Part I

FOUNDATIONS

1 Definitional and axiomatic theories of truth

Philosophers have been very optimistic about the prospects of defining truth. The explicit definability of truth is presupposed in many accounts of truth: only whether truth is to be defined in terms of correspondence, utility, coherence, consensus, or still something else remains controversial, not whether truth is definable or not. The advocated definitions usually take the form of an explicit definition. Hence, if one of these proposed definitions is correct, truth can be fully eliminated as explicit definitions allow for a complete elimination of the defined notion (at least in extensional contexts). It is a quirk in the history of philosophy that many of these definitional theories, according to which truth is eliminable by an explicit definition, have come to be known as *substantial* theories as opposed to *deflationary* theories of truth, although most proponents of deflationist accounts of truth reject explicit definitions of truth and in most cases also the eliminability of truth.

A common complaint against traditional definitional theories of truth is that it is far from clear that the *definiens* is not more in need of clarification than the *definiendum*, that is, the notion of truth. In the case of the correspondence theory one will not only invoke a predicate for correspondence, but one will also use facts or states of affairs as relata to which the objects that are or can be true are supposed to correspond; in the case of states of affairs one will then also have to distinguish between states of affairs that obtain and those that do not. Of course, proponents of the various varieties of the correspondence theory propounded theories of facts, states of affairs, obtaining, and correspondence in which the assumptions on which they rely in their reasoning about facts and states of affairs are made explicit. But these theories are controversial at best and most people are much clearer and firmer in their views about truth itself than in their views about facts and states of affairs. Therefore it seems sensible to make explicit the assumptions about truth rather than to take the detour via a definition in terms of notions less accessible than truth.

The decision to take truth as a primitive notion that is not defined in terms of other notions need not necessarily clash with definitional approaches. To begin with, one can take truth to be a primitive notion and postulate certain

principles or axioms for truth without taking a stance towards the question whether truth is definable or not. Choosing an axiomatic approach might well be compatible with the view that truth is definable; the definability of truth is just not presupposed at the outset. So an axiomatic approach might only differ from a definitional account in its methodology, and in the end both might converge to the same theory of truth.

I do not think, however, that there is only a methodological motivation for an axiomatic approach to truth. In this respect the situation with truth is fundamentally different from knowledge, for instance. In the case of knowledge, epistemologists have tried for a long time to provide an adequate definition in terms of truth, belief, justification, and some further condition that allows one to handle Gettier cases. Providing an adequate definition of knowledge has proven to be very hard and some epistemologists have abandoned the enterprise of finding such: some have declared the notion of knowledge to be marginal and put justification at the centre of epistemology; still others are happy to study knowledge as a primitive notion. The main reason to view knowledge as a primitive notion and to doubt that definitional theories are feasible seems to be that nobody has been able to come forward with an generally accepted definition of knowledge. The main evidence for the undefinability of knowledge is the observation that convincing counterexamples are known against most if not all proposals for explicit definitions that are non-trivial.

In the case of truth, in contrast, not only is there similar evidence that truth cannot be defined for a language within that language, but there is a theorem: Tarski's theorem on the undefinability of truth rules out the possibility of a definition of truth under certain conditions. It states that under fairly generally applicable conditions, the assumption that there is a definition of truth within a given theory for the language of that same theory leads to a contradiction.

For a sketch of Tarski's theorem I assume that a classical first-order language $\mathcal{L}$ is fixed and that it contains a closed term $\ulcorner e \urcorner$ for each expression e of the language $\mathcal{L}$. If a consistent theory $\mathcal{S}$ in the language $\mathcal{L}$ can prove certain basic facts about substitution of expressions in expressions and if it can describe a function taking each object to a closed term for that object, then there cannot be a formula $\tau(x)$ of $\mathcal{L}$ such that $\tau(\ulcorner \phi \urcorner) \leftrightarrow \phi$ is provable in $\mathcal{S}$ for each sentence ϕ of $\mathcal{L}$. But many philosophers agree with Tarski (1935) that a theory of truth for the language $\mathcal{L}$ should at least prove these equivalences, which are called T-sentences or disquotation sentences depending on how exactly they are formulated.

I will not try to make Tarski's theorem more precise, although marking out the limits of Tarski's theorem would be worthwhile as it would illustrate just how widely applicable it is (but see almost any textbook on Gödel's incompleteness theorems for an account of Tarski's theorem). The amount of syntax needed is very little and can be represented in very weak arithmetical theories.

If closed terms for the expressions of the language $\mathcal{L}$ are not available in $\mathcal{L}$, then the above equivalences can be replaced with the following claims:

$$\forall x \left(\mathrm{Sent}_{\phi}(x) \rightarrow (\tau(x) \leftrightarrow \phi) \right)$$

In these sentences the formula $\mathrm{Sent}_{\phi}(x)$ expresses that x is the sentence ϕ or its code. So the notation becomes more convoluted, but Tarski's theorem can be proved for languages like that of set theory that lack closed terms for sentences or their codes.

Also Tarski's theorem does not rely on any assumption about what a definition would look like, except that it would have to yield the disquotational or T-sentences. It is often stated for arithmetical theories and used to show that arithmetical resources do not suffice for defining the truth of arithmetical sentences. But Tarski's theorem applies in other settings as well. For instance, the proof of Tarski's theorem does not require that truth be attributed to sentences, or their arithmetical codes. If truth is attributed to propositions and the operations on propositions corresponding to the syntactic operations mentioned above can be expressed in the theory, then Tarski's theorem can be proved for such a setting as well. If the underlying theory $\mathcal{S}$ contains an axiomatic account of propositions, facts, states of affairs, or the like, Tarski's theorem shows that truth for the propositions of the theory cannot be defined on the basis of the theory $\mathcal{S}$. So Tarski's theorem does not affect traditional definitional theories of truth any less than more mathematical theories, although the impact of Tarski's theory may be felt less in the case of traditional theories because they are often presented in terms that are vague enough to make an application of formal results appear impossible or at least implausible. But Tarski's theorem applies to any sufficiently precise version of the correspondence theory of truth and all the other traditional theories of truth. At any rate, Tarski's theorem is a threat to all definitional theories whether they rely on a notion of correspondence or some other notion.

I do not want to claim that any satisfactory theory of truth has to prove the equivalences $T\ulcorner\phi\urcorner \leftrightarrow \phi$, but if it does, Tarski's theorem strikes and clearly truth cannot be defined in the sense just sketched, at least in a setting satisfying certain minimal conditions. If it is assumed that a theory of truth

has to be a definition of truth, then one is excluding many theories, and in fact many of the theories that will be studied below. All I am asking for is that undefinable notions of truth are not excluded from the outset as suitable accounts of truth. First one should become clear about which properties a notion of truth should have. Once one has become clear about what is expected from the notion of truth, one can investigate in a second step whether truth is definable.

In semantic theories of truth – by these I mean Tarski's theory but also, for instance, Kripke's (1975) – truth is defined. If the metalanguage contains the object language, the equivalence $T\ulcorner\phi\urcorner \leftrightarrow \phi$ will be provable for all sentences ϕ of the object language but not for all sentences of the metalanguage. So Tarski's theorem is evaded by restricting the possible instances in the schema $T\ulcorner\phi\urcorner \leftrightarrow \phi$ to sentences of a proper sublanguage of the language in which the equivalences are formulated.

So it might seem that in semantic truth theories one can proceed in a different order: the definition of truth comes first, and only after truth has been defined, one explores the consequences of the definition. But in fact when one is looking at the various semantic theories of truth, they very often start from certain assumptions about truth. Philosophers often appraise semantic theories by pointing out that certain sentences or, as I would like to call them, principles are satisfied in the proposed semantics. Tarski, for instance, justifies his semantic theory by pointing out that his defined truth predicate satisfies the above mentioned equivalences for all sentences of the object language (see Chapter 3 below for a discussion of Tarski's theory). So a certain syntactic principle stands at the beginning. Tarski's definition of truth is then designed to show that a notion of truth satisfying the equivalences can be defined in the metalanguage, or more precisely, in a metatheory, assuming again that the metalanguage contains the object language.

I propose then to focus on these principles – whether they are Tarski's equivalences or some other principles – and to discuss them before trying to eliminate them, for instance, by providing model-theoretic semantics for a language satisfying these principles or by defining truth in terms of correspondence.

One reason for pausing at this stage is that there is little agreement over which principles should be adopted. As I will show in Chapter 3, Tarski took his own equivalence to be insufficient. So even he did not fully believe in the adequacy of his principles. In the meantime many logicians and philosophers have rejected Tarski's approach as insufficient because it excludes any application of the truth predicate to sentences that contain it. This has been

the point of devising semantic theories of self-applicable truth; it is easy to show that Tarski's own restrictive solution is neither plausible nor useful for many purposes. So I would like to discuss first the principles that should be satisfied by truth as there is such a wide variety of them.

Moreover, once these principles have been formulated, defining the truth predicate contained in them is not the only way to eliminate the notion of truth. It is not too hard to come up with situations where truth is not definable but remains eliminable in some other way, such as being conservative over an underlying theory. In such a situation truth could be shown not to contribute anything to our knowledge outside semantics. Truth would, so to speak, supervene on the underlying base theory without contributing anything to it and truth would be in this sense eliminable without being definable. So perhaps a definition of truth is dispensable even if one aims at an eliminative theory of truth.

Also, one need not provide model-theoretic semantics for analysing various properties of these principles. In some cases one will be able to prove their consistency and many other properties without appealing to model-theoretic semantics. In particular, one will be able to see what commitments are tied to these principles. If one is using a defined notion of truth it can be difficult to see which properties flow from the postulated principles of truth and which come from the particular chosen definition of truth.

Furthermore, I would like not to exclude the situation where truth is added to one's overall mathematical theory where truth is not definable. Of course, there is an opposed reductive view according to which a notion is only acceptable if it can be defined in set theory. At least I would like to consider alternatives to set-theoretic reductionism in which truth is not and cannot be defined away.

After this plea for the axiomatic approach I assure the reader that I will also use model-theoretic methods throughout the book. As I mentioned above, the axiomatic approach is not opposed to definitional approaches. In fact, both approaches complement one another.

For instance, one may start by formulating truth-theoretic principles like Tarski's T-sentences for a theory like Peano arithmetic and then show that a suitable truth predicate for the language of arithmetic can be defined in set theory by defining a model for Peano arithmetic expanded by these truth-theoretic principles. Hence one knows that these truth-theoretic principles are consistent, at least if set theory is to be trusted. Then one might try to formulate analogous principles for the language of Zermelo–Fraenkel set theory. There is no hope to *define* this truth predicate for set theory, but the

fact that other theories such as Peano arithmetic possess a nice model when expanded by principles of this ilk supports the view that the expansion of set theory by the corresponding principles is consistent as well although Tarski's theorem rules that this cannot be shown, unless one goes beyond Zermelo–Fraenkel set theory by introducing class quantifiers or other devices. At any rate, the model-theoretic constructions can illuminate, motivate, and to some extent support axiomatic truth theories.

Model-theoretic approaches are also important for the proof theory of axiomatic truth theories. In many cases I will use formalizations of model-theoretic constructions to provide proof-theoretic analyses of axiomatic theories of truth. So axiomatic and semantic approaches complement one another also on the technical side.

2 Objects of truth

The axioms for truth will be added to what is called the base theory. In the main part of this book I will use Peano arithmetic as the base theory, but applications to other more comprehensive base theories are intended, and the base theory may contain empirical or mathematical or still other axioms together with the appropriate vocabulary. At any rate, a base theory must contain at least a theory about the objects to which truth can be ascribed.

Truth theories have been proposed where the need for objects to which truth can be ascribed and for a theory of these objects seems to be avoided. If truth were analysed in terms of special quantifiers as in the so-called prosentential theory of truth by Grover et al. (1975), for instance, it might initially appear that such objects are avoided, but it is not at all clear that the new quantifiers avoid any ontological commitment.

I have no ambition to avoid ontological commitment to objects that can be true. If the axiomatic theories of truth I am going to discuss are intertranslatable with an approach without such ontological commitment, so be it. If such a translation is not possible, then I suspect that something is wrong with the approach. Here I will stick to the usual approach that takes truth to be a predicate.

In almost all cases, the axioms for truth can only serve their purpose when combined with a suitable base theory. If the truth axioms, on which the theories in this book are based, are separated from the base theory, the result is a very weak theory. Since I have not introduced the theories or their axioms yet, I can only sketch some trivial results. But one might think about the theories with the disquotation sentences as axioms or with compositional axioms for truth like the one stating that a conjunction is true if and only if both conjuncts are true.

Pure theories of truth without a base theory will not prove that there are more than two objects, so they will be weak in this sense. For the purely truth-theoretic axioms do not allow one to distinguish between different true objects, or different untrue objects. Therefore one can usually obtain a model for a truth theory without a base theory by taking some model of the full theory and identifying all objects in the extension of the truth predicate, on

the one hand, and all objects not in its extension, on the other hand. The truth predicate and its axioms will only show their potential when combined with a suitable base theory.

Of course, one could strengthen the truth-theoretic axioms by building certain ontological claims and axioms about the structure of truth bearers into them. But that seems difficult if one is sticking to fairly natural axioms.

However, I do not want to suggest that truth axioms do not bring *any* ontological commitment, as some deflationists might hope. The theory of truth cannot be completely separated from the underlying ontology of objects that can be true, as even very weak axioms for truth will imply that there are at least two different objects, if the axioms imply that something is true and something not true, as even very axioms do. This will be shown on p. 55 below.

If a theory about the objects to which truth can be ascribed is required as base theory, the question arises what these objects are and objects of which kind should be described in the base theory.

Philosophers have ascribed truth to propositions conceived as objects that are independent from language, or to sentence types or sentence tokens. Of course, there are also thoughts, beliefs, contents of sentences, and so on, where it is not obvious how these relate to propositions and sentences. Even once one has settled on propositions or sentences many decisions remain.

Since I am not aiming at a complete (recursive) axiomatization of the ontology of the objects to which truth is ascribed – which would be impossible under even very weak assumptions – I avoid questions concerning their nature – at least to a certain extent. I do, however, presuppose that the objects possess an ontological structure analogous to that of sentences, more specifically of the sentences of the language of the base theory in the case of typed theories of truth, and of the language of the base theory expanded by the truth predicate in the case of type-free theories.

In order to state an axiom to the effect that a conjunction is true if and only if both conjuncts are true, one needs to assume that the operation of conjunction is defined on the objects to which truth is ascribed. Similarly, if one wants to say that a universally quantified sentence is true if all instances are true, the operations of universal quantification and of instantiation must be defined. The axioms of the truth theory can serve their purpose only if the base theory allows one to express certain facts about the syntactic operations; otherwise they may remain void. For instance, the axiom that a conjunction is true if and only if both conjuncts are true becomes void if no information is supplied on what a conjunction is. For instance it should be provable in