an explicit definition *in the special case where there is a set of all things x such that U(x)*, since in that special case the interpretation specifies a model as defined above.

Another way to paper over the problem is to say that it doesn't arise for "class models", but this brings us to a second linguistic vacillation, between two ways of understanding talk of classes.

First, there are theories which postulate, in addition to sets, additional "set-like entities" called proper classes. (In the usual formulations, sets are a special kind of class: the classes are partitioned into the sets and the proper classes.) According to such theories, although there is no set of all sets, there is still a proper class of all sets; but there is no class of all classes. In such theories we can allow models to be proper classes, and if we do so then there will be a model that contains all sets. Still, no model even in this extended sense can contain all *classes*, and hence we won't have essentially changed the predicament: truth in all models has no obvious bearing on genuine truth.

Second, there is also a standard practice of using the term 'class' informally in theories that don't literally posit proper classes, but only posit sets; talk of classes is viewed as just a dispensable manner of speaking. And in this manner of speaking, it is common to speak of these classes as the domains of "class models". (This is for instance what set theorists usually have in mind when they speak of "class models" of Zermelo–Fraenkel set theory.) But 'class' in this informal sense is just another term for 'interpretation' as understood several paragraphs back; and truth in a "class model" is not explicitly definable in such theories.

(It would be of no help to propose that we define validity not within set theory or class theory but within second order logic, viewing that as a basic logic not explainable in (and more powerful than) the first order theory of sets and classes. Putting aside doubts about the intelligibility of second order logic so conceived, let's grant that it could be used to explain *first order validity*. But the question of interest is defining validity *for the language being employed*, which on the current proposal isn't first order logic but second order logic. And I think it's pretty clear that similar difficulties to those in the first order case would arise for any explicit second order definition of second order validity.)

So however we twist and turn we have a problem to deal with: if validity is defined in a standard way, the claim that a sentence is logically valid doesn't entail that it's true, and the claim that an argument is logically valid doesn't entail that if the premises are true then so is the conclusion. The problem was first adequately addressed by Kreisel (1967).

2.3. THE KREISEL SQUEEZING ARGUMENT

Kreisel's solution to the puzzle introduced in the previous section was essentially this:

1. We should think of the intuitive notion of validity not as literally *defined* by the model theoretic account, or in any other manner; rather, we should think of it as a primitive notion.
2. We then use intuitive principles about validity, together with technical results from model theory, to argue that validity *extensionally coincides with* the technical notion.

This is somewhat similar in spirit to the approach of the Validity Argument in Section 2.1, but considerably different in detail since the idea is to sustain the extensional correctness of the model-theoretic definition rather than of a definition along the lines of (VAL).

To elaborate Kreisel's approach, consider some specific formalization F of (first order) quantification logic (with identity) that employs axioms and rules that seem self-evident. Call an argument *derivable in* F if its conclusion can be derived from its premises together with the axioms of F, using only the rules of F. Then the idea is that intuitive principles about validity guarantee

(Intuitive soundness of F) Any argument that is derivable in F is $\text{valid}_{\text{intuitive}}$.

Intuitive principles about validity also seem to guarantee

If there is a model in which the premises of an argument are true and the conclusion isn't, then the argument is not $\text{valid}_{\text{intuitive}}$;

or to put this another way,

(Model assumption) Any argument that is not $\text{valid}_{\text{technical}}$ is not $\text{valid}_{\text{intuitive}}$;

So the picture is as shown in Diagram 2.1.

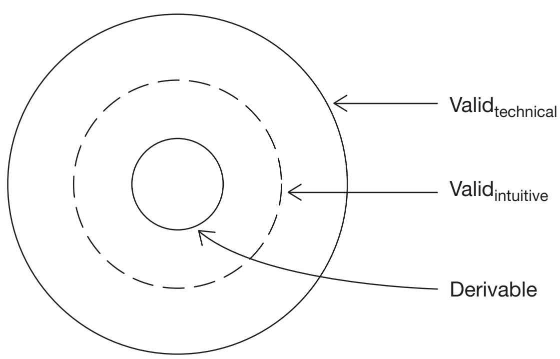

Diagram 2.1.

But now we appeal to the fact that (if the formalization F is properly chosen) we will have a completeness theorem, which says:

(Completeness Theorem for F) Any argument that is $\text{valid}_{\text{technical}}$ is derivable in F.

And this together with the previous two guarantees that intuitive validity extensionally coincides both with validity in the technical model-theoretic sense and with derivability in F: the outer circle and inner circle are squeezed together, forcing the middle one to line up with both of them.

This analysis (an example of what Kreisel called "informal rigour") seems to me to give quite a convincing reason for accepting the standard model-theoretic "definition" of validity as extensionally adequate for classical (first order) logic, even for unrestricted languages.

An interesting feature of the analysis is that the standard "Soundness Theorem" for F, the converse of the Completeness Theorem, plays no explicit role. (Of course, the Intuitive Soundness Principle and the Model Principle together imply the conclusion of the Soundness Theorem.) This is connected to the issues to be discussed in the following section.

2.4. THE UNPROVABILITY OF SOUNDNESS

In standard metalogic courses we teach a theorem that is billed as showing the soundness of some typical proof procedure for standard predicate logic. This is a bit of a hoax. What this "soundness theorem" actually says is this:

(M) for any set of sentences Γ and any sentence B, if B is derivable from Γ in the proof procedure then B is true in every model in which all members of Γ are true.

This is "argument soundness". As a corollary, we get the weaker claim of "statement soundness":

(M_w) for any sentence B, if B is derivable (without assumptions) in the proof procedure then B is true in every model.

But I hope it's clear from the discussion in Section 2.2 that (M) and (M_w) aren't really sufficient for soundness in the intuitive sense. For instance, (M_w) doesn't entail that sentences derivable in this proof procedure are actually true. To what extent can this lacuna be filled?

There are, actually, two quite distinct worries here, only one of which is my immediate concern.

The worry that is *not* my immediate concern (but which is worth mentioning so as to clearly distinguish it from the one that is) is that any argument for the genuine soundness of a logic is obviously going to have to assume that very logic. (I'll call this *the radical worry*, for want of a better term.) For instance, if we are to argue that all instances of excluded middle are true, the obvious way to do it is to say:

(i) if a sentence x is true, then the disjunction of it with anything else is true; so in particular, the disjunction of x with the negation of x is true;