1 Context

As far as academic disciplines go, *logic* is strange. In the western academy, its roots go back to Aristotle, to Euclid, to the Stoics, through medievals, the Arabic world, and into a flowering complexity in the nineteenth and twentieth centuries, as philosophers and mathematicians grappled with understanding the power and limits of deductive reasoning. The field we now know as modern logic took root in the project of systematising and securing the foundations of mathematics[1] and in giving an account of the relationship between those mathematical theories and our experience of the world around us. In the twentieth century, new connections emerged with the nascent fields of linguistics, digital systems and computer science. There is no way that an Element on the use of *proofs* and *models* in *philosophical* logic could do justice to anything more than a tiny fragment of this sprawling edifice.[2]

So what small fragment of the discipline of logic will this Element address? As the title states, our focus is *philosophical logic* and the twin roles of *proofs* and *models* in the development of logic. The *philosophical* concern will also be twofold: we will reflect on the application of logic to some questions in philosophy and, at the same time, consider a philosophical reflection on the discipline of logic itself. Philosophical logic provides both a set of sensibilities and processes and tools for application in philosophical discourse (among other kinds of discourse), *and* at same time, it is a site of philosophical reflection. We will maintain these dual perspectives on our topic throughout this Element.

In this first section, I set the scene by way of an introduction to how we can use the different tools of *proofs* and *models* to form judgements about logical validity and invalidity. Then I outline how attention to proofs and models plays a role in some of the current debates in philosophical logic, concerning the *semantic paradoxes* and *vagueness*. I then end the section by looking ahead to the argument of the remainder of the Element.

1.1 Proofs and Models

There are many ways to look at *logic* and the constellation of concepts that logicians have attempted to analyse using proofs and models. One way to do

[1] J. Alberto Coffa's *The Semantic Tradition from Kant to Carnap* (1993) tells the compelling story of the growth of modern logic beyond its Aristotelian bounds as Bolzano, Weierstrass and others attempted to make sense of the mathematically important notions of convergence and continuity.

[2] So I will not cover the rich tradition of proof complexity, Gentzen's consistency proof for arithmetic, the connections between proof search and decidability and many more interesting topics in proof theory. Neither will I discuss a great deal of model theory, such as significant metatheoretical results including compactness, cardinality of models or ultrapowers, and other model construction techniques. This Element is only *so* long.

this is to focus on the production and the evaluation of argumentation, reasoning and inference. Let's start with a simple example, involving two mathematicians, who are reasoning about some newfangled binary relation R they are exploring. They have discovered that the relation R is *transitive* (i.e., for any objects x, y and z if R relates x to y and R relates y to z, then R relates x to z too, so *is older than* is an example of a transitive relation, while *is a parent of* is not) and *symmetric* (if R relates x to y, then R relates y back to x too, so neither *is a parent of* nor *is older than* are symmetric relations but *is a sibling of* is), and finally, the relation is *directed* (for each object x, there is some object y where R relates x to y, so, on the collection of non-negative natural numbers (*natural numbers* for short) $0, 1, 2, \ldots$, the relation *is smaller than* what is directed, since for every number x we can find some whole number y where x is smaller than y, but the relation *is larger than* what is not directed if we restrict our attention to the natural numbers. There is no natural number y where 0 is larger than y).

So our two mathematicians agree that this newfangled relation R is transitive, symmetric and directed. One of our mathematicians exclaims: 'The relation R is reflexive, too!' (A relation R is *reflexive* if, for every object x, R relates that object to *itself*). Our second mathematician asks: why is that? The first responds:

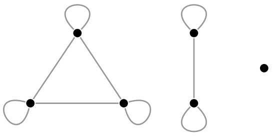
(1) R is transitive. It's symmetric. It's directed. It must be reflexive, too!

The second mathematician doesn't see why this is the case, so they ask for the reasoning to be spelt out. Can the leap from the premises to the conclusion be broken down into smaller steps? It can. Our quick thinker responds:

(2) Take an object a. Since R is directed, there is some object b where R relates a to b. Since R is symmetric, R also relates b to a. Now, R relates a to b and R relates b to a, so since R is transitive, R relates a to a. So we have just shown that for any object a at all, R relates a to itself. That is, R is reflexive.

This elaboration is one way to spell out the jump from the premises to the conclusion. It is what we call a *proof*. It fills out that large jump in thought in terms of smaller steps. In this case, the smaller steps involve the applications of agreed-upon definitions (unpacking the definitions of the terms *reflexivity*, *transitivity*, *directedness* and *symmetry*), the operations of individual logical concepts like the universal and existential quantifiers (in the concept of directedness, e.g., for *every x*, *there is some y* where R relates x to y), and other logical concepts like conjunction (the *and* in the definition of transitivity) and conditionality (the *if* in the definitions of transitivity and symmetry). If you were to question this proof at any of the steps in the explanation, it would seem to be a very different problem than not understanding a large leap in thought. It would be a problem of understanding the concepts in use and not a problem of understanding how those concepts are combined. (That is the idea, anyway. Exactly how proofs work, and what options we have in understanding them is the topic of the next section.)

Proofs are one side of the logical coin. Models are the other. To explain models, we might consider another possible transition in thought: suppose our mathematicians have another newfangled relation S, and they have concluded that S is transitive and symmetric, but they do not know whether S is directed or not. They wonder, does it follow from these two properties (transitivity and symmetry) that S is reflexive? We can see that the reasoning supplied previously concerning R does not apply in the case of S, since we do not know that S is directed. But having one potential proof that *doesn't* work does not mean that there isn't another proof that *does*. Is it the case that when a relation S is symmetric and transitive, then it must also be reflexive? Our mathematicians think for a while, and sketching some ideas on a blackboard, they draw the following diagram:



This diagram represents a way a relation S could be. Each dot is a different object in the domain of the relation, and a line connecting dot x with dot y indicates that S relates object x to object y. By design, the relation depicted is symmetric, since if there is a line connecting x with y it is line that connects y with x. We can see, too, that whenever we can get from x to y and from y to z by lines, there is a direct line from x to z. This holds even in the case where we can get from x to y and back – whenever there is a line from x to anywhere at all, there is a line looping around from x to itself. So the relation depicted here is *transitive* too. However, it is not reflexive because we have an isolated dot in our diagram. This lonely object is not a counterexample to the claim of symmetry or of transitivity for S, but it shows that if S were like this, it would not be reflexive. We have a *counterexample* to the rule that symmetric and transitive relations must be reflexive.

This counterexample is a *model*. It is not a claim about how the original relation S is. It is a sketch, a representation, showing that if we want an explanation why a relation S *is* reflexive, we cannot appeal merely to its transitivity and its symmetry, since the argument breaks down in circumstances like those depicted in this model. A counterexample is a barrier through which our argument cannot pass.

Cambridge University Press & Assessment
978-1-009-04538-4 — Proofs and Models in Philosophical Logic
Greg Restall
Excerpt
More Information

Here, in a nutshell, we have the distinction between *proofs* and *models*. For the first argument, we have provided a proof, leading from the premises to the conclusion, and for the second, we have provided a *counterexample*, a *model* that renders the premises true and the conclusion false.[3]

The proofs and models we have seen so far are relatively simple, involving reasoning with the quantifiers *all* and *some* and familiar logical connectives like *if* and *and*. There are important questions about the role of these concepts in our argumentation and in our construction of proofs and models. There is broad agreement that these quantifiers and connectives are important. There is less agreement over whether there is anything categorically *distinctive* about those concepts.[4]

≻ ≺

However, there are concepts other than the familiar connectives and quantifiers that have proved important for philosophical logic, and which are amenable to treatment by way of *proofs* and *models*. One example is provided by modal concepts, such as possibility and necessity. However, the proofs and models appropriate for modal operators seem qualitatively different to the models we have seen so far. They do not just represent *a* way things could have been but also represent more than one such way that things could be. To see how they arise, consider the difference between two different arguments:

(3) It's possible that either p or q. So either it's possible that p or it's possible that q.

We can fill in this reasoning into a *proof* in the following way:

(4) Since it's possible that either p or q, we grant some possibility where either p or q holds. Suppose it's p. In that case, we can conclude that (back where we started) it's possible that p, and so it's either possible that p or it's possible that q. On the other hand, suppose that the possibility we granted makes q hold. In that case (also, back where we started), it's possible that q, and again, either it's possible that p or it's possible that q. So, in either case, it's possible that p or it's possible that q, and we're done.

[3] Distinguishing validity as defined by way of proofs and validity defined by way of models was a great conceptual advance in the twentieth century. Consult Zach (1999) for a discussion of the early days of that development of the distinction.

[4] There is extensive literature attempting to characterise the *logical constants* from other concepts. We will not explore it here. Gila Sher's *The Bounds of Logic* is a good place to start (1991).

In this proof, we broke down the leap from premise to conclusion into smaller steps, using more fundamental principles governing possibility and disjunction. This is a good candidate for being a *proof*.

Suppose, on the other hand, we tried a similar argument, with *necessity* in place of possibility.

(5) It's necessary that either p or q. So either it's necessary that p or it's necessary that q.

This argument is less convincing. We can propose a model as a counterexample.

(6) Suppose we have a range of possibilities, where *some* of them (not all) make p true, and the *all the others* (again, not all of possibilities) make q true. In that case, in each possibility, we have either p or q – so from the point of view of any possibility at all, it is *necessary* that either p or q. Nonetheless, we don't have that it is necessary that p (since in some possibilities, p fails), and we also do not have that it is necessary that q (since in other possiblities, q fails). So, in any possibility in our model, we do not have either that it is necessary that p or that it is necessary that q. So this model, at any possibility, the premise is true, and the conclusion is not.

Modal reasoning works just like the other reasoning we have seen. Valid arguments can be broken down into proofs, while invalid arguments can be given models as counterexamples. The models have a richer structure than the models we saw at first. We used not only a representation of one way that things might be but also a range of such representations, a system of different *possible worlds*.

There are many more concepts that can be rigorously explored with proofs and models, like the identity predicate; definite and indefinite descriptions; quantifiers (over objects) beyond the existential and universal quantifier; quantifiers ranging over other domains, such as *functions*, *propositions* or *properties* and much more. However, this will be more than enough to be going on with for what follows.

1.2 Paradoxes

As soon as the field developed accounts of proofs and models – in fact, before these tools took distinct shapes – some natural questions arose. How do we *evaluate* those tools? Do they distinguish good and bad arguments *correctly* (whatever that would mean), or should the main candidates for the correct account of proofs or the correct account of models be revised or rejected? Some of the most active *revisionary* arguments concerning proofs and models have involved different kinds of paradoxes. After all, a paradoxical argument is one where the premises seem true, the argument seems valid and the conclusion seems false. If we want to find a good reason to take some argument that is traditionally thought to be valid to be, in fact, *invalid*, then the paradoxes are where we should look.

Example 1: The liar paradox

Consider this sentence, which says of itself that it is not true.

(7) Sentence (7) is not true.

It seems that we can reason like this. Suppose (7) is true. Then, since (7) says that (7) is not true, then it would follow that (7) is not true, which would contradict (7) being true. In other words, if (7) is true, we have a contradiction. This means that (7) is *not* true, since we reduced the supposition that (7) is true to a contradiction. We've refuted it. But this means that we have proved that (7) is not true and that is what (7) itself says. So we've proved (7). It's true. And so, it isn't. We have proved a contradiction.

This is the *liar paradox*, one example of a *semantic paradox*, and a *paradox of self-reference*.[5] As we will see in the next section, we have used very few logical principles in this line of reasoning, and it very much looks like we have made some kind of mistake, though it has proved very difficult to *locate* the mistake to everyone's satisfaction. For some, the semantic paradoxes like the liar have been seen as reasons to curtail our rules of proof for the logical connectives in some way or other, so as to stop the contradictory conclusion or to render the contradictory conclusion palatable. For others, the fact that logical principles like these are involved in the proof means that the problem must lie elsewhere, either in the so-called definition of the liar sentence (say, we attempt to ban self-reference) or to say that despite appearances, the logic of the truth predicate cannot satisfy the rules used here in the derivation of the paradoxes. There are many different kinds of response to the liar paradox, and we will discuss a representative sample of these in the coming sections, since doing so will give us a range of perspectives on what we are *doing* when we use the logical tools of proofs and models.

[5] Another example, which we will also consider, is Curry's paradox, which uses the conditional, where the liar uses negation. Pick some statement p. Consider c, the class of all classes x where if x is a member of itself, then p. Suppose c is a member of itself. Then, it follows that if c is a member of itself, then p. So, combining those two facts, we have p. In other words, we have just proved that *if* c is a member of itself, then p. So, it follows that c is a member of itself. Again, putting these together, we have proved p. And we made no assumption about p at all.

Example 2: The sorites paradox

Consider a colour strip of colour, shading evenly from red on the left to yellow on the right. Let's divide the strip up into 10 000 evenly sized tiny patches, from left to right, labelled 1 to 10 000. For each number *n* from 1 to 10 000, consider the claim that *patch number n looks red* (*to me*). The first such claim, '*patch* 1 *looks red* (*to me*)', is true. The last such claim, *patch* 10 000 *looks red* (*to me*), is false. The claim '*if patch* 1 *looks red to me, so does patch* 2' also seems true, not just because both patches look red but also because they look indistinguishable to me. This generalises: three distinct features conspire to make it that each claim of the form '*if patch n looks red to me, so does patch n* + 1' seems just as true. First, the strip shades evenly from red to yellow, with no sharp changes in observable colour. Second, we chose very many subdivisions, so each patch differs from its neighbours by at most a tiny difference, and third, my powers of visual discrimination have their limits. So the premises of this argument seem true:

(8) Patch 1 looks red to me.
If patch 1 looks red to me, so does patch 2.
If patch 2 looks red to me, so does patch 3.
⋮
If patch 9 999 looks red to me, so does patch 10 000.
Therefore, patch 10 000 looks red to me.

From these premises, we can draw the conclusion that patch 10 000 looks red to me, using one very simple principle of logic, the inference rule of *modus ponens*, which takes us from a conditional claim of the form *if A then B* and its *antecedent A* to its *consequent*, *B*. Unfortunately, for us, we seem to have a logically valid argument from premises that seem true to a conclusion that seems false.

≻ ≺

If we wish to find a *counterexample* to the validity of the sorites argument, we need to find some model in which the premises hold and in which the conclusion fails. If there is no such counterexample, then either we grant that the argument is valid or we reject the constraint that invalidity must be witnessed by a model as a counterexample. Classical 'two-valued' logical systems have proved difficult to adapt to this task. A two-valued model will assign 'true' or 'false' to each sentence 'patch *n* looks red to me', which will involve assigning either 'true' to each such claim (so this model represents the whole strip as looking red to me) or ‘false’ to each claim (so the model represents the whole strip as *not* looking red to me), or there are two adjacent patches, and the model represents one as looking red to me and the other as not looking red to me. But, as we said, the set-up is designed to make each patch indistinguishably different from its neighbours. So, if *that* is unpalatable, a natural reaction involves expanding the picture of semantic evaluation to allow for more than the two values of ‘true’ and ‘false’: logics with truth-value gaps or a whole panoply of degrees of truth might provide ways to understand the sorites paradoxes. In Section 3, we will examine options for the sorites paradox, as well as other reflections on models that have proved fruitful in philosophical logic in recent decades.

1.3 The Plan

The semantic paradoxes and the sorites paradox are two examples of paradoxes over which a great deal of ink has been spilt in recent decades. The philosophical literature concerning the paradoxes provides one entry point – among many – to the different approaches to understanding the foundations of logical consequence, and this is our entryway into the broader landscape of the use of proofs and models in philosophical logic. The paradoxes are sites where what seemed for all the world to be fundamental principles about proofs and about models come into tension and give rise to what seem to be absurd conclusions. Different proposals for revising those fundamental principles or for defusing the tension provide different approaches to understanding these principles of logic, and they will provide a suitable set of lenses through which to view key ideas in logic as they have developed.

≻ ≺

In Section 2, we will discuss logic from the standpoint of *proof*, giving a quick introduction to the kinds of techniques philosophical logicians have adopted in the study of proof and its application to issues in semantics, epistemology and metaphysics. In this section, we will keep an eye on the kinds of responses people have made to the semantic paradoxes, as these paradoxes provide ample motivation for us to inquire into the costs and benefits of different fundamental proof-theoretical principles. Then, in Section 3, we will do the same thing for *models*, introducing not only the debate over the applicability of the standard two-valued ‘classical’ semantic picture in the light of the paradoxes, but also our sights that will involve discussions of other kinds of models of use in philosophical logic, such as models featuring *possible worlds*, which have proved so useful, and so controversial, in giving an account of the meanings of modal expressions.

After those two sections, we will wrap up with Section 4, in which we explore not only the ways that these tools are used in the discussions of the paradoxes but also some other natural questions, including the relationship between proofs and models themselves. In particular, we will ask which *proofs* and *models* should be taken to be fundamental. But, first, let us turn our attention to proofs.

2 Proofs

Let's start with the proof labelled (2) on page 2. That is a *proof* that R is reflexive. It has three premises: R is directed, R is symmetric and R is transitive. It lays out a path from those premises to the conclusion, leaving nothing out. The aim of a proof is not just to convince us *that* some conclusion is true but also, in some sense, to make explicit *how* that conclusion follows from the premises.

Notice that this proof does not just start from the premises and lead to the conclusion, with each intermediate step following from the ones granted before it.[6] There are some other important features of our reasoning that are worth examining. First, our proof includes an *imperative*: 'take an object a'. This sentence is not a premise, nor is it a conclusion, and it is not another statement that follows from the premises. It is an invitation. It cannot be true or false. We cannot assert or deny it. Second, the term 'b' in the proof also has an interesting status. We moved from the claim that R is directed (so, in particular, there must be some object to which R relates a) to *calling* one such object b. The fact that R relates a to a given object b does not logically follow from the claim that R is directed. R could be directed without R relating a to this particular b (whichever b it happens to be).[7] So much more is going on in this proof than simply working out conclusions from things we have granted. There are different steps where objects are given names, and speech acts, other than asserting, are involved as well. Proofs have complex structure. A crucial constraint, though, is that a proof is not simply a statement of the premises and the conclusion – at least, not in most cases. To *prove* some conclusion C from some premises $P_1, \ldots, P_n$, you must somehow trace the connection *from* $P_1, \ldots, P_n$ *to* C.

[6] Proofs with that direct 'linear' structure *Hilbert Proofs*. We will see these in the next section.

[7] If you find this puzzling, think of a concrete case. I have a son. As a matter of fact, Zachary is my son. The fact that *Zachary* is my son does not follow as a logical consequence of the fact that I have a son because there are *other* ways I could have had a son, other than Zac. Similarly, if R relates a to some object, in any particular circumstance in which that is true, we could call that object b. Given that choice, it would still not follow that in every circumstance where R relates a to something, that R must relate a to *that* object b. It might have related a to some other object instead.

2.1 Proof Structures

What kind of structure can a proof have? How can such connections be made, leading from premises to conclusions? In philosophical logic, over recent decades, the focus of research has not been on any general overarching theory of the structure and properties of proofs *as such*.[8] We have seen one example, but in the study of proofs and the formal development of proofs through the years, formal accounts of different kinds of proofs have been represented in many different ways, each with different costs and benefits.

Let's start with the standard logical connectives such as ∧ (conjunction), ∨ (disjunction), → (the conditional) and ¬ (negation) and the quantifiers such as ∀ (universal) and ∃ (existential) and consider proofs for statements exploiting the concepts in this vocabulary. As mentioned previously, a simple analysis of proofs is given by the structure of *Hilbert Proofs*. A Hilbert Proof from premises $P_1, \ldots, P_n$ to conclusion C is a *list* of formulas, ending in the conclusion C, and each of them is (a) one of the premises P_i, (b) one of the *axioms* or (c) a formula that follows from earlier formulas in the list by way of the *rules*. Whether something counts as a Hilbert proof depends on the choices of *axioms* and *rules*. If we focus just on the connectives → and ¬, we can provide a very simple Hilbert proof system for propositional logic. There are three axiom schemes:[9]

WEAKENING: $A \to (B \to A)$
DISTRIBUTION: $(A \to (B \to C)) \to ((A \to B) \to (A \to C))$
CONTRAPOSITION: $(\neg B \to \neg A) \to (A \to B)$

These are *schemes* rather than individual formulas because we count *any* formula that fits the shape as an instance of the axiom. So, for example, $p \to (q \to p)$ and $p \to (p \to p)$, and $\neg(p \to q) \to (p \to \neg(p \to q))$ are each instance of the WEAKENING axiom scheme. Our Hilbert proof system has just one *rule*:

MODUS PONENS: $A \to B, A \implies B$

We understand the rule like this: if, in a proof, we have already written down $A \to B$ and A (in either order), then we can add B to our proof. With this choice of axioms and rules, we can construct proofs in our axiom system

[8] See some of Dag Prawitz's papers to see what *could* be done in this direction (1973, 1974, 2019).

[9] Hilbert proofs are named after German mathematician David Hilbert (1862–1943), for whom the *axiomatisation* of mathematics, with a precise formally specified notion of proof, was a central task to the foundation and justification of mathematical methods (see Sieg, 2013; Zach, 2019).