Chapter 4
Truth and Trustworthiness

Michael Sheard

Abstract In the course of ordinary communication, people transmit messages (i.e., say things) which may involve the application of a truth predicate. The receiver of such a message needs to have a method which allows the extraction of non-truth-theoretic information from uses of the truth predicate; such a method can be modeled with an axiomatic system. On close examination, the choice of which axiomatic system to employ can be seen to depend on whether or not the source of the message is considered trustworthy—that is, whether the information in the message can simply be accepted, or if it must first be examined for consistency with previously known information and, on the basis of that determination, possibly be rejected. This paper explores some of the consequences involved in this framework.

4.1 Setting the Problem

Three philosophers—a deflationist, an advocate of the correspondence theory of truth, and a believer in the coherence theory—go out to dinner. They have a riotously good time debating politics, sports, gossip about other employees of their university—anything except philosophy. Throughout the evening, their conversation is peppered with phrases like "That's true!", "That can't be true," and "Nothing that man has ever said is true." Remarkably, they all understand each other completely at those moments, even though they do not agree in the least on what it means for something to be "true". How is it that they are able to communicate so effectively using a concept about which they so thoroughly disagree?

To a deflationist, of course, there is nothing surprising here. Since truth (for a deflationist) is a logical or linguistic concept that operates at a surface level, of course it can be employed in ordinary conversation without need of deep analysis or occasion for disagreement. But while the deflationist may not be surprised *that* the process works, such confidence is not the same as having an explanation of *how* the process works. Meanwhile, a substantivist may maintain very emphatically that at its core, truth is a much deeper phenomenon that the deflationist mistakenly dismisses.

M. Sheard
Rhodes College, Memphis, TN, USA
e-mail: sheardm@rhodes.edu

T. Achourioti et al. (Eds.), *Unifying the Philosophy of Truth,* Logic, Epistemology, and the Unity of Science 36, DOI 10.1007/978-94-017-9673-6_4

Nonetheless, all but the most extreme substantivist will have to admit that, whatever the deep issues may be, those issues are not engaged in a meaningful way when people use words like "true" in the course of ordinary conversation. While truth may be deep, it must have some shallow features that allow it to serve as a mechanism of day-to-day speech. In either case, we still have work to do to uncover how truth functions in ordinary conversation. We need a logical system that can model the way the concept of truth is used to convey information.

If we were to stipulate that a truth predicate can only be applied to sentences that do not involve truth itself, then there would be few problems. In formal terms, we could adopt as an axiom the Tarski T-sentence T('A') ↔ A for each sentence A which does not contain the truth predicate (e.g., " 'snow is white' is true if and only if snow is white"), and modulo a simple ability to unpack sentences, we would be done. This restriction is unrealistic, however. I ought to be able to say that everything Bob said in his lecture today was true, even if one of the things Bob said was that Emily's statement was true. Applying the truth predicate to sentences that themselves involve the truth predicate is completely natural. We need an analysis which is robust enough to account for an untyped truth predicate.[1]

Obviously, however, insistence on an untyped truth predicate raises other problems. Given our ability to formulate Liar-like sentences, both in ordinary language and in its formal analogues, we know that there are instances of the unrestricted T-sentences which lead in short order to contradictions. One response (one which I will admit I find somewhat attractive) is to grant that people actually work day-to-day on a basis of the unrestricted T-sentences, and simply do not carry their reasoning with inconsistent hypotheses far enough to derive explicit contradictions.[2] As a practical matter, this is a reasonable position to maintain about actual people, as part of a much more general phenomenon—probably most people hold inconsistent beliefs of one sort or another, yet rarely get into trouble simply because they do not reason through to an explicit contradiction. While there may be merit in this hypothesis, it is unsatisfactory as an exercise in logical modeling. If we adopt this attitude, there appear to be only two places to go next. One is to announce that humans are irrational creatures and walk away, which does nothing useful to address the original question about how people are able to communicate effectively using the concept of truth. The other option is to begin an empirical study of how real people manage, maintain, and apply actual inconsistent assertions about truth, but such a study is more suitable for psychology than for philosophy. Moreover, there is nothing in such a project that is specific to the study of the logic of truth *per se*, since the same questions could be asked any time someone holds inconsistent beliefs in any context. Neither of these approaches advances our understanding of how a truth predicate can be used in communication.

Instead, then, let us assume that each person has some logically consistent framework for processing information that is transmitted by means of a truth predicate.

[1] Kripke (1975) makes this point in much more detail.

[2] Horwich (1990) speaks about our *inclination* to accept the T-sentences.

Becoming more formal, we can capture this framework as a set of axioms and rules of inference concerning truth which are overlaid on whatever base of factual information the person has and whatever ordinary rules of logical inference the person employs. When someone speaks, we can think of what he says as a message which is transmitted with the intent that it be added to the listener's base of factual information. The listener applies her own truth-specific axioms and rules to extract the non-truth-theoretic content of the message. Our goal will be to explore what axioms and rules are needed for this process, and how they are to be applied.

4.2 The Form of a Message

Ideally, a successful analysis would account for all possible uses of a truth predicate in ordinary communication. There are good reasons, however, for believing that such a far-reaching goal may simply be impossible. Our aspirations will be more modest at the outset, so let us focus on three very common uses of truth in communication:

1. Direct attribution of truth
2. Denial
3. Generalization

Let us briefly consider each of these in turn.

A direct attribution of truth states that a specified sentence is true, and can take several forms. Direct attribution can occur in quotational (or equivalent) form, in which the sentence to which truth is being attributed is immediately displayed: "The sentence 'Amsterdam is in the Netherlands' is true". While there are grammatical, linguistic, and perhaps logical differences between this example and "It is true that Amsterdam is in the Netherlands", it is hard to argue that they convey any different information, either implicitly or explicitly.[3] Note that in this regard attributions of truth differ from some other kinds of linguistic communication. There is a meaningful distinction between "Catherine said 'Amsterdam is in the Netherlands' " and "Catherine said that Amsterdam is in the Netherlands"—are we repeating her exact words, or paraphrasing?—but there is no difference in the context of our analysis of the uses of truth in communication.

Direct attribution need not be quotational or nearly-quotational. It can proceed by indexical reference: "That is true", where the referent of "that" is unambiguous in context. It can proceed by some sort of definite description: "The first sentence in Susan's essay is true." It can also proceed—perhaps a bit artificially—via description of the construction of the sentence to which truth is attributed; such an approach is valuable in creating unassailably self-contained examples of self-reference, like the Liar sentence.

[3] The explicitly quotational version as written does imply, or at least seems to assume, that sentences are appropriate bearers of truth. That discussion can be left for a different occasion.

One may wonder why direct attribution of truth would pose a problem at all in the context of communication. At some level, of course, it does not. The redundancy theory, the disquotational theory, the prosentential theory, and the minimal theory of truth will all tell you that to communicate a direct attribution of truth is to communicate the underlying sentence itself. If we could stop there, the answer would be fully sufficient. When we build a theory robust enough to handle denials and generalizations, however, the mechanisms we put in place may not be sufficient to achieve what we would hope to achieve in our analysis of direct attribution.

It is tempting to regard a denial of truth as an immediate variant of a direct attribution of truth—no more and no less problematic. Certainly attributing falsity to a statement has much the same feel as attributing truth—the same action with just a negation operator inserted at some appropriate point. In fact, though, the nature of that negation operation, and the question of exactly what is the "appropriate" point for its insertion, turn out to make a huge difference in the logical analysis; this observation will come into sharper focus later. For now, let us just look at what happens when we entertain the possibility that a sentence might—for whatever reason—lack a truth value, that is, might be neither determinately true nor determinately false. Certainly there are theories of truth that include the possibility of indeterminate truth values in their structure. A direct attribution of truth unquestionably asserts that the sentence in question has a determinate truth value. But if someone says that a sentence is "not true", does that mean "false or possibly indeterminate"? Or does it mean "determinately false"? The former seems a more reasonable reading on purely logical grounds (given that we already accept that there is a category of sentence which is neither true nor false), but possibly not in the spirit of the way the expression might be used in ordinary communication.[4] If someone chooses to say that a sentence is "not true" —rather than "meaningless", or "impossible to determine", for example—might we not be justified in the inference that the speaker believes that the sentence in question does indeed have a determinate truth value? There are many other ways to signal that a statement fails to have a truth value for some intrinsic reason, which are more specific and perhaps more sensible than simply saying it is "not true". Alternatively, if we decide that we are not justified in assuming that "not true" excludes the possibility of indeterminacy, then can we run the train in the other direction, and also interpret "false" not as "determinately false" but simply "not true"? After all, if asked, most people would define the word "false" as "not true".

We may choose to dismiss these questions as a linguistic muddle arising from the ambiguity of ordinary language, but we will have no such luxury when we try to formalize the logic which is used in the process of communication. In any case, what is apparent is that the question of denial in the use of a truth predicate is *not* merely the mirror image of the question of direct attribution of truth. For these reasons, it is appropriate to keep denial as a separate prototypical example.

[4] Similarly, in English, "I don't think it will work" really means "I think it will not work."

Last, we come to generalization. Some philosophers have claimed that generalization is the whole reason that there is a problem about the theory of truth at all; if all we had available in our language were direct attributions and denials, most of the problems would disappear via some sort of redundancy interpretation, quibbles around the margins notwithstanding. At the very least, the problems to be solved without generalization would be substantially smaller in number and magnitude. Generalization applies the truth predicate to a defined list of sentences, where membership on the list is given by specification of a shared property rather than enumeration: "Every sentence in the book is true." It is worth exploring different features that the specification can have. If the specification unambiguously specifies a finite list, then the generalization can be read as a mere stand-in for the conjunction of direct attributions of truth to each of the sentences individually: "The first sentence in the book is true and the second sentence in the book is true and . . ." If the list is actually or potentially infinite, then this interpretation cannot be sustained: the claim "Every theorem of Peano Arithmetic is true" attributes truth to a list of sentences which is not only infinite but also provably undecidable. Finally, there is the situation in which the list, while presumably finite, is not known or not yet established. "Everything Bob has ever said about chemistry is true" specifies a large finite list of which neither the speaker nor the listener is likely to know the exact membership. "Everything Bob will ever say about chemistry will be true" specifies a list which does not (yet) even have an exact membership. Why the speaker would make such a reckless claim may be an interesting question, but it in no way impinges on the use of the truth predicate itself as a way of conveying information. The content the speaker aims to convey is clear enough.

4.3 Logical Systems

Now we need a logical system for our idealized listener to apply to decode incoming messages. There are three standout candidates for the role: the systems known in the literature as FS, KF, and VF.[5] (To be precise, these designations all denote systems of arithmetic augmented with a truth predicate, in which the underlying axiomatization is Peano Arithmetic. Working informally, I will use the same designations for the corresponding truth-theoretic apparatus overlaid on any system with sufficient expressive power to permit discussion of linguistic elements like sentences, which is a necessary precondition for applying a truth predicate.) While all three have been studied thoroughly for their truth theoretic properties, and KF in particular has been the focus of some discussion concerning its suitability as a theory *about* truth (most notably by Reinhardt (Reinhardt 1986)), there has not been much discussion of their

[5] For a comprehensive presentation of all of these systems, see Halbach (2011).

relative merits as a tools for the kind of communication described here. Let us take a quick look at the principal features of each.[6]

FS is Halbach's axiomatization (Halbach 1994) of a theory which replaces the T-sentences with corresponding rules of inference:

From A, deduce T(‘A’)
From T(‘A’), deduce A
From ¬A, deduce T(‘¬A’)
From T(‘¬A’), deduce ¬A

In the analysis that follows, it is important to remember that these rules can only be applied to sentences that are given as axioms or have already been proved; they may not be used in conditional subproofs. Thus in general there is no deduction theorem for FS.

KF is Feferman's axiomatization (Feferman 1991) of Kripke's basic fixed-point model. It is most smoothly axiomatized with both a truth predicate and a falsity predicate. The salient axioms are compositional; for example, here are the axioms for the truth and falsity of negations, conjunctions, and truth-attributions:

T(‘¬A’) ↔ F(‘A’)
F(‘¬A’) ↔ T(‘A’)
T(‘A&B’) ↔ T(‘A’) & T(‘B’)
F(‘A&B’) ↔ F(‘A’) ∨ F(‘B’)
T(‘T(A)’) ↔ T(‘A’)
F(‘T(A)’) ↔ F(‘A’)

There are similar sets of axioms for disjunctions, material conditionals, quantifiers, and falsity-attributions. In addition, KF contains the T-Consistency axiom: ¬(T(A) & T(¬A)). One consequence of the T-Consistency axiom is that by induction on the build-up of formulas, one can prove T(‘A’) → A (that is, one direction of the biconditional T-sentence) for each sentence A. Critically, KF does not assert the truth of validities of first-order logic: for one significant example, not every sentence of the form T(‘A∨¬A’) is provable.

VF is Cantini's axiomatization (Cantini 1990) of the Kripke/ van Fraassen supervaluation model. The central axiom of VF is again the one direction of the T-sentences: T(‘A’) → A, for all sentences A. Unlike KF, it also contains axioms guaranteeing that the set of true sentences is closed under logical implication, although in exchange it gives up some of the principles of compositionality, such as T(‘A∨B’) ↔ T(‘A’) ∨ T(‘B’).

[6] Each system has additional axioms, including ones that establish which basic sentences are to be declared true. For simplicity, I will suppress mention of most of those here. Moreover, my notation is intentionally over-simplified in some regards to improve readability.

4.4 Trustworthiness

If a speaker conveys a message with intent that it be added to a listener's base of factual knowledge, the listener faces a decision. If (in our idealized model) it can be assumed with certainty that the message will not be in conflict with information already known to the listener, then the listener can decode the message and add the new message directly to the knowledge base. I will call the source of the message in this case *trustworthy*. Alternatively, if there is no assumption that the message will not be in conflict with preexisting knowledge, then the user may need to evaluate the information as an additional step in the process, and perhaps may even draw inferences from the outcome of that evaluation (such as perhaps concluding that the speaker is a liar). I will call such a source *untrustworthy*. The distinction will matter in selecting an appropriate logical system for the task.

4.5 Decoding the Messages

Let us look first at the situation of a trustworthy source, and consider in turn each of the three principal kinds of messages. Imagine that a trustworthy source sends a message which makes a direct attribution of truth, which we can represent as T(‘A’). All three systems are strong enough extract the information conveyed. In FS, one applies the rule of semantic descent to T(‘A’) to derive A. In VF or KF, one grabs the axiom/theorem T(‘A’)→A and applies *modus ponens*.

As I have suggested already, the situation for a denial is a little more delicate. If we represent a denial as ¬T(‘A’), then FS is fully equipped to handle it: there is a rule of inference to derive ¬A. The systems VF and KF, however, pose more of a stumbling block. Both systems prove all instances of T(‘A’)→A, but not in general A→T(‘A’), so that there is no generally valid way to deduce ¬A from ¬T(‘A’). Here, perhaps the best solution is to fall back on the suggestion to render a denial as T(‘¬A’) rather than ¬T(‘A’), which solves the problem immediately. It also can be applied uniformly to include FS, since the inference from T(‘¬A’) to ¬A is also valid there.

Finally, generalization turns out to pose little problem. The formal representation of the use of truth for generalization as we have defined it has the form ∀x(R(x)→ T(x)). If ‘A’ is any sentence which falls into the list defined by R(x), then all three systems will allow one to move directly from R(‘A’) and ∀x(R(x)→ T(x)) to T(‘A’). One can then apply the system's method for handling direct attribution of truth to extract A.

If a source is not trustworthy, then any message received from it needs to be checked for the possibility of inconsistency with existing knowledge (or internal self-contradiction) before it can be added to the listener's base of knowledge. For systems like KF and VF which are purely axiomatic—i.e., having no additional auxiliary rules of inference—this is simple, since they are closed under *reductio ad absurdum*: if sentence B (whether truth-theoretic or not) is inconsistent with existing

knowledge, then already ¬B is a logical consequence of existing knowledge. In principle, then, under these systems screening information from an untrustworthy source is logically straightforward. Any sentence, whether containing a truth predicate or not, can be accepted if it is not a direct contradiction of a known consequence of existing knowledge.

For FS, with its auxiliary rules of inference that do not admit a deduction theorem, such a *reductio* is not available in general. To take an extreme example, let L be the Liar sentence. If one tries to add L to an existing knowledge base, then the rule of semantic assent applied to L produces T('L'), and an immediate contradiction. Nonetheless, ¬L is not a theorem of FS, nor can it even be a member of any consistent set of sentences closed under the rules of FS, since it leads to a contradiction of its own in much the same manner. In the end, however, the difference is of minor import for the purposes of evaluating messages. As a model, one can envision a person who uses the rules of FS as a truth mechanism provisionally accepting a message from an untrustworthy source, and then testing to see if a contradiction emerges. If one does, then the new information and anything that followed from it is rejected and the *status quo ante* is restored. If no contradiction emerges, then the new information remains. For a simple decision about accepting or rejecting a message, the full force of a *reductio* argument is not needed.

4.6 Next Steps

So far, the distinctions proposed here may seem to be much ado about little. One does not have to go much further beyond the restricted realm of direct attribution, denial, and generalization, however, to reach a point where the complications begin to mount up. Consider the case of a corporate rumor claiming that Smith lied to the president of the company, to which the wise and trustworthy old-timer announces, "If the rumor is true, then Smith will be fired." The wise and trustworthy old-timer's statement has the form T('A') → B, but the obvious inference to A → B is beyond the scope of the basic inferential mechanism of *any* of our three systems. The system FS fares slightly better than the others, in that if it turns out to be the case that Smith did indeed lie to the president of the company, then FS can deduce T('A') from A by semantic ascent, and then draw the accurate conclusion that Smith is on his way out. Without semantic ascent, KF and VF cannot do even that, unless someone chooses to announce specifically that the rumor was true.

Before reading too much into the preceding example, however, note that FS tends to underperform in situations where a message from an untrustworthy source can be examined logically not just for acceptance or rejection, but as a basis of logical inferences to acquire new additional information. Elsewhere (Sheard 2008) I have offered as an example a variation of a problem of Smullyan (Smullyan 1978):

> There is an island on which every inhabitant either always tells the truth or always lies, but the two types are otherwise indistinguishable. You encounter two of them; one says "At least one of us always lies." Which type is each of them?

In either VF or KF, some easy conditional reasoning allows one to answer the question. (The speaker is a truth-teller.) In FS, however, without *reductio* available for conditional inferences involving the truth predicate, the speaker's statement, while consistent and therefore not to be rejected out of hand, cannot be followed to its apparent logical conclusion. In this example FS fails a simple test for suitability as a logical system for reasoning about information from untrustworthy sources.

As these examples suggest, it is likely that in the end no axiomatic system will prove ideal for handling all reasonable uses of a truth predicate in the communication process. By carving out a large and productive fragment of instances where the tools available to us *can* be applied, and by uncovering issues like trustworthiness that shape the context in which these systems operate, we should be in a better position to assess the merits of axiomatic systems for truth.

References

Cantini, A. (1990). A theory of formal truth arithmetically equivalent to ID_1. *Journal of Symbolic Logic, 55,* 244–259.

Feferman, S. (1991). Reflecting on incompleteness. *Journal of Symbolic Logic, 56,* 1–49.

Halbach, V. (1994). A system of complete and consistent truth. *Notre Dame Journal of Formal Logic, 35,* 311–327.

Halbach, V. (2011). Axiomatic theories of truth. Cambridge: Cambridge University Press.

Horwich, P. (1990). Truth. Oxford: Basil Blackwell.

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

Reinhardt, W. N. (1986). Some remarks on extending and interpreting theories with a partial predicate for truth. *Journal of Philosophical Logic, 15,* 219–251.

Sheard, M. (2008). A transactional approach to the logic of truth. In C. Dimitracopoulos, L. Newelski, D. Normann, & J. Steel (Eds.), Logic colloquium 2005 (pp. 202–220). Cambridge: Cambridge University Press.

Smullyan, R. (1978) What is the name of this book? The riddle of dracula and other logic puzzles. Englewood Cliffs: Prentice-Hall.