11

Completeness and Soundness

The reduction principle is the linchpin of the model-theoretic account of consequence. If some version of it were correct, the account would certainly deserve the esteem in which it is held. But we have seen that this principle is irremediably flawed, and that, as a result, we have no guarantee that an application of the account to any particular language will be extensionally correct. The definition can both overgenerate and undergenerate, declaring arguments logically valid that in fact are not, and declaring them not when they actually are.

In the case of first-order languages, the completeness and soundness theorems have been taken as providing a proof that a particular deductive calculus correctly characterizes the logical consequence relation for these languages. The soundness theorem is traditionally viewed as showing that the calculus does not overgenerate, that whenever a sentence S is derivable from a set K of assumptions, S does indeed follow logically from K. Conversely, the completeness theorem is thought to show that the calculus does not undergenerate, that if S follows logically from K, then there is a proof of this fact within the calculus. The goal of this chapter is to reconcile the morals of the preceding chapters with the intuitions at work here.

It is clear that the traditional construal of completeness and soundness can no longer be maintained. If the model-theoretic analysis can overgenerate, a soundness theorem by itself does not guarantee the soundness of the deductive calculus in question. Just so, if the model theory can undergenerate, a completeness theorem does not, by itself, guarantee the completeness of the deductive system. The usual interpretation of these theorems clearly presupposes the extensional adequacy of the model-theoretic account of consequence, a presumption that is simply unjustified.

We have seen that the model-theoretic account will get things right in one set of circumstances. If, first of all, none of the substantive generalizations on which the output of the account depends turns out to be true, then the definition will not overgenerate. Second, if none of the valid arguments expressible in the language depend for their validity on expressions whose interpretations we vary, then the definition will not undergenerate, either. However, a second's thought shows that the relationship between these principles and the soundness and completeness theorems is far from straightforward.

What work are the soundness and completeness theorems doing? Do they in fact guarantee anything at all about the intuitive notions of logical truth and logical consequence? To answer these questions I will make a slight detour. I am not the first person to raise questions about the significance of the completeness theorem for first-order logic. In a well-known article entitled "Informal Rigour and Completeness Proofs," Kreisel distinguishes between what he calls "intuitive" validity and the model-theoretic notion of truth in all set-theoretic structures. This distinction leads Kreisel to an alternative view of the significance of the completeness theorem. Although Kreisel's starting point is incorrect, for reasons that will become clear, his strategy is one we will find useful in our own reconciliation.

Kreisel's Observation

The main thrust of Kreisel's article is to emphasize that we can prove rigorous results about informal notions, a contention with which I wholeheartedly agree. As a case study, he considers the intuitive notions of logical validity (what I have been calling "logical truth") and logical consequence. Kreisel's aim is to show that, in the case of first-order logic, we can rigorously establish that the intuitive notion of validity, which he abbreviates as *Val*, is extensionally equivalent to the set-theoretic definition standardly given.

The definition that Kreisel has in mind (which he denotes by V) is that a sentence has property V just in case it is true in all models (or structures, as Kreisel prefers to call them), *where the domain of quantification is a set in the cumulative hierarchy*. Kreisel's worry is that this does not correspond exactly to the notion *Val*. As he expresses the problem:

> The intuitive meaning of *Val* differs from that of V in one particular: $V(\alpha)$ (merely) asserts that α is true in all structures in the cumulative hierarchy, . . . while $Val(\alpha)$ asserts that α is true in *all* structures. (1969, p. 90)

To drive home the difference between these two notions, Kreisel considers a sentence α in the language of set theory. Intuitively, it

seems that if $Val(\alpha)$ then α must be true as a statement *about* the cumulative hierarchy—that is, where the domain of quantification is the collection of all sets. But $V(\alpha)$ assures us only that it must be true *in* all set-theoretic structures. But the cumulative hierarchy itself is "too big" to be among these structures. Because of this, Kreisel rightly points out, it is not at all clear that sentences true in all *set-theoretic* structures will be true in *all* structures.

Kreisel's main point is that, insofar as V takes into consideration fewer structures than *Val*, it cannot be a trivial matter to go from $V(\alpha)$ to $Val(\alpha)$. His use of a set-theoretic example introduces an additional point, though. For in this case, it turns out that the *intended* interpretation of the language is among the structures canvassed by *Val*, but not among those canvassed by V. Of course, as I have emphasized several times, one of Tarski's key requirements is that the intended interpretation of an expression be in the satisfaction domain associated with that expression, otherwise we risk declaring sentences logically true that in fact are false. This requirement is violated when our language is the language of set theory and our model-theory uses only structures from within the cumulative hierarchy.

Kreisel claims that in spite of these seeming difficulties, the completeness theorem allows us to establish that V has the same extension as the intuitive notion of logical validity. His reasoning is as follows. He first notes that, since the standard deductive rules of first-order logic are intuitively sound, we know that if a first-order sentence α is derivable, $D(\alpha)$, then it is logically valid.[1] That is,

(1) $\forall\alpha(D(\alpha) \rightarrow Val(\alpha))$.

Second, since it is obvious that all set-theoretic structures are structures, truth in all structures (*Val*) implies truth in all set-theoretic structures (V). That is,

(2) $\forall\alpha(Val(\alpha) \rightarrow V(\alpha))$.

However, the completeness theorem for first-order logic tells us that any sentence true in all set-theoretic structures is derivable. That is,

(3) $\forall\alpha(V(\alpha) \rightarrow D(\alpha))$.

Putting these three together, we see that *Val*, V, and D are, for first-order languages, extensionally equivalent:

$$\forall\alpha(Val(\alpha) \leftrightarrow V(\alpha) \leftrightarrow D(\alpha)).$$

For our purposes, there is a serious flaw in Kreisel's argument. The problem has to do with the interpretation of *Val*. If *Val* simply *means* truth in all structures, then the argument is correct, though its moral is not exactly what Kreisel implies. But if *Val* really is the intuitive notion of logical validity (or logical truth), then step (2) is quite dubious. The problem is that Kreisel simply identifies, without argument, the intuitive notion with the model-theoretic notion of truth in all structures. Needless to say, this is precisely the identification against which I have been arguing.

Let us reexamine the two steps of Kreisel's argument that involve *Val*. But to avoid the above conflation, I will reserve *Val* for the notion of truth in all structures, and introduce *LTr* for the intuitive notion of logical truth or validity. With this disambiguation, step (1) splits into two possible claims, namely:

(1) $\forall\alpha(D(\alpha) \rightarrow Val(\alpha))$

(1′) $\forall\alpha(D(\alpha) \rightarrow LTr(\alpha))$.

It turns out that both of these are legitimate, though they require slightly different justifications. (1′) holds simply because the deductive system is intuitively sound—that is, it allows us to derive only logically true sentences. To recognize the truth of (1), however, we need to observe that the validity of the rules of our deductive system holds in all of the interpretations canvassed by *Val*. In the case of a standard first-order system of deduction, both of these follow by a routine examination of the rules on a case-by-case basis. But notice that (1) is more sensitive than (1′) to the details of the deductive system in question. For example, if our deductive system included the ω-rule, or a rule allowing us to conclude 'Man(x)' from 'Bachelor(x),' then (1) would certainly fail, even though (1′) might not.

How about step (2)? Here we have the following split:

(2) $\forall\alpha(Val(\alpha) \rightarrow V(\alpha))$

(2′) $\forall\alpha(LTr(\alpha) \rightarrow V(\alpha))$.

Clearly, (2) follows trivially from the fact that every model in the cumulative hierarchy is a model, the same reason we gave before. But (2′) is quite another matter: it is simply the bald assertion that logical truths are true in every model in the cumulative hierarchy. But in fact we do not know that the logical truths of any given first-order language will be a subset of either V or *Val*. To suppose that they are is just to suppose that the model-theoretic account (whether V or *Val*) does not undergenerate. If there is an argument for this contention, it must be something quite specific to the first-order language in question, since we have seen that it does not hold in general.

Kreisel's argument goes through for the notion *Val*. What it shows is that, in the first-order case, truth in *all* structures is equivalent to truth

in a restricted collection of structures. To drive this point home, let us generalize the argument a bit. Let $\mathcal{M}$ be any class of models and write $Val_{\mathcal{M}}$ for truth in all structures in $\mathcal{M}$. Thus V is the special case of $Val_{\mathcal{M}}$ where $\mathcal{M}$ consists of all structures whose domain is a set in the cumulative hierarchy.

What do we need in order for the argument given above to generalize? We need to know that the class $\mathcal{M}$ is rich enough to serve as a basis for the proof of the completeness theorem. More fully, let us call a class *rich* if it satisfies the condition that any first-order sentence S which is true in all models in $\mathcal{M}$ is derivable. For example, the collection of countable structures is rich. We can clearly replace V by $Val_{\mathcal{M}}$ throughout Kreisel's proof, and the argument goes through so long as $\mathcal{M}$ is rich, since then we have $Val(\alpha) \rightarrow Val_{\mathcal{M}}(\alpha) \rightarrow D(\alpha) \rightarrow Val(\alpha)$. Thus, one moral of Kreisel's argument is that when we apply the model-theoretic account to first-order languages, it does not matter whether we use all structures, all set-theoretic structures, or all countable structures. These will all have precisely the same extension.

What Kreisel's argument does not show, however, is that this extension coincides with the set of logical truths of any given first-order language—say, the logical truths of the language of elementary arithmetic. For, in spite of the argument, it is far from clear that $Val_{\mathcal{M}}$ (or *Val* itself) does not undergenerate. To see this, we need only recall that rich classes are in general forced to contain unintended models for our first-order language. Thus, for example, if our language is the language of first-order arithmetic, then a rich class will perforce contain nonstandard models of arithmetic. But what guarantee do we have that an intuitive logical truth in the language of arithmetic will be true in these nonstandard models, or that an intuitively valid argument will preserve truth in them? For example, if some version of the ω-rule is logically valid, as Tarski argued, then there will indeed be logical truths which are not true in all models in $\mathcal{M}$, let alone true in *all* models.

The Problem of Overgeneration

Recall that the goal of this chapter is to determine the significance of completeness and soundness theorems for the intuitive notions of logical truth and logical consequence. In particular, what bearing do they have on the question of whether a given application of the model-theoretic account either overgenerates or undergenerates?

If Kreisel's argument were correct when construed as an argument about *LTr*, then it would settle both of these questions; as it is, though, it does not directly address either. All it shows us is that the three notions *Val*, *V*, and *D*, are coextensive. This is a significant observation,

to be sure, but not the one we are after. It tells us nothing about how the intuitive notions of logical truth and logical consequence relate to their model-theoretic (or proof-theoretic) counterparts.

Still, it does suggest a partial solution. Indeed, as the reader may already have noticed, Kreisel's argument can be combined with (1′) to settle the overgeneration question, at least in the first-order case. Recall that (1′) is the observation that the deductive system used in the proof of completeness is intuitively sound, that only genuine logical truths are derivable in the system.

(1′) $\forall\alpha(D(\alpha) \rightarrow LTr(\alpha))$.

This observation holds for any first-order language, whether the language of elementary arithmetic, the language of set theory, or the simple language of Chapter 5. But Kreisel has shown, using completeness, that any first-order sentence that is true in all models is derivable:

$$\forall\alpha(Val(\alpha) \rightarrow D(\alpha)).$$

Combining these two, we get the result we need. In the case of first-order languages the model-theoretic account does not overgenerate:

$$\forall\alpha(Val(\alpha) \rightarrow LTr(\alpha)).$$

How does this bear on our observation that the model-theoretic account overgenerates only when some of the substantive generalizations associated with sentences of the language turn out to be true? The relationship is roughly this.[2] Suppose we have a first-order sentence

(4) $S(\vec{P}, \vec{c})$

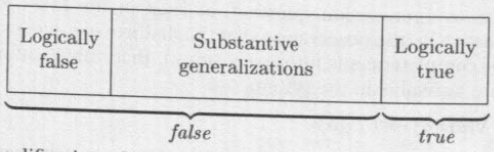
where the displayed $\vec{P}$ and $\vec{c}$ are the only constituent expressions not in the set $\widetilde{\mathcal{S}}$ of fixed terms. Sentence (4) will be declared logically true by the model-theory only when the following closure is true:

(5) $\forall\vec{X}\forall\vec{x}[\, S(\vec{X}, \vec{x})\,]$

Now the completeness theorem tells us that whenever we have such a true closure, the original sentence (4) is provable. By the fact that our deductive system is intuitively sound, this is enough to guarantee that (4) is a genuine logical truth. This is all we need for the above argument to go through.

But note that there is something more that we can recognize. Since (4) is provable in our system without any assumptions, we can also prove (5) in the same system (or in a minimal second-order extension of it, if $\vec{P}$ is not degenerate), by generalizing on the parameters $\vec{P}$ and $\vec{c}$. In other words, the completeness theorem (plus the recognizable

soundness of our deductive system) guarantees that *if* (5) is true, then it is itself a *logical* truth. This is how we can establish that all of the substantive generalizations that the model-theory associates with sentences of the language are indeed false. This is how we can show that our application is of the "fortuitous" sort:



Our modification of Kreisel's argument obviously generalizes, yielding a useful strategy for showing that a particular model-theoretic account does not overgenerate. The strategy is simple to state, though not always possible to implement. Find a set of derivation rules for the language in question that, first of all, are intuitively valid and, second, are provably complete with respect to the model-theoretic account. When this can be done, we are assured that the model theory does not wrongly declare sentences logically true or arguments logically valid. The recognizable soundness of the deductive calculus transfers over, via the completeness theorem, to the semantic account.

Of course, we know the strategy cannot always succeed, because the model-theoretic account does sometimes overgenerate. For example, I argued in Chapter 9 that there are second-order sentences which are not logical truths but which are declared such by the model-theoretic account. If so, then it follows that there is no sound deductive system (effective or not!) that is complete with respect to the standard, second-order model theory. This is partially substantiated by the well-known result that no such *effective* system exists, a consequence of Gödel's incompleteness results.

In Chapter 9, we saw that there is no "internal" guarantee that an application of the model-theoretic account will not overgenerate. Even when it does not, there is no way to recognize this fact from the analysis itself, from characteristics of the language, or from the expressions held fixed. What we can now see, though, is that an external guarantee can sometimes be found, a guarantee derived from the presumed soundness of our deductive calculus, in tandem with a completeness theorem showing that the semantic account reaches no further than the syntactic.

The Problem of Undergeneration

The reason the completeness theorem is so called is that it purports to establish that a given deductive calculus does not undergenerate, that

it is "complete." We have used the theorem, in contrast, to show that an application of the *model-theoretic* account of consequence does not *overgenerate*, in effect to show that our semantic account is, in the first-order case, *sound*. Is there any way to prove the converse, to show that our first-order model theory (or an extensionally equivalent deductive system) does not undergenerate?

The most straightforward answer, unfortunately, is no. If our aim is to characterize the set of logical truths (or the logical consequence relation) for an antecedently given first-order language, then there is no general way, short of fixing all of the expressions in the language, to guarantee that the model theory captures them all. Indeed, once we focus on any interesting first-order language, such as the language of elementary arithmetic, it seems clear that the standard model theory does undergenerate. It is only our uncritical adoption of the model-theoretic analysis that has obscured this simple point.

Still, it is possible to extract from the model-theoretic account some substantive observations about the intuitive notions of logical truth and logical consequence. The trick is to shift attention from the logical properties of any particular language to the logical properties common to a range of languages.

In Chapter 7, we noted that Tarski's unmodified definition of logical truth (that is, prior to the use of cross-term restrictions) provides a necessary condition for what we there called the *relativized* concept of logical truth. That is, if a sentence S is logically true, and if this fact depends only on the meanings of some subset $\mathfrak{F}$ of its constituent expressions, then S will remain true however we reinterpret the other expressions (so long as our reinterpretations do not change the semantic categories of those expressions). We took this as showing that Tarski's original definition would never undergenerate with respect to the notion of logical truth *relativized* to the set $\mathfrak{F}$ of fixed terms.

We can reconstrue this as an observation about the logical truths common to a collection of languages, those languages canvassed by the model theory. In the case of Tarski's original account, these are the languages that arise when our models provide (semantically well-behaved) interpretations of the expressions not in $\mathfrak{F}$. Construed this way, the observation is that the set of sentences that are logically true in *every* such language will be a subset of the set of sentences declared logically true by the model theory.

It turns out that when we cast our observation in this form, it becomes completely independent of any details of Tarski's account. Indeed, suppose we have any collection $\mathfrak{L} = \{L_M\}_{M \in \mathcal{M}}$ of languages that share the same set of sentences, but differ in how these sentences are interpreted. Note first of all that for any language L_M in this collection,

the logical truths of L_M must clearly be a subset of the truths of L_M. Modifying our earlier notation in an obvious way, we can express this as follows:

(6) $\quad LTr(L_M) \subseteq Tr(L_M).$

It follows from this simple fact that the logical truths *common* to the languages in $\mathcal{L}$ must be a subset of the common *truths* of the languages. That is:

(7) $\quad \bigcap_{M \in \mathcal{M}} LTr(L_M) \subseteq \bigcap_{M \in \mathcal{M}} Tr(L_M).$

Or, equivalently:

$$\bigcap_{M \in \mathcal{M}} LTr(L_M) \subseteq Val_M.$$

Now, the model-theoretic account takes the set appearing on the right-hand side of (7) to be the set of logical truths of each and every language in $\mathcal{L}$. While this simple identification is mistaken, what (7) shows us is that, at least as an account of the *common* logical truths of the canvassed languages, the account will not undergenerate. And unlike our earlier observation, this observation holds even if our semantics employs cross-term restrictions. Interestingly, it is entirely independent of how the model theory specifies the collection $\mathcal{L}$ of languages: it does not even matter if expressions retain the same semantic categories as we move from interpretation to interpretation.

This puts us in a position to give a Kreisel-like argument showing that, in the first-order case, we can characterize exactly the set of logical truths common to all languages of the form L_M, where M ranges over any rich collection $\mathcal{M}$ of models. Let us write $CLTr_{\mathcal{M}}(\alpha)$ to indicate that α is one of these common logical truths—that is,

$$CLTr_{\mathcal{M}}(\alpha) \leftrightarrow \alpha \in \bigcap_{M \in \mathcal{M}} LTr(L_M).$$

Our argument will use the following three steps.

(1″) $\quad \forall\alpha(D(\alpha) \rightarrow CLTr_{\mathcal{M}}(\alpha))$

(2″) $\quad \forall\alpha(CLTr_{\mathcal{M}}(\alpha) \rightarrow Val_{\mathcal{M}}(\alpha))$

(3″) $\quad \forall\alpha(Val_{\mathcal{M}}(\alpha) \rightarrow D(\alpha)).$

All of these observations have, in fact, been made earlier in the chapter. Step (1″) is simply the observation that our deductive system is sound, independent of which first-order language L_M is under consideration. Thus, if α is derivable in the system, it must be a common

logical truth of these languages. Step (2″), on the other hand, is a restatement of (7), which holds for absolutely any collection of languages. Finally, step (3″) is a statement of the completeness theorem, and follows from our assumption that $\mathcal{M}$ is a rich collection of models. Combining these gives us a result analogous to Kreisel's:

(8) $\quad \forall\alpha(CLTr_{\mathcal{M}}(\alpha) \leftrightarrow Val_{\mathcal{M}}(\alpha) \leftrightarrow D(\alpha)).$

It is not entirely clear how significant this result really is, for all its elegance. If our concern is to explicate the logical properties of a specific first-order language, then (8) is of limited interest. Indeed, it seems likely that the most significant logical truths and logically valid arguments of a given language will be filtered out by shifting attention to that portion of its logic common to a rich collection of languages. From this perspective, we have done little more than redefine the notions under investigation, and in such a way that the resulting task has been stripped of many of the intuitions that motivated the pioneers of modern logic, intuitions clearly at work in Tarski's original attempt to characterize the consequence relation.

On the other hand, there is a different project in the context of which (8) is of considerable interest. It would be misleading to think of model theory as motivated solely by the goal of analyzing logical properties and relations. A large part of its motivation can be understood only in relation to modern algebra. Indeed, a central concern of the discipline from Tarski and Robinson on has been the systematic understanding of notions and techniques of abstract algebra.

One of the most striking features of modern algebra is the technique of simultaneously studying a wide collection of mathematical structures, as when we investigate the properties of abelian groups. A key insight was that one and the same proof can often be interpreted as applying to all structures in the specified collection. By isolating the common truths on which such a proof depends, we can obtain results of striking generality. As a result, the practice in algebra is to group structures together by means of a set of core truths called "axioms," and to construct proofs that rely solely on the core truths together with the logical properties common to any interpretation of these truths.

From this perspective, a key concern is exactly the logical properties common to a collection of interpreted languages, and so (8) acquires added significance. It assures us that so long as our collection of algebraic structures can be characterized by first-order axioms,[3] the consequence relation simultaneously captured by the model theory and proof theory coincides with the specialized notion of consequence used by the algebraist when reasoning about a range of structures. This positive result is in striking contrast to the case where the collec-

tion of structures in question cannot be characterized using first-order axioms, as in the case of torsion groups, archimedean fields, or finite division rings. In these cases, the notion of consequence used by the algebraist clearly outstrips that captured by the notions related in (8).

Recapitulation

In previous chapters, we saw that the model-theoretic account of logical truth and logical consequence will regularly and predictably go astray: some applications overgenerate, others undergenerate, and in some cases it fails both ways at once. In this chapter, I have tried to reconcile these general observations with the intuition that, at least in the first-order case, the analysis gets something right, and that the completeness and soundness theorems play an important role in demonstrating that fact.

By modifying an argument of Kreisel's, we saw that for first-order languages the model-theoretic account does not overgenerate: no argument declared valid by the model theory will be invalid. The crucial observation is that the completeness theorem allows us to transfer the intuitive soundness of the deductive system over to the model theory. The theorem assures us that any model-theoretically valid argument is provable in the deductive system, and so is genuinely valid if this system is sound. Note that, somewhat ironically, the real guarantee of validity is carried by the presumed soundness of the deductive calculus, and not by the declarations of the model theory itself.

Reassuring as this is, it is also a bit unsatisfying. After all, one thing we might hope for in a semantic account of consequence is an explanation of why valid arguments are valid, an explanation not given to us by syntactic characterizations of this notion. But we now see that even in cases where we can demonstrate that the model theory does not overgenerate, our proof hinges on the presumed soundness of the syntactic characterization.

Still, a proof is a proof, and we are better off in this case than we are in the absence of a completeness theorem. In those cases where the model theory outstrips the deductive calculus, we have no general way of determining whether it is because the model theory overgenerates or the deductive calculus undergenerates. Indeed, with second-order logic, we seem to have both. The model theory declares the continuum hypothesis (or perhaps its negation) to be a logical consequence of the pair-set axiom—hardly a plausible assessment. On the other hand, any effective deductive calculus for the language will, if sound, fall short of the intuitive consequence relation for the language.[4] Here, the genuine consequence relation must fall somewhere in between the deductive and model-theoretic accounts.

When we turn from the problem of overgeneration to the problem of undergeneration, the situation is even less satisfactory. If we maintain our original interest in the notion of consequence for a fixed language, the model-theoretic account does undergenerate for all but the most trivial languages, and so of course there is no way to show that it does not. The only way to get around this is, in effect, to define away the problem, by shifting attention to the notion of the logical truths and logically valid arguments *common* to the range of languages canvassed by the model theory. Then, although the model theory can still overgenerate, it is guaranteed not to undergenerate for very simple and straightforward reasons.

In cases where we have a completeness theorem, this trick allows us to view our model theory (and our deductive calculus) as both sound and complete, relative to this alternative notion of "common logical validity." The thing to remember here is that the role of the completeness theorem is to show the soundness of the model theory relative to this new notion. The "completeness" of the model theory is simply built into the definition of the alternative notion. Still, this gives us a construal of the first-order completeness theorem that sheds some light on the notion of consequence that is of interest to practitioners of modern algebra.