other classical features.

6 Conclusion

This paper has presented and explored a logical framework, STT, for adding transparent truth to classical logic. By building on the familiar Kripke construction, but using an unfamiliar definition of countermodel, and so of logical consequence, STT allows us both to retain every classically-valid argument and to allow for a fully transparent truth predicate. This is possible because some familiar metainferences, crucially including transitivity, fail for STT.

It's been claimed [Leitgeb, 2007] that the following eight desiderata for a theory of truth are not jointly satisfiable: 1) that it include a truth predicate and a theory of syntax; 2) that, when added to a mathematical or empirical theory, it allow for that theory to be proven true; 3) that it be type-free; 4) that it include the full T-schema; 5) that it be compositional; 6) that it allow for standard interpretations; 7) that its outer and inner logics coincide (that is, that A entails B iff $T\langle A\rangle$ entails $T\langle B\rangle$); and 8) that its logic be classical.

When one considers $\mathrm{STT}_{\mathrm{PA}}$ (as in §3.4), it turns out that all eight of these desiderata *are* satisfied. (Arithmetic is important here to get a theory of syntax, for desideratum 1, and to formulate the compositional principles, for desideratum 5.) The argument that they cannot be jointly satisfied turns crucially on the assumption of transitivity, but transitivity is not among the eight desiderata, nor does it follow from them. (STT shows that a logic can be classical (and thus satisfy desideratum 8) without being transitive.) As Leitgeb says, "In the best of all (epistemically) possible worlds, some theory of truth would satisfy all of these norms at the same time" (283). We might yet live there, unless transitivity is seen as an additional desideratum. However, as we've tried to argue, the loss of transitivity is minimally disruptive; transitivity continues to hold in nonparadoxical cases.

There is much left to do. We have not here explored an STT-based theory's prospects for avoiding revenge paradoxes, or description-based paradoxes like Berry's. We also have not drawn very many connections between this treatment of truth and our treatments of vague predicates in [Cobreros et al., 2011b, Cobreros et al., 2011a], although the approaches are intimately related. Although we've sketched some relations between our approach

and other approaches in the literature, we have not given the issue the detailed exploration it deserves. These issues await future research. For now, we are content to put STT on the table as suggesting a promising avenue for approaching the paradoxes.

References

[Anderson and Belnap, 1975] Anderson, A. R. and Belnap, N. D. (1975). *Entailment: The Logic of Relevance and Necessity*, volume 1. Princeton University Press, Princeton, New Jersey.

[Beall, 2009] Beall, J. (2009). *Spandrels of Truth*. Oxford University Press, Oxford.

[Beall and Murzi, 2011] Beall, J. and Murzi, J. (2011). Two flavors of Curry's paradox. *Journal of Philosophy*. To appear.

[Boolos, 1995] Boolos, G. (1995). *The Logic of Provability*. Cambridge University Press, Cambridge.

[Brady, 1989] Brady, R. (1989). The non-triviality of dialectical set theory. In Priest, G., Routley, R., and Norman, J., editors, *Paraconsistent Logic: Essays on the Inconsistent*, pages 437–471. Philosophia Verlag, Münich.

[Brady, 2006] Brady, R. (2006). *Universal Logic*. CSLI Publications, Stanford, California.

[Cobreros et al., 2011a] Cobreros, P., Égré, P., Ripley, D., and van Rooij, R. (2011a). Tolerance and mixed consequence in the s'valuationist setting. *Studia Logica*. To appear.

[Cobreros et al., 2011b] Cobreros, P., Égré, P., Ripley, D., and van Rooij, R. (2011b). Tolerant, classical, strict. *Journal of Philosophical Logic*. To appear.

[Dunn and Hardegree, 2001] Dunn, J. M. and Hardegree, G. M. (2001). *Algebraic Methods in Philosophical Logic*. Oxford University Press, Oxford.

[Field, 2008] Field, H. (2008). *Saving Truth from Paradox*. Oxford University Press, Oxford.

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

[Gupta and Belnap, 1993] Gupta, A. and Belnap, N. (1993). *The Revision Theory of Truth*. MIT Press, Cambridge, Massachusetts.

[Halbach, 2011] Halbach, V. (2011). *Axiomatic Theories of Truth*. Cambridge University Press, Cambridge.

[Halbach and Horsten, 2006] Halbach, V. and Horsten, L. (2006). Axiomatizing Kripke's theory of truth. *Journal of Symbolic Logic*, 71(2):677–712.

[Hyde, 1997] Hyde, D. (1997). From heaps and gaps to heaps of gluts. *Mind*, 106:641–660.

[Kleene, 1952] Kleene, S. C. (1952). *Introduction to Metamathematics*. North-Holland Publishing Co., Amsterdam.

[Kremer, 1988] Kremer, M. (1988). Kripke and the logic of truth. *Journal of Philosophical Logic*, 17(3):225–278.

[Kripke, 1975] Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy*, 72(19):690–716.

[Leitgeb, 2007] Leitgeb, H. (2007). What theories of truth should be like (but cannot be). *Philosophy Compass*, 2(2):276–290.

[Maudlin, 2004] Maudlin, T. (2004). *Truth and Paradox*. Oxford University Press, Oxford.

[McGee, 1985] McGee, V. (1985). How truthlike can a predicate be? A negative result. *Journal of Philosophical Logic*, 14(4):399–410.

[Nait-Abdallah, 1995] Nait-Abdallah, A. (1995). *The Logic of Partial Information*. Springer, Berlin.

[Negri and von Plato, 2001] Negri, S. and von Plato, J. (2001). *Structural Proof Theory*. Cambridge University Press, Cambridge.

[Priest, 2006a] Priest, G. (2006a). *Doubt Truth to be a Liar*. Oxford University Press, Oxford.

[Priest, 2006b] Priest, G. (2006b). *In Contradiction*. Oxford University Press, Oxford.

[Priest, 2008] Priest, G. (2008). *An Introduction to Non-Classical Logic: From If to Is*. Cambridge University Press, Cambridge, 2nd edition.

[Quine, 1970] Quine, W. V. O. (1970). *Philosophy of Logic*. Prentice-Hall, Englewood Cliffs, New Jersey.

[Ripley, 2011a] Ripley, D. (2011a). Conservatively extending classical logic with transparent truth. To appear.

[Ripley, 2011b] Ripley, D. (2011b). Paradox and failures of cut. *Australasian Journal of Philosophy*. To appear.

[Weir, 2005] Weir, A. (2005). Naïve truth and sophisticated logic. In Beall, J. and Armour-Garb, B., editors, *Deflationism and Paradox*, pages 218–249. Oxford University Press, Oxford.

[Zardini, 2011] Zardini, E. (2011). Truth without contra(di)ction. To appear.