Reaching Transparent Truth

Pablo Cobreros
University of Navarra,
pcobreros@unav.es

Paul Égré
Institut Jean-Nicod (CNRS-EHESS-ENS)
paul.egre@ens.fr

David Ripley
University of Connecticut
davewripley@gmail.com

Robert van Rooij
Institute for Logic, Language, and Computation, University of Amsterdam
R.A.M.vanRooij@uva.nl

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 STTT (for Strict-Tolerant Transparent Truth), is that it supports a non-transitive relation of consequence. At the same time, it can be seen that the only failures of transitivity STTT 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. 2012b, Cobreros et al. 2012a. 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 a sense that will be made precise below.

In section 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.

Mind, Vol. 122 . 488 . October 2013
doi:10.1093/mind/fzt110 Advance Access publication 1 April 2014

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. Section 3 introduces our target logic, which we will call STTT, and elaborates on its relation to *T*-free classical logic. Section 4 outlines a theory of paradoxical sentences based on STTT. Section 5 considers the advantages of our approach, comparing it to a number of other approaches in the literature. Finally, section 6 concludes.

2. Some desiderata for truth

Theories of the truth predicate differ on whether it 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 this view, the reason why truth should be transparent is related to the function of the truth predicate in natural language, namely to allow expression of generalizations we could not otherwise express (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; namely, 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: $A \leftrightarrow T\langle A\rangle$ (for at least some conditional $\rightarrow$). 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 can prove 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 behaviour 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 STTT

A number of approaches to maintaining transparent truth have been tried in response to the well-known paradoxes that inevitably arise. Many of these (e.g. 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 section 3.1, before proceeding to present our logical framework in section 3.2. (Although we here present our logic model-theoretically, it is susceptible to a proof-theoretic treatment as well; see Ripley 2012 for a three-sided sequent calculus, or Ripley 2013 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 will not present them; what is 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 $\{1, ½, 0\}$ of values, with Kleene's strong valuation schema.[1] According to this schema, negation maps 1 to 0, 0 to 1, and ½ onto itself; conjunction & 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 (see 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.

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

[1] Actually, Kripke considers the case where the value ½ 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.

[2] Note that this means we will use only models with infinite domains, since there are infinitely many wffs to talk about. To define a naive satisfaction predicate, we should also require the models to be acceptable in the sense of Moschovakis 1974, allowing for encoding finite sequences of members of the domain into individual members of the domain. (We will not worry more about satisfaction here, as the relevant features of the Kripke construction are already present with truth, and satisfaction poses no additional problems.)

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.[4] 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.[5] Moreover, M can be very rich indeed, and include predicates and terms appropriate for any subject matter 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: no amount of swapping As for $T\langle A\rangle$s, or vice versa, will ever affect the validity of any argument. This is 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 premisses 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.

[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] For experts: this is the stronger *model-theoretic* notion of conservative extension.

[5] 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.

Some of these ways result in relatively familiar logics. One way, resulting in the logic we will call K3TT (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 will call LPTT (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 of Δ. A third way, resulting in the logic we will call S3TT (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 S3TT valid, then, iff it is both K3TT valid and LPTT 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 K3TT and LPTT) 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, K3TT does not validate excluded middle ($\models A \vee \neg A$) or, equivalently, identity ($\models A \supset A$), LPTT does not validate material modus ponens ($A, A \supset B \models B$), and S3TT 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.[6]

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 STTT

The definition we consider stays very close to the familiar classical definition. We say a model is an ST countermodel to an argument from premisses Γ to conclusions Δ iff the model assigns 1 to every member of Γ and 0 to every member of Δ. The logic STTT (for ST

[6] On the other hand, defenders of S3TT, as far as we can see, do not take this route. This is odd, since S3TT is weaker than either K3TT or LPTT, and so it seems to need even more help than they do. It might be explained by the lack of well-developed theories of truth based on S3TT; Kremer 1988 and Halbach and Horsten 2006 both explore the logic, but neither spends much time defending it.

with transparent truth) is the logic that results from this definition over KK models.[7]

It is immediate that STTT is stronger than both K3TT and LPTT: any STTT countermodel is automatically both a K3TT and an LPTT countermodel, but there are K3TT and LPTT countermodels that are not STTT countermodels.

In fact, STTT is a strong logic indeed. First, consider its *T*-free fragment, ST. ST is exactly classical logic augmented with quote-names. To see this, consider the usual sort of two-valued classical model for $\mathcal{L}$, imposing the requirement (as ever) that $\langle A \rangle$ denote A, for every wff A of $\mathcal{L}^+$. Let CL be the usual classical consequence relation defined over these classical models.

Then an argument from premisses Γ to conclusions Δ is ST valid iff it is CL valid; we will refer to this as 'classically valid' throughout this paper.[8] For proof, see Ripley 2012; the rough idea is this. Any CL counterexample to an argument immediately provides an ST counterexample, since (by Kripke's result) any CL model can be extended to a KK model. Similarly, any ST counterexample can be used to provide a CL counterexample: we build a 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 does not matter which) where it assigns ½. It can be shown that this always results in a CL counterexample to the argument.

So ST captures classical validity. This means that STTT conservatively extends CL: the only difference comes in arguments that involve *T*. On its own, this might still leave us worried about STTT 's strength: STTT 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, STTT preserves all classical validities: if $\Gamma \models^{CL} \Delta$, then $\Gamma^\star \models^{STTT} \Delta^\star$, for any uniform substitution * on the full language. (For proof, see Ripley 2012.) This

[7] We have considered (in Cobreros et al. 2012b) 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 premisses and 0 (the bottom value) to all the conclusions. Related ideas are also explored in Frankowski 2004, Nait-Abdallah 1995, among other places.

[8] It is actually a slight extension of pure classical logic, at least in the presence of =, since all these models have infinite domains, and since sentences like $\langle p \rangle \neq \langle q \rangle$ are valid. See also footnote 2.

ensures that arguments valid in the base language retain their validity in the full (*T*-involving) language. Thus, STTT 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 STTT is defined on KK models, it includes a fully transparent truth predicate. So STTT is a logic with some interesting features; it is a conservative extension of CL with a transparent truth predicate, which allows classical reasoning to be used over the full language. This also shows that STTT includes the unrestricted *T*-schema; since $\vDash^{CL} A \equiv A$, by the above results we have $\vDash^{STTT} A \equiv A$, and thus by transparency $\vDash^{STTT} A \equiv T\langle A\rangle$. STTT shows that we can use KK models to define a logic for transparent truth that does not suffer from the excessive weakness of K3TT, LPTT, and S3TT, without adding any extra connectives or other vocabulary.

Despite its considerable affinities with classical validity, however, STTT holds some surprises. First among these is that it is *non-transitive*. There are wffs A, B, and C such that $A \vDash^{STTT} B$ and $B \vDash^{STTT} C$, but $A \nvDash^{STTT} C$. For example, consider a liar sentence λ equivalent to $\neg T\langle\lambda\rangle$. This sentence must take value ½ 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, $p \vDash^{STTT} \lambda$. Similarly, there is no ST countermodel to the argument from λ to q; $\lambda \vDash^{STTT} q$. 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, $p \nvDash^{STTT} q$. STTT consequence is not transitive.

This non-transitivity, though, is quite limited. It is restricted in the following way. Let *generalized transitivity* be the move from $\Gamma \vDash^{STTT} A, \Delta$ and $\Gamma, A \vDash^{STTT} \Delta$ to $\Gamma \vDash^{STTT} \Delta$ (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 $\Gamma \nvDash^{STTT} \Delta$: 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 non-empty. Now, if A takes value 1 on any model in $\mathfrak{M}$, then $\Gamma, A \nvDash^{STTT} \Delta$, so we do not have a counterexample to generalized transitivity; similarly, if A takes value 0 on any model in $\mathfrak{M}$, then $\Gamma \nvDash^{STTT} A, \Delta$, so we again do not have a counterexample. It follows that, in any counterexample to generalized transitivity, A must take value ½ 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 — $\Gamma \vDash^{STTT} A, \Delta$ and $\Gamma, A \vDash^{STTT} \Delta$ but $\Gamma \nvDash^{STTT} \Delta$ — iff: there is some KK model that assigns 1 to everything in Γ and 0 to everything in Δ, and every such model assigns ½ to A. This is not a situation that often arises. In particular, it can be shown (by standard cut-elimination, for example), that this situation *never* arises when the arguments from Γ, A to Δ and from Γ to A, Δ are both classically valid. (For some other conditions also sufficient to guarantee transitivity, see Ripley 2012.) As a result, our endorsement of a non-transitive logic in no way amounts to a criticism of any classical uses of transitivity. We merely resist the assumption that transitivity can continue to operate freely once transparent truth is taken account of.[9]

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.[10]

It is important to be clear on the difference between a valid argument and a validity-preserving metainference, so we pause here to look at an example of each. Consider *modus ponens*. In its most basic form, it is an argument from premisses A and $A \supset B$ to the conclusion B. A logic can validate this argument or not; as examples, STTT validates every instance of it (so $A, A \supset B \vDash^{STTT} B$), and LPTT does not (so $A, A \supset B \nvDash^{LPTT} B$).

There is a metainference that also travels under the name 'modus ponens', however, and it is importantly distinct. This metainference moves from $\Gamma \vdash A$ and $\Gamma \vdash A \supset B$ to $\Gamma \vdash B$. Here, $\vdash$ should be read as a consequence relation; some consequence relations are, and some are not, closed under this metainference. For example, classical validity is closed under this metainference. (This is so whether classical validity is embodied by CL or by ST, since, as we mentioned above, these two give completely equivalent results.) That is, whenever both $\Gamma \vDash^{CL} A$ and $\Gamma \vDash^{CL} A \supset B$, it is also the case that $\Gamma \vDash^{CL} B$.

STTT, on the other hand, is *not* closed under this metainference. There are cases in which $\Gamma \vDash^{STTT} A$ and $\Gamma \vDash^{STTT} A \supset B$, but $\Gamma \nvDash^{STTT} B$.

[9] Thanks to Sam Butchart, Graham Priest, and an anonymous referee for discussion here.

[10] Sometimes 'rule' is used in the same sense, as an anonymous referee reminds us.

For example, consider the liar sentence λ, discussed above. As one can quickly verify, we have $\models^{STTT} \lambda$ and $\models^{STTT} \lambda \supset p$, but $\not\models^{STTT} p$. In fact, Negri and von Plato 2001, p. 19 show that this metainference is equivalent to generalized transitivity (given certain assumptions, which hold for STTT). Since STTT is not closed under generalized transitivity, we can conclude that it is also not closed under this metainference.

As this example demonstrates, it is possible to break a metainference by *adding* validities to a logic. Although every classically-valid argument is also valid in STTT, classical validity is closed under some metainferences that STTT validity is not closed under. The additional valid arguments in STTT give new opportunities for counterexamples to various metainferences. It is possible for a logic to retain all classically valid arguments across its full vocabulary while still failing certain classical metainferences, by adding new valid arguments, and STTT does just this. This immediately leads to two questions about STTT, one technical and one philosophical. First, just how many familiar metainferences does STTT fail? Second, how classical can STTT be if it fails metainferences that classical validity is closed under? Here, we answer each question in turn.

We have already seen two familiar metainferences failed by STTT: generalized transitivity and the metainferential relative of modus ponens. (We emphasize: modus ponens itself is an STTT-valid argument, as STTT preserves the validity of every classically-valid argument.) These two, we have also noted, are not independent. There is a third related issue as well: a bit of care is called for around the metainference of reductio. (Since double-negation rules hold without restriction in STTT, 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 preserves validity in STTT. In another familiar form, it moves from $\Gamma, A \vdash \bot, \Delta$ to $\Gamma \vdash \neg A, \Delta$; this form also preserves validity in STTT. In a third form, though, reductio moves from $\Gamma, A \vdash B \,\&\, \neg B, \Delta$ to $\Gamma \vdash \neg A, \Delta$, and this form does not preserve validity in STTT. (For example, $p \models^{STTT} \lambda \,\&\, \neg\lambda$, but $\not\models^{STTT} \neg p$.)

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 $\Gamma, A \vdash B \,\&\, \neg B, \Delta$ and $B \,\&\, \neg B \vdash \neg A$ that $\Gamma, A \vdash \neg A, \Delta$, or from $\Gamma, A \vdash B \,\&\, \neg B, \Delta$ and $B \,\&\, \neg B \vdash \bot$ that $\Gamma, A \vdash \bot, \Delta$; one is then in a position to apply one of the forms of reductio that does preserve validity in STTT. 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 is 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 STTT (for proofs, see Ripley 2012):[11]

Monotonicity:
If $\Gamma \models^{STTT} \Delta$, then $\Gamma, \Gamma' \models^{STTT} \Delta, \Delta'$.

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

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

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

Conjoining premisses, disjoining conclusions:
$\Gamma \models^{STTT} \Delta$ iff $\Gamma' \models^{STTT} \Delta'$, 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 is 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 and Hyde 1997) give up proof by cases and disjoining conclusions, the non-classical approaches in Beall 2009 and 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 classical theories of truth, such as the theory FS described in Friedman and Sheard 1987 and discussed in section 5.1, may retain all of these, but (as we will see presently) they must give up others. What is 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

[11] Sequent calculi are a way to present a logic almost entirely through metainferences, and Ripley 2013 shows that STTT retains all the rules of usual (cut-free) classical sequent calculi as well.

that $\Gamma \models \Delta$ for any Γ, Δ. (For further discussion of these theories, see Sect. 5.)

This is enough to give a sense of the situation with familiar metainferences in STTT. The question remains: Is it appropriate to see STTT as preserving classical logic, given that it fails some metainferences that preserve classical validity? 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 though STTT keeps all classically-valid arguments, if it loses some metainferences, then it might seem to have weakened some aspect of classical logic, and this could be enough to put it in with other non-classical approaches to paradox.

Even if this claim were right, it would not be too much trouble; it is not a bad crowd to be lumped in with. Non-classical 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 $\models^{S5} \Diamond p \supset \Box \Diamond p$ 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 premisses 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 $\vdash_1$ is closed under some metainferences that $\vdash_2$ is not closed under. That is, strengthening a logic always involves losing metainferences, unless we strengthen all the way to the universal consequence relation.

To see this, 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 $\vdash_1$ is closed under the metainference: if $\Gamma \vdash \Delta$, then $\Gamma' \vdash \Delta'$, but $\vdash_2$ is not. 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 — every time we strengthen a logic, we lose metainferences, so long as we do not strengthen all the way to the universal consequence relation.[12]

In other words, if STTT gives up something important about *T*-free classical logic, it cannot be 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 STTT, 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 STTT on the grounds of its non-transitivity, it cannot be denied that STTT is a conservative extension of classical logic that allows its users to recognize that *every* classically-valid argument is valid over the full vocabulary.[13] We take this to suffice for preserving classical logic.

3.4 Coding, induction, and compositionality

This far, we have worked with a simple quote-name approach, on which $\langle A \rangle$ names the wff A, and that is that. However, an ideal theory of truth should include more than this: we want a full theory of syntax. In this subsection, we will discuss how to achieve this within STTT. We use Peano arithmetic and Gödel coding to get the job done; for details, see e.g. Boolos 1995. We will write $\ulcorner A \urcorner$ for the code of a piece of vocabulary A. We use a predicate $\mathsf{sent}(x)$ true of all and only the codes of sentences, a predicate $\mathsf{var}(x)$ true of all and only the codes of variables, and functions $\dot{\neg}$, $\dot{\&}$, $\dot{\forall}$, and $\dot{T}$ such that for

[12] The S4/S5 example above fits this mould; so too does the following example. Classical predicate logic fails some metainferences that hold in classical propositional logic; for example, if $\forall xPx \vdash Pa$, then $p \vdash q$. It would be a serious abuse of terminology to hold that classical predicate logic does not preserve classical propositional logic 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 $\forall xPx$ as atoms.)

[13] An anonymous referee objects that arguments to $\perp$ stemming from semantic paradox are classically valid, and so, given our refusal to accept such arguments, we should not claim to preserve all classically-valid arguments. However, such arguments are not classically valid: they turn on applications of truth rules not contained in classical logic, and on chaining those applications together with classically-valid subarguments. (It is the chaining that we take to be the source of the problem.) As such, our avoidance of these problematic arguments is in no way a rejection of any classically-valid arguments.

any formulas A, B, and variable v: $\dot{\neg}\ulcorner A\urcorner = \ulcorner \neg A\urcorner$, $\ulcorner A\urcorner \dot{\&} \ulcorner B\urcorner = \ulcorner A \& B\urcorner$, $\dot{\forall}\ulcorner v\urcorner\ulcorner A\urcorner = \ulcorner \forall vA\urcorner$, 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 $\forall x \forall y(\mathsf{sent}(x \dot{\&} y) \supset (T(x \dot{\&} y) \equiv (Tx \& Ty)))$. 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 $\forall x(\mathsf{sent}(x) \supset (T\dot{\neg}x \equiv \neg Tx))$ and $\forall x \forall y(\mathsf{sent}(\dot{\forall}xy) \supset (T\dot{\forall}xy \equiv \forall t(y(t/x))))$, 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 $\models_{PA}^{STTT}$ of consequence analogously to $\models^{STTT}$, 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.[14] Thus, every instance I of the induction schema, even extended to those instances involving T, is such that $\models_{PA}^{STTT} I$; they are all theorems.

Moreover, the compositionality principles alluded to above are also theorems of $\models_{PA}^{STTT}$. 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 $\mathsf{sent}(x \dot{\&} y), T(x \dot{\&} y) \vdash Tx \& Ty$. It can be shown that (1) if these principles hold in PKF, then they hold in $\mathrm{STTT}_{\mathrm{PA}}$, and (2) if these principles hold in turnstile form in $\mathrm{STTT}_{\mathrm{PA}}$, then they hold in quantified theorem form as well (due to $\mathrm{STTT}_{\mathrm{PA}}$'s obeying a deduction theorem and allowing for the sequent metainference introducing $\forall$ on the right). As a result, STTT, when restricted to fixed points over the

[14] The instances are all of the form $(A(0) \& \forall x(A(x) \supset A(x+1))) \supset \forall x A(x)$. The only way for this sentence to get value 0 on a KKP model M is for $A(0) \& \forall x(A(x) \supset A(x+1))$ to get value 1 and $\forall x A(x)$ to get value 0. This cannot happen, given the constraints on $\supset$ and $\forall$ — and remembering that KKP models are built over the standard model of PA.

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 STTT-valid as well, and STTT 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 $\neg T\langle\lambda\rangle$:

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

If indeed $\top \models^{STTT} \bot$, 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 $\Gamma \models^{STTT} \Delta$ for every Γ, Δ. We know, since STTT conservatively extends classical logic, that this is not the case, but how is it avoided?

Every step in the above proof is STTT-valid: all but the T steps are classically valid, and the T steps are covered by transparency. (After all, $A \models^{STTT} A$, so transparency guarantees that $A \models^{STTT} T\langle A\rangle$ and $T\langle A\rangle \models^{STTT} A$.) It is the attempt to chain these steps together that has gone wrong, as we will presently show.

Remember, an STTT-valid argument is one that can never go from value 1 to value 0. The present argument, however, by moving from $\top$ to $\bot$, *always* goes from 1 to 0 — every KK model is a countermodel. Despite this, no KK model is a countermodel to any particular step of the argument. The descent from value 1 to value 0 happens in two stages, neither of which would be sufficient on its own. Let us look at this in more detail.

The first of these two stages is the application of LEM — concluding $T\langle\lambda\rangle \vee \neg T\langle\lambda\rangle$ from $\top$. As a classically-valid argument, this is

STTT-valid as well; it cannot go from value 1 to value 0. In this case, though, it goes from value 1 to value ½ on every KK model (since λ takes value ½ on every KK model). The second of the two stages is the application of Explosion — concluding $\bot$ from $T\langle\lambda\rangle \mathbin{\&} \neg T\langle\lambda\rangle$. As a classically-valid argument, this too is STTT-valid; it cannot go from value 1 to value 0. In this case, though, it goes from value ½ to value 0 on every KK model.

So although every step is valid — no step can go from 1 to 0 — chaining them together in this way has resulted in an invalid argument. The descent from 1 to 0 is split across different steps.

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

$$
\dfrac{
\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\top}{(T\langle\kappa\rangle \mathbin{\&} (T\langle\kappa\rangle \supset \bot)) \supset \bot}\,\text{PC}}{(T\langle\kappa\rangle \mathbin{\&} \kappa) \supset \bot}\,\text{Def. }\kappa}{(T\langle\kappa\rangle \mathbin{\&} T\langle\kappa\rangle) \supset \bot}\,\text{Transparency}}{T\langle\kappa\rangle \supset \bot}\,\text{PC}}{\kappa}\,\text{Def. }\kappa}{T\langle\kappa\rangle}\,\text{Transparency}
\qquad
\dfrac{\dfrac{\dfrac{\dfrac{\top}{(T\langle\kappa\rangle \mathbin{\&} (T\langle\kappa\rangle \supset \bot)) \supset \bot}\,\text{PC}}{(T\langle\kappa\rangle \mathbin{\&} \kappa) \supset \bot}\,\text{Def. }\kappa}{(T\langle\kappa\rangle \mathbin{\&} T\langle\kappa\rangle) \supset \bot}\,\text{Transparency}}{T\langle\kappa\rangle \supset \bot}\,\text{PC}
}{\bot}\,{\supset}\text{E}
$$

Again, every step is STTT-valid, but the proof seems to show that $\top \models^{STTT} \bot$. We know, since STTT conservatively extends classical logic, that this is not the case, so the trouble must have again come from linking the steps together. Although no single step can go from value 1 to value 0, the whole argument does manage to go from 1 to 0. Again, we can narrow the problem down to two steps, one of which goes from 1 to ½ and the other of which goes from ½ to 0. (Again, this works for every KK model, as all agree in assigning κ the value ½.) The descent from 1 to ½ happens in the first step of each subproof: $(T\langle\kappa\rangle \mathbin{\&} (T\langle\kappa\rangle \supset \bot)) \supset \bot$ only has value ½. The descent from ½ to 0 happens at the very end: both $T\langle\kappa\rangle$ and $T\langle\kappa\rangle \supset \bot$ have value ½, but $\bot$ always takes value 0. Again, the problem with this argument is not in any particular step, but rather in chaining these steps together.

Since STTT 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[15] will have treatments like the liar and Curry

[15] 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.

above; somewhere in the derivation of the troublesome conclusion, if every individual step is valid, there will be an illicit use of transitivity. The descent from 1 to 0 will not happen all at once, but it will happen bit by bit instead.[16]

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 ($\vDash^{STTT} \lambda$) and refutable ($\lambda \vDash^{STTT}$). Similarly, the claim that it is true is both a theorem and refutable, as is the claim that it is false. What do we say about such sentences, then?

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 the 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 2013, 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 STTT-based approach in terms of designated values, and we do not impose this two-way division.[17]

Instead, we can see a direct connection between model-theoretic value and assertibility. A sentence is either both strictly and tolerantly

[16] For example, in the Jones/Nixon case explored in Kripke 1975, if the circumstances are such as to render the case paradoxical, it will emerge that *both* Jones's and Nixon's utterances can be demonstrated to take value ½.

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

assertible (value 1), tolerantly but not strictly assertible (value ½), or not assertible at all (value 0). 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. 2012b. 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.[18]

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 is 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.

[18] If we like, we can call sentences whose tolerant meanings are true 'tolerantly true' and 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).

This reasoning, though, assumes transitivity throughout, and we have 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 STTT-based theory like the one we have 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 STTT as a formal approach to truth by comparing it and contrasting it to some of its relatives in the literature. One key difference between STTT and most other approaches is clear: transitivity. Almost all existing approaches to truth are based on transitive logics (but see Sect. 5.4), while STTT, quite crucially, is not. The other main distinction is STTT's preserving classical logic while adding transparent truth; no other theory combines these features.

5.1 FS

The first relative of STTT we look to is FS, or the Friedman-Sheard theory of truth. (This theory is presented in Friedman and Sheard 1987 and discussed in e.g. 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 (vitally, these are rules of *proof*, not of inference):

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

FS validates every classically-valid argument; in addition, if it is put in sequent form, it is closed under every rule of, say, Gentzen's LK. So it is quite classical indeed. It also validates the compositional principles that we might want to govern a truth predicate. But, as

is common among classical approaches, its approach to truth still falls short of what we might want, and short of what is achievable in STTT.

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 STTT. (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 STTT.

Further, FS is ω-inconsistent, and so can have no standard models. STTT, on the other hand, is shown to have standard models by the Kripke construction. In this regard, it is worth noting that STTT_{PA}, which contains the compositional principles, PA, and a transparent truth predicate, seems to more than satisfy 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 to STTT_{PA}, as McGee's argument depends on assuming transitivity.

5.2 Extra-arrow theories

One subfamily of STTT's non-classical relatives includes the logics of Priest 2006b, Beall and Ripley 2004, Brady 2006, Field 2008, and Beall 2009. While these logics differ from each other in various ways, their differences from STTT are more uniform; here, we will discuss them together, paying more attention to their common features than to what differentiates them.

Like STTT and unlike FS, most of these logics support full transparency.[19] All these logics include, in addition to the defined conditional $\supset$, a new conditional $\rightarrow$, and most validate the *T*-schema, at least in the form $A \leftrightarrow T\langle A\rangle$ (Beall and Ripley's system rather validates its contraposition, $\neg A \leftrightarrow \neg T\langle A\rangle$). Priest's, Beall and Ripley'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 non-classical logics; Priest's, Beall and Ripley's, and Beall's logics are extensions of LP, Field's is an extension of K3TT, and Brady's, as a relevant logic, is an extension of the logic FDE (see Anderson and Belnap 1975 or Priest

[19] Priest 2006b, Beall and Ripley 2004 are exceptions, for philosophical rather than technical reasons; we believe that transparency can be added to these systems without triviality.

2008 for details of FDE). The most apparent non-classicalities involve negation; none of the logics validates both excluded middle ($A \vdash B \vee \neg B$) and explosion ($A \& \neg A \vdash B$), and Brady's validates neither. The situation around reductio is also delicate. While the LP-based 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 five logics supports reductio in a different form. None allows passage from $\Gamma, A \vdash B \& \neg B, \Delta$ 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, STTT supports both excluded middle and explosion, as well as the first two forms of reductio. As we mentioned in section 3.3, it also does not support the third form of reductio — there, STTT 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 five 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 these 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 ($\vdash (A \& (A \rightarrow B)) \rightarrow B$); 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 $\vdash A \supset B$); all but Field fail the right-to-left direction, while Brady and Field both fail the left-to-right direction.

By contrast, STTT'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 STTT'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 STTT and these extra-arrow theories. Consider the liar sentence λ. Priest, Beall and Ripley, 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.[20]

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 2013 has also offered some arguments in favour of such a view.

Amongst non-classical approaches, this is probably the closest to STTT. 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 STTT have as a theorem every instance of the claim that modus ponens is truth-preserving: $\vdash (T\langle A \supset B\rangle \,\&\, T\langle A\rangle) \supset T\langle B\rangle$.[21]

There are some notable differences, however. First, $\mathbf{IKT}^{\omega}$ is weaker than classical logic, even on some very basic arguments: for example, $A \not\models^{\mathbf{IKT}^{\omega}} A \,\&\, A$, 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 validities is perhaps less drastic than in the case of many other non-classical systems, it is still very much a part of Zardini's approach.

Second, while $\mathbf{IKT}^{\omega}$ is known to be non-trivial, its relation to models of PA has not yet been explored. This leaves in question the status of the compositional principles mentioned in section 3.4. STTT, by building on the well-explored Kripke construction, can provide these principles.

[20] There is also a real connection 'under the hood'. The extra-arrow logics are proved non-trivial by 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 section 3.1; the transfinite series is only necessary to handle the extra arrow.

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

5.4 Non-transitive

Finally, we mention the relation between STTT and the non-transitive 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 is the closest to classical of the non-classical systems we consider here. However, it still exhibits some non-classical, and we think odd, behaviour.

A number of crucial arguments, such as modus ponens, are valid in Weir's logic only under restricted conditions. In addition, theoremhood cannot be defined in the usual way (being a consequence of the empty set of premisses); rather, Weir says, 'The notion of theoremhood … has to be: ϕ is a theorem iff for some A, B, we have that $A \rightarrow A, B \rightarrow B \vdash \phi$ is provable' (p. 246). (Here, $\rightarrow$ 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, STTT shows that there is no need to make these further modifications. It is possible, as we have shown here, to give up transitivity while preserving 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, STTT, 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, STTT 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 STTT.

It has been claimed in 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 STTT_{PA} (as in section 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. (STTT 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' (p. 283). We might yet live there, unless transitivity is seen as an additional desideratum. However, as we have tried to argue, the loss of transitivity is minimally disruptive; transitivity continues to hold in non-paradoxical cases.

There is much left to do. We have not here explored an STTT-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. 2012b, Cobreros et al. 2012a, although the approaches are intimately related. Although we have 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 STTT on the table as suggesting a promising avenue for approaching the paradoxes.[22]

References

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

Beall, J. 2009: *Spandrels of Truth*. Oxford: Oxford University Press.

Beall, J. and B. Armour-Garb (eds) 2005: *Deflationism and Paradox*. Oxford: Oxford University Press.

[22] For helpful discussions, we would like to thank audiences at Truth Be Told 2011, Truth at Work 2011, and the Seventh Barcelona Workshop on Issues in the Theory of Reference. Conversations with Sam Butchart, Hartry Field, Michael Glanzberg, Anil Gupta, Volker Halbach, Leon Horsten, Hannes Leitgeb, Toby Meadows, Graham Priest, Greg Restall, Philippe Schlenker, Zach Weber, and Elia Zardini greatly helped the paper along. We are also grateful for detailed and insightful comments from an editor and four anonymous referees. This research was partially supported by the Agence Nationale de la Recherche, grant ANR-07-JCJC-0070, programme 'Cognitive Origins of Vagueness', as well as by grants ANR-10-LABX-0087 IEC and ANR-10-IDEX-0001-02 PSL*, and by the Government of Spain, programme 'Borderlineness and Tolerance' ref. FFI2010-16984, MICINN.

Beall, J. and J. Murzi 2013: 'Two Flavors of Curry's Paradox'. *Journal of Philosophy*, 110, pp. 143–65.
Beall, J. and D. Ripley 2004: 'Analetheism and Dialetheism'. *Analysis*, 64, pp. 30–35.
Boolos, G. 1995: *The Logic of Provability*. Cambridge: Cambridge University Press.
Brady, R. T. 1989: 'The Non-triviality of Dialectical Set Theory'. In Priest, Routley, and Norman 1989, pp. 437–71.
—— 2006: *Universal Logic*. Stanford, California: CSLI Publications.
Cobreros, P., P. Égré, D. Ripley, and R. van Rooij 2012a: 'Tolerance and Mixed Consequence in the S'valuationist Setting'. *Studia Logica*, 100, pp. 855–77.
—— 2012b: 'Tolerant, Classical, Strict'. *Journal of Philosophical Logic*, 41, pp. 347–85.
Dunn, J. M. and G. M. Hardegree 2001: *Algebraic Methods in Philosophical Logic*. Oxford: Oxford University Press.
Field, H. 2008: *Saving Truth from Paradox*. Oxford: Oxford University Press.
Frankowski, S. 2004: 'Formalization of a Plausible Inference'. *Bulletin of the Section of Logic*, 33, pp. 41–52.
Friedman, H. and M. Sheard 1987: 'An Axiomatic Approach to Self-referential Truth'. *Annals of Pure and Applied Logic*, 33, pp. 1–21.
Gupta, A. and N. Belnap 1993: *The Revision Theory of Truth*. Cambridge, Massachusetts: MIT Press.
Halbach, V. 2011: *Axiomatic Theories of Truth*. Cambridge: Cambridge University Press.
Halbach, V. and L. Horsten 2006: 'Axiomatizing Kripke's Theory of Truth'. *Journal of Symbolic Logic*, 71, pp. 677–712.
Hyde, D. 1997: 'From Heaps and Gaps to Heaps of Gluts'. *Mind*, 106, pp. 641–60.
Kleene, S. C. 1952: *Introduction to Metamathematics*. Amsterdam: North-Holland Publishing Co.
Kremer, M. 1988: 'Kripke and the Logic of Truth'. *Journal of Philosophical Logic*, 17, pp. 225–78.
Kripke, S. 1975: 'Outline of a Theory of Truth'. *Journal of Philosophy*, 72, pp. 690–716.
Leitgeb, H. 2007: 'What Theories of Truth Should Be Like (but Cannot Be)'. *Philosophy Compass*, 2, pp. 276–90.
Maudlin, T. 2004: *Truth and Paradox*. Oxford: Oxford University Press.

McGee, V. 1985: 'How Truthlike can a Predicate Be? A Negative Result'. *Journal of Philosophical Logic*, 14, pp. 399–410.
Moschovakis, Y. N. 1974: *Elementary Induction on Abstract Structures*. Amsterdam: North-Holland Publishing.
Nait-Abdallah, A. 1995: *The Logic of Partial Information*. Berlin: Springer.
Negri, S. and J. von Plato 2001: *Structural Proof Theory*. Cambridge: Cambridge University Press.
Priest, G. 2006a: *Doubt Truth to be a Liar*. Oxford: Oxford University Press.
—— 2006b: *In Contradiction*. Oxford: Oxford University Press.
—— 2008: *An Introduction to Non-Classical Logic: From If to Is*, 2nd edition. Cambridge: Cambridge University Press.
Priest, G., R. Routley, and J. Norman (eds) 1989: *Paraconsistent Logic: Essays on the Inconsistent*. Munich: Philosophia Verlag.
Quine, W. V. O. 1970: *Philosophy of Logic*. Englewood Cliffs, NJ: Prentice-Hall.
Ripley, D. 2012: 'Conservatively Extending Classical Logic with Transparent Truth'. *Review of Symbolic Logic*, 5, pp. 354–78.
—— 2013: 'Paradoxes and Failures of Cut'. *Australasian Journal of Philosophy*, 91, pp. 139–64.
Weir, A. 2005: 'Naive Truth and Sophisticated Logic'. In Beall and Armour-Garb (eds) *Deflationism and Paradox*, pages 218–49.
Zardini, E. 2011: 'Truth Without Contra(di)ction'. *Review of Symbolic Logic*, 4(4), pp. 498–535.

