STUDIES IN THEORETICAL PHILOSOPHY

vol. 6

VITTORIO KLOSTERMANN

ALEXANDRA ZINKE

The Metaphysics of Logical Consequence

VITTORIO KLOSTERMANN

Gedruckt mit freundlicher Unterstützung der
Geschwister Boehringer Ingelheim Stiftung für Geisteswissenschaften
in Ingelheim am Rhein.

Bibliographische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über *http://dnb.dnb.de* abrufbar.

Gedruckt auf Eos Werkdruck von Salzer,
alterungsbeständig ISO 9706 und PEFC-zertifiziert.
Druck: docupoint GmbH, Magdeburg
Printed in Germany
ISSN 2199-5214
ISBN 978-3-465-04345-4

– to Stromi and Misriy –

Acknowledgment

This book is based on my PhD thesis, which I submitted in 2013 at the University of Konstanz. While working on the thesis – but also before and after – Wolfgang Spohn was a real *Doktorvater* to me, caring and daring. There was never any pressure, but a continuous warmhearted encouragement to think things through. My gratitude also goes to Holger Sturm, my second supervisor, for sharing his expertise and for invaluable literature suggestions.

Moreover, I wish to thank the editors of *Studies in Theoretical Philosophy*, Tobias Rosefeldt and Benjamin Schnieder, for being so uncomplicated, Anastasia Urban at *Klostermann* for her kind guidance during the publication process, Christopher von Bülow for helping with the typesetting, and Peter Anstee for English proofreading. I am indebted to the *Studienstiftung des deutschen Volkes* for financial support (but please change your name).

Somehow, I always wanted to write this book. And then again, I definitely didn't. I am thankful to my mother and my brother, and to my friends Sandra and Katrin for lending me their ears again and again when I once more doubted my doings.

Then there is this one person where words fail me. Nobody influenced and influences my philosophical thinking nearly as thoroughly and deeply as he did and does. He is the Logic of my Life. And I love the consequences!

Contents

Introduction

Logic is the art and heart of reasoning. As such it pervades all our theoretical endeavors. It is omnipresent in philosophy, in the sciences and humanities, in everyday life. Logic is not only all-pervasive, but also most general. It is concerned with the structure of reasoning as such. Logic is 'topic-neutral', independent of the specific contents of our reasoning. It determines which arguments are valid and which are invalid. Logic itself is impeccable and thus is beyond any dispute. Where our favorite empirical or metaphysical theories may fail, logic still prevails.

The special status of logic has given rise to high expectations. Many thinkers had the hope that it could play a foundational role and maybe replace metaphysics, the most fundamental, but also contentious field of theoretical research. Where metaphysics even lacks a uniform methodology, logic was meant to provide a firm basis and a neutral, general, and immaculate scheme. Although time has shown these expectations to be exaggerated, logic has had a great influence on philosophy and metaphysics itself. It has purified our thoughts and put those that survived onto secure footing. It has done so, people claim, in an unbiased way, independently of any metaphysical assumptions.

My aim in this book is to challenge this view and to show that logic is by no means independent of metaphysics. Both are deeply intertwined. More specifically, I will argue that a proper definition of logical consequence, the core concept of logic, rests on metaphysical presuppositions. To the extent that logic can replace metaphysics, it is able to do so only because it is imbued with metaphysical considerations. Logic is not as nearly as neutral a frame of thought as has often been assumed.

I expect serious and immediate resistance, based not only on the desire to preserve the purity of logic, but on the following obvious argument. The classical model-theoretic account of logical consequence involves only well-defined set-theoretic concepts. It defines logical consequence as truth-preservation in all models: a formula φ logically follows from a set of formulas Γ iff it holds that φ is true in

every model in which all elements of Γ are true. No contentious metaphysical notions seem to be involved.[1]

It would, however, be premature to conclude that logic does not have any metaphysical presuppositions. In so far as the model-theoretic notion is to play the role we assign to it, it must capture our pretheoretical concept of logical consequence. Or, if we are skeptic that there is a pretheoretical or intuitive notion of logical consequence, the definition must at least capture some of the essential features we assign to logically valid arguments. Logic has its special status only because we consider logically valid arguments to have certain distinguished modal, epistemic and formal properties. *Prima facie* nothing whatsoever secures that truth-preservation in all models has anything to do with this. As Timothy Williamson (2007: 65) puts it: "[…] the mathematical rigor, elegance, and fertility of model-theoretic definitions of logical consequence depend on their freedom from modal and epistemological accretions. As a result, such definitions provide no automatic guarantee that logical truths express necessary or a priori propositions." In order to establish that only arguments with certain distinguished features are truth-preserving in all models, we have to spell out what truth-preservation in all models amounts to, i.e., we have to lay out what is modeled by the models of model theory. It is here, I claim, that metaphysics enters the stage.

I will plunge right in and draw your attention to some of the central characteristics of the pre-theoretic notion of logical consequence and therefore those desiderata that a formal definition has to meet. I do so by way of example. The argument:

(A1) All humans are mortal. Plato is human. Therefore, Plato is mortal.

is a paradigm case of an argument that is pretheoretically logically valid.[2] The same holds for the two following arguments:

(A2) All humans are male. Plato is human. Therefore, Plato is male.

1 I ignore here that sets themselves are ontologically problematic entities.

2 Throughout this book, I am concerned solely with arguments that have exactly one conclusion and bracket multiple-conclusion arguments. I allow the set of premises to be empty, however. As I take the totality of premises to be a set, I also do not take into account the order of the premises and that they might have multiple occurrences.

(A3) All humans are female. Plato is human. Therefore, Plato is female.

Valid arguments allow for different combinations of truth values. Importantly, however, we will not find a logically valid argument containing only true premises and a false conclusion. Logically valid arguments are *truth-preserving*. Yet that characteristic seems insufficient for validity, as the following argument shows:

(A4) Plato is human. Therefore, Plato is male.

Both, the premise and the conclusion of (A4) are true. Nevertheless, the argument is invalid. It is truth-preserving, but only – as one might say – accidentally so. In contradistinction to the arguments (A1)-(A3), where the conclusion must be true if the premises are true, the conclusion of (A4) is not guaranteed to be true by the truth of the premise. In a logically valid argument, the premises somehow necessitate the conclusion. Logically valid arguments are not merely truth-preserving, but *necessarily* truth-preserving.

Yet this amendment still does not suffice. Consider (A5):

(A5) Gaia is a mother. Therefore, Gaia is female.

In (A5), the conclusion also follows from the premises by necessity. Nevertheless, according to the standard view, (A5) does not qualify as logically valid. In reaction to cases like (A5), it is usually claimed that logically valid arguments must not only be necessarily truth-preserving, but also *formal*. They are said to be truth-preserving in virtue of their *form*. The notion of form then is evoked in explanations like the following.

The difference between (A1) and (A5) is a difference with respect to formality. (A5) is truth-preserving only in virtue of the particular meaning of the terms "mother" and "female". If we were to substitute "male" for "female", the argument would no longer be truth-preserving. The fact that (A5) is truth-preserving rests on certain semantic facts. The validity of (A1) does not, however, depend on (A1) being about Plato or about humanity or mortality. The argument is necessarily truth-preserving not because of its particular semantic content, but because of its form alone. To illustrate this position, consider a further example:

(A1*) All planets are eternal. Pluto is a planet. Therefore, Pluto is eternal.

The logically valid argument (A1*) is of the same form as (A1). In fact, the arguments (A1) and (A1*) are both instances of the following schema:

(S1) All *F* are *G*. *a* is *F*. Therefore, *a* is *G*.

To account for the validity of (A1*) and (A1), we need not know anything about the meaning of the particular terms occurring in the respective arguments; it is fully sufficient to know that these arguments are instances of (S1). They are valid in virtue of their being an instance of a valid schema. (A5), on the other hand, is not valid in virtue of it being an instance of a valid schema, but in virtue of its content.

These examples suggest two intuitive desiderata on logical validity. Logically valid arguments must be *necessarily truth-preserving* and they must be *formal*. It is far from obvious that the classical model-theoretic definition of consequence as truth-preservation in all models meets these requirements. Whether it does, will depend on our conception of a model.

There are competing conceptions. John Etchemendy, in his groundbreaking monograph *The Concept of Logical Consequence* (1990) makes two proposals for what the models of model theory might model. According to one conception, we may understand a model as representing a way the world might be. This is known as the *representational* notion of a model. Alternatively, we might conceive of it as representing a possible interpretation of the linguistic items involved. This constitutes the so-called *interpretational* notion. Truth-in-a-model, conceived of representationally, amounts to truth with respect to a way the world might be. Truth-in-a-model, understood interpretationally, amounts to truth with respect to a certain interpretation of the language.

If one understands a model representationally, the definition of logical consequence as truth-preservation in all models seems to obviously fulfill the desideratum of necessary truth-preservation. If models stand for worlds, then, so it seems, only metaphysically necessary arguments are truth-preserving in all models. However, the second criterion, emphasizing formality, appears to be violated. Indeed,

the representational conception seems to declare argument (A5) logically valid. For this very reason, there is a wide consensus in the literature that the representational definition of logical consequence is extensionally inadequate and thus fails to capture our intuitive notion of logical consequence. Besides extensional inadequacy, there is a second kind of objection. The representational definition, the argument goes, rests on dubious modal notions: it grounds the notion of logical consequence on the notion of metaphysical necessity, which itself stands in need of analysis. For these reasons, the representational definition is usually considered to present an inadequate definition of our concept of logical consequence.

It is therefore no surprise that the interpretational conception of logical consequence has been dominant in philosophical logic. According to this view, a model provides a specific interpretation of the sentences used in the argument. Truth in a model then is truth under a certain interpretation of these sentences. Given this construal of the model-theoretic definition, logical consequence amounts to truth-preservation under all interpretations.

One major advantage of this conception is that it appears to be 'free of metaphysics'. In contrast to the representational definition, no unexplained modal notion seems to play a role. Furthermore, an interpretational reading clearly enforces the formality condition. Argument (A5), for example, is obviously not truth-preserving under all interpretations: if you reinterpret the predicate "is female" as having the meaning of "is male", the conclusion will be false, while the premise remains true. There is a drawback to this theory, however. The interpretational reading struggles with the criterion of necessary truth-preservation. Etchemendy (1990) famously argues that there is no conceptual reason why truth-preservation in all interpretations should guarantee truth-preservation in all worlds. He actually provides examples that aim at showing that the interpretational definition is not even materially adequate. There seem to be arguments which satisfy the interpretational definition, but fail to be necessary truth-preserving.

Furthermore, there is yet another problem for the interpretational definition: the so-called *problem of the logical constants*. If reinterpretation is unrestricted, the intuitively valid argument (A1) would no longer come out as valid. Not only the terms "human", "mortal" and "Plato" could be given a novel meaning, but also the term "all". In that case,

there would be a reinterpretation of the argument such that it is no longer truth-preserving. In fact, if the interpretational definition allows a reinterpretation of the so called "logical constants" ("all", "and", "not", etc.), no argument whatsoever would be declared logically valid by this definition. On pain of trivialization, the interpretational definition must declare such interpretations inadmissible and thereby presuppose a demarcation of the logical from the non-logical terms.

The problem of logical constants is generally held to be the most pressing problem of the interpretational definition of logical consequence. Tarski himself mentioned this problem at the end of his seminal article "On the concept of logical consequence":

> I am not at all of the opinion that in the result of the above discussion the problem of a materially adequate definition of the concept of consequence has been completely solved. On the contrary, I still see several open questions, only one of which – perhaps the most important – I shall point out here. Underlying our whole construction is the division of all terms of the language discussed into logical and extra-logical. This division is certainly not quite arbitrary. […] On the other hand, no objective grounds are known to me which permit us to draw a sharp boundary between the two groups of terms. […] Further research will doubtless greatly clarify the problem which interests us. Perhaps it will be possible to find important objective arguments which will enable us to justify the traditional boundary between logical and extra-logical expressions. (Tarski 1936: 420)

Three decades after this remark, in a lecture from 1966, published 1986 as "What are Logical Notions?", Tarski displayed a more optimistic attitude and characterized the logical constants by reference to the mathematical property of permutation invariance. Yet it has ever since been disputed whether permutation invariance yields an adequate criterion for logical constants. It is fair to say that the problem of logical constants has not yet been solved.

Reviewing the situation, we are presented with a dilemma. There are two readings of the model-theoretic definition of logical consequence. Under a representational conception, the definition of logical consequence seems to analyze logical consequence with unexplained modal notions and the definition fails to satisfy the criterion of for-

mality. Using an interpretational reading, it remains questionable whether the criterion of necessary truth-preservation is satisfied, and the definition depends on a solution to the problem of logical constants which so far has been lacking.

The dilemma is serious given the alternatives presented so far. I venture to show, however, that it can be avoided if we adopt a different perspective. Upon closer consideration, it will turn out that the interpretational and the representational definition do not constitute real alternatives. Given certain plausible assumptions they turn out to be extensionally equivalent. Furthermore, they struggle with analogous problems, which I dub the *problem of admissible interpretations* for the interpretational view and, as far as the representational definition is concerned, the *problem of admissible states*. I will argue that both definitions rely on a prior demarcation of the admissible from the inadmissible interpretations and states, respectively. In particular, the problem of logical constants turns out to be merely a special case of this more general and deeper problem type. The demarcation problem is essentially the same for both approaches to logical consequence. The central question of the philosophy of logic is the demarcation of the admissible from the inadmissible models. I will claim that such a demarcation requires substantial semantical and metaphysical considerations.

Before going into the sundry details of the discussion, some preliminary methodological remarks will be helpful. I am primarily interested in definitions of logical consequence, but discussions in terms of logical truth are often less cumbersome. Fortunately, we can treat the notions of logical consequence and logical truths as interdefinable. The standard model-theoretic definition of logical consequence construes logical consequence as truth-preservation in all models. The model-theoretic definition of logical truth understands logical truth as truth in all models. We can thus translate the one definition into the other as follows: a sentence is logically true ("$\models \varphi$") iff it is a logical consequence of the empty set. Conversely, a sentence φ follows logically from a set of sentences Γ ("$\Gamma \models \varphi$") iff the material conditional containing the conjunction of the members of Γ as antecedent and φ as consequent is a logical truth. In short: $\Gamma \models \varphi \Leftrightarrow \models \gamma_1 \wedge \ldots \wedge \gamma_n \rightarrow \varphi$

(where $\Gamma = \{\gamma_1, \ldots, \gamma_n\}$).[3] I am confident that this also captures our pretheoretic intuitions about the relation of logical consequence and logical truth. Therefore, I allow myself to switch back and forth between the two notions.

Let me also address a methodological worry having to do with the relation of natural and formal languages. Our pretheoretic intuitions about the concept of logical truth pertain to natural language sentences like (A1)–(A5). Definitions of logical truth, such as our standard model-theoretic definition, are, however, formulated for formal languages. One might thus doubt the legitimacy of discussing these definitions with recourse to natural language examples on the grounds that, in order to test whether or not a certain natural language sentence is declared logically true by a given definition of logical truth, we have to first *formalize* the respective sentence. It is therefore important to say something about formalization.

A formalization is a function[4] f from a class N of natural language sentences on a class F of sentences in our chosen formal language. A formal definition D of logical truth determines for each sentence of the formal language whether or not it is logically true. We can think of this classification as a function g^D from the class F on the set $\{0, 1\}$: the function g^D maps a sentence to 1 just in case it is declared a logical truth according to the definition D, and to 0 otherwise. Our intuitive judgments whether or not a given sentence is logically true apply to sentences in natural language. We can also model this intuitive classification of natural language sentences by a function. Let h be a function mapping a natural language sentence to 0 or 1. We then judge the definition D of logical truth as extensionally adequate iff $g^D(f(p)) = h(p)$ for every sentence $p \in N$.

[3] This equivalence only holds if we assume that the set Γ of premises is finite or assume that the language allows infinite conjunctions. This limitation will not play any role in the present discussion.

[4] By calling f a *function*, I make the simplifying assumption that a natural language sentence is mapped on exactly one formula. By letting the domain of f be the class of all natural language sentences, I furthermore presume that there is a formula in the target formal language for every natural language sentence. I do not necessarily want to subscribe to these assumptions, but we can safely disregard these peculiarities in the present context.

Trivial as this reconstruction may be, it makes the role of the formalization explicit. It shows that whether or not a natural language sentence is declared logically true by a given definition of logical truth crucially depends on the way this sentence is formalized. If there are no restrictions on formalizations, i.e., if arbitrary functions f are allowed, it holds that for any given functions g^D and h, and arbitrary sentence p, we can provide an f such that $g^D(f(p)) = h(p)$ and an f such that $g^D(f(p)) \neq h(p)$. In other words: for any natural language sentence and any definition of logical truth, we can formalize the sentence such that it comes out logically true (or not logically true). If there are no restrictions on the formalization f, we can yield any arbitrary definition of logical truth extensionally adequate or inadequate, simply by choosing a formalization function f that yields this result.

Which lesson is to be drawn from these considerations? The classification of natural language sentences as logically true does not, by itself, support or undermine a certain formal definition of logical truth. All that we can put to the test is a package consisting of a certain formalization f of the natural language sentence and a formal definition of logical truth D. If the natural language sentence is classified contrary to intuition, the formalization or the formal definition of logical truth (or both) are to be rejected. If, on the other hand, the natural language sentence is classified in agreement with the pretheoretic notion of consequence, this does not speak in favor of either the definition or the formalization: in such case, it may be that both are adequate, or that both inadequate, or that only one of them is correct.

We could break out of this quandary if there were a reliable means to evaluate a given formalization in isolation. However, the conditions for adequate formalizations are themselves highly contentious.[5] Not only are there no generally accepted rules for generating the formalization of a natural language sentence, but there are no strict criteria to evaluate it. We formalize sentences according to our gut feelings and by means of some rules of thumb. Nothing even remotely resembling a consensus on the exact adequacy conditions of formalization has been reached.

[5] See especially Brun 2003 for a detailed investigation of the adequacy conditions of formalizations.

A lot would be gained if – at least at this early stage of the debate – questions concerning the adequacy of formalization could be disregarded. Indeed, I think that we can apply the definitions of logical consequence and logical truth to natural language arguments and sentences directly and thereby sidestep the topic of formalization. In this respect, I am a student of Richard Montague, who famously applied the model-theoretic concepts and the rules of formal semantics to natural language sentences. He thought that there is no relevant difference between formal and natural language – at least as long as we only consider a limited fragment of natural language.[6] I agree and use the same method which I will briefly illustrate. To check whether a given sentence turns out to be logically true under the model-theoretic definition, I will skip formalizing it, but apply the formal interpretation function to the original sentence straightaway. Consider exemplarily the sentence "Plato is a philosopher". This sentence could be formalized as "Fa". One could then check whether "Fa" is true in all models, i.e., whether $i(a) \in i(F)$ for all interpretation functions i (and all domains). I will avoid this detour by applying the interpretation function to "Plato" and "is a philosopher" directly, i.e., I examine whether $i(\textit{Plato}) \in i(\textit{is a philosopher})$ for all interpretation functions i. The sentence "Plato is a philosopher" is a logical truth only if this condition is fulfilled. I sidestep the problem of formalization by eliminating the necessity of formalizations.

Importantly, this procedure is unproblematic only because I restrict the discussion to a small and regimented portion of natural language. I confine myself to declarative sentences not involving any complicated adverbial constructions, minor sentences, intensional operators, deictic or indexical elements. Also, I do not take into account paradoxical sentences like, e.g., the Liar sentence. I only examine ordinary and harmless subject-predicate constructions like "Plato is a philosopher" or "Gaia is a mother". One could almost say that I discuss a hybrid of natural and formal language. In fact, I will not hesitate to use sentences like "$\forall x$ (x is a planet $\rightarrow$ x is eternal)," which combine elements from both natural language and the formal language of first-order predicate logic.

6 See especially Montague 1970[a]: 188, and 1970[b]: 222.

This is a quite common methodology. Even Tarski, who prominently confined his research to formal languages, remarks that the formal languages he is interested in can be understood as 'fragments of natural language':

> I should like to emphasize that, when using the term 'formalized languages', I do not refer exclusively to linguistic systems that are formulated entirely in symbols, and I do not have in mind anything essentially opposed to natural languages. On the contrary, the only formalized languages that seem to be of real interest are those which are fragments of natural languages (fragments provided with complete vocabularies and precise syntactical rules) or those which can at least be adequately translated into natural languages. (Tarski 1960: 68)

I do not in any way want to suggest that the logical form of a sentence can be read off its surface appearance. Quite to the contrary: I hold the restriction to a very small portion of natural language to be necessary, because I am, like so many others are, convinced that "grammatical form misleads as to logical form" (Strawson 1952: 51).[7] The logical form of a sentence or an argument cannot be easily read off its surface structure and I do not intend to conceal that it often requires deep linguistic and philosophical efforts to uncover the logical form of natural language expressions. Indeed, I consider the relation between logic and natural language as one of the most intriguing and fundamental topics in philosophy. Here I simply bracket any associated problems and confine myself to a more unproblematic part of natural language in order to engage with the central questions concerning logical truth and consequence without being compelled to make unnecessary or distracting detours.

Let me outline the structure of the book. Chapter 1 explains the notions of interpretational and representational semantics in detail. It will be argued that, while Tarski's original definition of logical consequence has to be read interpretationally, our contemporary model-theoretic definition is ambiguous between both readings. The stand-

[7] There are innumerable examples to support the *misleading form* thesis. Let me here only depict one of the earliest ones from Plato's *Euthydemus*. The following arguments have the same surface structure, but not the same logical form: (i) This is a pen. This is blue. Therefore, this is a blue pen. (ii) That dog is a father. That dog is his. Therefore, that dog is his father.