Trends in Logic 43

Thomas Piecha
Peter Schroeder-Heister *Editors*

Advances in Proof-Theoretic Semantics

Springer Open

Trends in Logic

Volume 43

TRENDS IN LOGIC
Studia Logica Library

VOLUME 43

SCOPE OF THE SERIES

The book series Trends in Logic covers essentially the same areas as the journal Studia Logica, that is, contemporary formal logic and its applications and relations to other disciplines. The series aims at publishing monographs and thematically coherent volumes dealing with important developments in logic and presenting significant contributions to logical research.

The series is open to contributions devoted to topics ranging from algebraic logic, model theory, proof theory, philosophical logic, non-classical logic, and logic in computer science to mathematical linguistics and formal epistemology. However, this list is not exhaustive, moreover, the range of applications, comparisons and sources of inspiration is open and evolves over time.

More information about this series at http://www.springer.com/series/6645

Thomas Piecha · Peter Schroeder-Heister
Editors

Advances in Proof-Theoretic Semantics

Springer Open

Editors
Thomas Piecha
Department of Computer Science
University of Tübingen
Tübingen
Germany

Peter Schroeder-Heister
Department of Computer Science
University of Tübingen
Tübingen
Germany

ISSN 1572-6126 ISSN 2212-7313 (electronic)
Trends in Logic
ISBN 978-3-319-22685-9 ISBN 978-3-319-22686-6 (eBook)
DOI 10.1007/978-3-319-22686-6

Library of Congress Control Number: 2015949264

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Contents

Advances in Proof-Theoretic Semantics: Introduction

Thomas Piecha and Peter Schroeder-Heister

Abstract As documented by the papers in this volume, which mostly result from the second conference on proof-theoretic semantics in Tübingen 2013, proof-theoretic semantics has advanced to a well-established subject in philosophical logic.

Keywords Proof-theoretic semantics

In the mid-1980s, the term "proof-theoretic semantics" (Schroeder-Heister 1991 [13], and before in lectures) was proposed (1) to explain meaning in terms of proof rather than denotation or truth and (2) to give a semantics for proofs. Though related to the meaning-as-use approach in the philosophy of language, and belonging to what in a more general setting has been called "inferentialism" (Brandom 1994 [1]), the intention of proof-theoretic semantics was to capture and continue the specific line of research that originated from the work of Gentzen (1934/35) [5] (and also Jaśkowski 1934 [6]) and was taken up and developed, amongst others, by Lorenzen (1955) [8], Prawitz (1965, 1971) [11, 12], von Kutschera (1968) [15], Martin-Löf (1975, 1984) [9, 10], and Dummett (1975, 1991) [2, 3]. Whereas in the 1980s proof-theoretic semantics was almost exclusively the business of proof-theorists, the field has since expanded into the wider area of philosophical logic. The first conference on proof-theoretic semantics was held in 1999 in Tübingen with a special issue of *Synthese* originating from it, which was published in 2006 (Kahle and Schroeder-Heister 2006 [7]). At the time of this first conference, the subject still belonged to a relatively small community of logicians and philosophers. This has changed in the meantime. One only needs to look at the multitude of papers published on issues of proof-theoretic semantics in the past decade and at the widespread usage of this term. "Proof-theoretic semantics" is no longer the provocative title it used to be, containing an alleged *contradictio in adjecto* between proof theory as dealing with syntax, and semantics as dealing with meaning. The link between proofs and meaning is well-established now. Given the growing interest in the subject, we organised a

T. Piecha (✉) · P. Schroeder-Heister
Department of Computer Science, University of Tübingen, Tübingen, Germany
e-mail: thomas.piecha@uni-tuebingen.de

P. Schroeder-Heister
e-mail: psh@uni-tuebingen.de

T. Piecha and P. Schroeder-Heister (eds.), *Advances in Proof-Theoretic Semantics*,
Trends in Logic 43, DOI 10.1007/978-3-319-22686-6_1

second conference on proof-theoretic semantics in Tübingen in 2013 to discuss the advances in the now well-established field (for overviews see Wansing 2000 [16], Schroeder-Heister 2012 [14]). Some speakers of the second conference had already spoken at the first, namely Došen, Dyckhoff, Hallnäs, Kahle, Prawitz, Sundholm, Tait and Usberti.

The presentations given at the conference were the following.

- Sergei N. Artëmov: On Brouwer-Heyting-Kolmogorov provability semantics
- Walter Dean and Hidenori Kurokawa: Kreisel's second clause and the Theory of Constructions
- Kosta Došen: Two ways of general proof theory
- Roy Dyckhoff: Generalised elimination rules
- Lars Hallnäs: On the proof-theoretic foundations of set theory
- Wilfrid Hodges: The choice of semantics as a methodological question
- Reinhard Kahle: The mode of presentation
- Yoshihiro Maruyama: On paradoxes in proof-theoretic semantics
- Jan von Plato: Explicit composition and its application in normalization proofs
- Dag Prawitz: Remarks on relations between Gentzen and Heyting inspired PTS
- Giovanni Sambin: Unification of logics by reflection
- Göran Sundholm: BHK and Brouwer's theory of the creative subject
- William W. Tait: Compositional semantics for predicate logic: Eliminating bound variables from formulas and deductions
- Gabriele Usberti: Intuitionism, the Paradox of Knowability and empirical negation
- Heinrich Wansing: A two-sorted typed lambda-calculus

This volume collects the contributions of many of the participants, not necessarily in the form presented, and also some additional papers by authors who did not speak at the conference. Therefore it exemplifies from many perspectives what proof-theoretic semantics is about. The papers by Prawitz and Dean and Kurakowa confront the proof-theoretic approach with the constructive semantics of Heyting and of Kreisel and Goodman, respectively. Došen pleads for a categorial approach to proof-theoretic semantics, arguing that it best exhibits the structures of deductions. Dyckhoff's paper is a critical examination of approaches in proof-theoretic semantics based on general elimination rules of a certain form. Maruyama gives a taxonomy of various forms of paradoxes based on a categorial approach to proof-theoretic harmony. Usberti proposes a solution to the epistemic knowability paradox from the standpoint of logic. Von Plato investigates the significance of a rule of explicit composition in natural deduction which makes the substitution of a derivation for an open assumption an inference step of its own. Kahle applies proof-theoretic semantics to the treatment of equality by elucidating the difference between extensional and intensional equality in a non-denotational way. Hallnäs sketches some ideas towards a proof-theoretic foundation for set theory using generalisations of definitional reflection. The paper by Hodges, which is the only one not written from the perspective of a proof-theoretic semanticist or constructivist, defends model theory against false claims from the proof-theoretic semantics 'camp'. It is followed by a reply by Došen.

Goldfarb's paper has been circulating for more than 15 years and has been referred to repeatedly. It was the subject of Dummett's presentation at the first conference in 1999. It represents significant results on the proof-theoretic semantics of intuitionistic logic, in particular on the question of completeness. Another contribution not presented at the conference is a chapter of Ekman's thesis (Ekman 1994 [4]). It uses an interesting way of associating labels with formulas that are proved, which is different from standard Curry–Howard term-annotation and particularly suited to analyse non-well-founded and paradoxical reasoning, a topic which has recently gained much attention in the proof-theoretic semantics community. There are also papers by us, the editors: An overview on results concerning completeness in proof-theoretic semantics and a presentation of three open problems that are considered significant for the further development of the subject.

We would like to thank all participants of the conference, and in particular all contributors to this volume, who made this work possible, as well as those who helped with reviewing and editing the papers. We are also very grateful to Marine Gaudefroy-Bergmann for invaluable organisational assistance. The second editor is particularly grateful to Reinhard Kahle and Thomas Piecha, who organised the conference as a present for his 60th birthday including a special colloquium with friends and colleagues.

Acknowledgments This work was supported by the French-German ANR-DFG project "Hypothetical Reasoning—Its Proof-Theoretic Analysis" (HYPOTHESES), DFG grant Schr 275/16-2. Its completion was supported by the French-German ANR-DFG project "Beyond Logic: Hypothetical Reasoning in Philosophy of Science, Informatics, and Law", DFG grant Schr 275/17-1.

References

1. Brandom, R.B.: Making It Explicit: Reasoning, Representing, and Discursive Commitment. Harvard University Press, Cambridge (1994)
2. Dummett, M.: The justification of deduction. In: Proceedings of the British Academy, pp. 201–232 (1975). Separately published by the British Academy 1973. Reprinted in Dummett, M.: Truth and Other Enigmas, Duckworth, London (1978)
3. Dummett, M.: The Logical Basis of Metaphysics. Duckworth, London (1991)
4. Ekman, J.: Normal proofs in set theory. Ph.D. thesis, University of Göteborg (1994)
5. Gentzen, G.: Untersuchungen über das logische Schließen. Mathematische Zeitschrift **39**, 176–210, 405–431 (1934/35). English translation in: Szabo, M.E. (ed.) The Collected Papers of Gerhard Gentzen, pp. 68–131. North Holland, Amsterdam (1969)
6. Jaśkowski, S.: On the rules of suppositions in formal logic. Stud. Log. **1**, 5–32 (1934). Reprinted in: McCall, S. (ed.) Polish Logic 1920–1939, pp. 232–258. Oxford (1967)
7. Kahle, R., Schroeder-Heister, P. (eds.): Proof-Theoretic Semantics. Springer. Special issue of Synthese **148**(3), 503–743 (2006)
8. Lorenzen, P.: Einführung in die operative Logik und Mathematik. Springer, Berlin (1955). 2nd edn. 1969

9. Martin-Löf, P.: An intuitionistic theory of types: predicative part. In: Rose, H.E., Shepherdson, J.C. (eds.) Logic Colloquium '73: Proceedings of the Logic Colloquium, Bristol, July 1973, pp. 73–118. North Holland, Amsterdam (1975)
10. Martin-Löf, P.: Intuitionistic Type Theory. Bibliopolis, Napoli (1984)
11. Prawitz, D.: Natural Deduction: A Proof-Theoretical Study. Almqvist &Wiksell, Stockholm (1965). Reprinted by Dover Publications, Mineola NY (2006)
12. Prawitz, D.: Ideas and results in proof theory. In: Fenstad, J.E. (ed.) Proceedings of the Second Scandinavian Logic Symposium (Oslo 1970), pp. 235–308. North-Holland, Amsterdam (1971)
13. Schroeder-Heister, P.: Uniform proof-theoretic semantics for logical constants (Abstract). J. Symb. Log. **56**, 1142 (1991)
14. Schroeder-Heister, P.: Proof-theoretic semantics. In: Zalta, E.N. (ed.) Stanford Encyclopedia of Philosophy (2012). http://plato.stanford.edu/entries/proof-theoretic-semantics/
15. von Kutschera, F.: Die Vollständigkeit des Operatorensystems $\{\neg, \wedge, \vee, \supset\}$ für die intuitionistische Aussagenlogik im Rahmen der Gentzensemantik. Archiv für mathematische Logik und Grundlagenforschung **11**, 3–16 (1968)
16. Wansing, H.: The idea of a proof-theoretic semantics. Stud. Log. **64**, 3–20 (2000)

On the Relation Between Heyting's and Gentzen's Approaches to Meaning

Dag Prawitz

Abstract Proof-theoretic semantics explains meaning in terms of proofs. Two different concepts of proof are in question here. One has its roots in Heyting's explanation of a mathematical proposition as the expression of the intention of a construction, and the other in Gentzen's ideas about how the rules of Natural Deduction are justified in terms of the meaning of sentences. These two approaches to meaning give rise to two different concepts of proof, which have been developed much further, but the relation between them, the topic of this paper, has not been much studied so far. The recursive definition of proof given by the so-called BHK-interpretation is here used as an explication of Heyting's idea. Gentzen's approach has been developed as ideas about what it is that makes a piece of reasoning valid. It has resulted in a notion of *valid argument*, of which there are different variants. The differences turn out to be crucial when comparing valid arguments and BHK-proofs. It will be seen that for one variant, the existence of a valid argument can be proved to be extensionally equivalent to the existence of a BHK-proof, while for other variants, attempts at similar proofs break down at different points.

Keywords Proof · Valid argument · Meaning · Semantics · Heyting · Gentzen

1 Introduction

The term "proof-theoretic semantics" was introduced to stand for an approach to meaning based on what it is to have a proof of a sentence. The idea was, at least originally, that in contrast to a truth-conditional meaning theory, one should explain

This is an elaborated version of a talk at the "Second conference on proof-theoretic semantics" at Tübingen in March 2013. Earlier versions have also been presented elsewhere and have been circulated among some colleagues, which has given me the benefit of several comments. I thank especially Per Martin-Löf, Peter Schroeder-Heister and Luca Tranchini for their suggestions, which have stimulated me to prove stronger results and to improve the presentation.

D. Prawitz (✉)
Department of Philosophy, Stochkolm University, Stockholm, Sweden
e-mail: dag.prawitz@philosophy.su.se

T. Piecha and P. Schroeder-Heister (eds.), *Advances in Proof-Theoretic Semantics*,
Trends in Logic 43, DOI 10.1007/978-3-319-22686-6_2

the meaning of a sentence in terms of what it is to *know* that the sentence is true, which in mathematics amounts to having a proof of the sentence.[1]

There are in particular two different concepts of proof that have been used in meaning theories of this kind, but the relation between them has not been paid much attention to. They have their roots in ideas that were put forward by Arend Heyting and Gerhard Gentzen in the first part of the 1930s. Their approaches to meaning are quite different and result in different concepts of proof. Nevertheless there are clear structural similarities between what they require of a proof. The aim of this paper has been to compare the two approaches more precisely, in particular as to whether the existence of proofs comes to the same.

I shall first retell briefly how Heyting and Gentzen formulated their ideas and how others have taken them. In particular, I shall consider how the ideas have been or can be developed so that they become sufficiently precise and general to allow a meaningful comparison, which will then be the object of the second part of the paper.

2 Heyting's Approach to Meaning

A mathematical proposition expresses according to Heyting the intention of a construction that satisfies certain conditions. He explained the assertion of a proposition to mean that the intended construction had been realized, and a proof of a proposition to consist in the realization of the intended construction (Heyting 1930 [5, pp. 958–959], 1931 [6, p. 247], 1934 [7, p. 14]). Thus, according to this explanation, to assert a proposition is equivalent with declaring that there is proof of the proposition. The notion of proof retains in this way its usual epistemic connotation: to have a proof is exactly what one needs in order to be justified in asserting the proposition.

As an important example, Heyting explained the meaning of implication, saying that "$a \supset b$ means the intention of a construction that takes any proof of a to a proof of b".

There are several proposals for how to develop Heyting's ideas more explicitly. One early proposal due to Kreisel (1959, 1962) [10, 11] suggests quite straightforwardly that the constructions intended by implications and universal quantifications are constructive functionals of finite type satisfying the conditions stated by Heyting.[2]

The so-called BHK-interpretation stated by Troelstra and van Dalen (1988) [24], which is less developed ontologically, defines recursively "what forms proofs of

[1] Schroeder-Heister (2006) [22], who coined the term and used it as the title of a conference that he arranged at Tübingen in 1999, writes that proof-theoretic semantics "is based on the fundamental assumption that the central notion in terms of which meanings can be assigned to expressions of our language … is that of *proof* rather than *truth*".

[2] Kreisel was interested in this interpretation as a technical tool for obtaining certain non-derivability results. For a foundation of intuitionistic logic he suggested another interpretation that took a proof of an implication to consist of a pair (α, β) where α is a construction satisfying the condition stated by Heyting and β is a proof of the fact that α satisfies this condition (Kreisel 1962 [12]).

logically compound statements take in terms of the proofs of the constituents".[3] What is here called a proof corresponds rather to what Heyting calls an intended construction, but it has become common in intuitionism to speak about proofs in this way, and I shall follow this way of speaking.

For my purpose here it is sufficient to stay roughly at the level of precision of the BHK-interpretation. I assume that we are given a set $\mathcal{P}$ of proofs of atomic sentences of a first order language and an individual domain D. What it is to be a *proof over* $\mathcal{P}$ of a closed compound sentence A in that language is then defined by recursive clauses like the ones below:

(1) α is a proof over $\mathcal{P}$ of $A \supset B$, if and only if, α is an effective operation such that if β is any proof over $\mathcal{P}$ of A then $\alpha(\beta)$ is a proof over $\mathcal{P}$ of B.
(2) α is a proof over $\mathcal{P}$ of $\forall x A(x)$, if and only if, α is an effective operation such that for any element e in the individual domain D, $\alpha(e)$ is a proof over $\mathcal{P}$ of the instance $A(e)$.

Instead of speaking of proofs of open sentences $A(x)$ under assignments of individuals to variables, I have here assumed for convenience that each element e in the individual domain D has a canonical name, and understand by $A(e)$ the closed sentence obtained by substituting in $A(x)$ this canonical name of e for x. Furthermore, I assume that if α is as stated in clause (2), then there is another effective operation α^*, effectively obtained from α, such that for any closed term t, $\alpha^*(t)$ is a proof of $A(t)$.

To distinguish proofs defined by recursive clauses of this kind, I shall sometimes refer to them as *BHK-proofs*.

3 Gentzen's Approach to Meaning

Gentzen's approach to meaning is commonly described by saying that he had the idea that the meaning of a logical constant is determined by its introduction rule in Natural Deduction, or as he put it himself: "the introductions present, so to speak, the 'definitions' of the symbols concerned" (Gentzen 1934–35 [4, p. 189]). However, this should not be confused with what has later become known as inferentialism, the view that the meaning of a sentence is given by the inference rules concerning the sentence that are in force, which was advocated by Carnap (1934) [1] at about the same time. For Gentzen only some of the inference rules are meaning constitutive, viz. the introduction rules. To indicate their special status, a proof or deduction whose last step is an introduction is now commonly called *canonical* or is said to be in *canonical form*.[4]

[3] BHK stands here for Brouwer-Heyting-Kolmogorov, but there is also another interpretation stated by Troelstra (1977) [23] that is called the BHK-interpretation, where BHK stands for Brouwer-Heyting-Kreisel. It is more akin to Kreisel's second proposal mentioned in footnote 2.

[4] Prawitz (1974) [19]. The term "canonical proof", which was used already by Brouwer in a different context, was applied to normal proofs by Kreisel (1971) [13] and to proofs mentioned in the intuitionistic meaning explanations (such as the BHK-interpretation) by Dummett (1975) [2].

Besides introduction rules there are elimination rules and about them Gentzen says "in an elimination we may use the constant only in the sense afforded to it by the introduction of that symbol". What is intended is clearly that we may use the constant only in this sense, if we are to justify the elimination inference. Gentzen is obviously concerned with what justifies inferences: the introductions stipulate what the logical constants mean, and the eliminations are justified because they are in accord with this meaning.

He clarifies how his ideas are to be understood by giving one example, saying that given an implication $A \supset B$ as premiss, "one can directly infer B when A has been proved, because what $A \supset B$ attests is just the existence of a proof of B from A".[5]

Three important principles can be distinguished here. *Firstly*, what a sentence "attests" is the existence of a canonical proof. An introduction is therefore immediately justified: given proofs of its premisses, the conclusion is warranted, since what the conclusion attests is just that there is a canonical proof of it—the introductions are self-justifying, as one says, when they are taken to be what gives the meanings of the logical constants. Thus, in view of what a sentence attests, a canonical proof is in order, or is valid, provided only that its immediate sub-proofs are.

Secondly, the justification of an elimination consists more precisely in the fact that given that there are proofs of the premisses of the elimination and that the proof of the major premiss is of the kind attested to exist, that is, is in canonical form, a proof of the conclusion can be obtained from these proofs without the use of that elimination. For instance, as Gentzen points out, a proof of the conclusion B of an implication elimination can be obtained from proofs of the premisses if the proof of the major premiss $A \supset B$ is in canonical form, because then there is a proof of B from A, and by replacing the assumption A in that proof by the proof of the minor premiss A, one obtains a proof of B, as is illustrated by the following figure:

$$\dfrac{\dfrac{\begin{array}{c}[A]\\ \vdots\\ B\end{array}}{A \supset B} \qquad \begin{array}{c}\vdots\\ A\end{array}}{B} \qquad \text{gives rise to (is reduced to)} \qquad \begin{array}{c}\vdots\\ [A]\\ \vdots\\ B\end{array}$$

$[A]$ stands for the set of assumptions that are discharged by the exhibited $\supset$-introduction in the first figure and become replaced by the proof of A in the second figure. The operation by which the proof to the left is transformed to the one to the right, that is, substituting in the proof of B from A the proof of the minor premiss A for the occurrences of A that belong to $[A]$, is what is called an $\supset$-*reduction*. These kinds of reductions, which were introduced explicitly in the proof of the normalization theorem for natural deduction (Prawitz 1965 [16]), but which Gentzen was already quite aware of,[6] have in this way a semantic import in being what shows

[5] "kann man … aus einem bewiesenen A sofort B schließ[en]. Denn $A \supset B$ dokumentiert ja das Bestehen einer Herleitung von B aus A" (Gentzen 1934–35 [4, p. 189]).

[6] Although Gentzen never stated these reductions in any published work, it seemed clear already from his example quoted here that he was aware of them. This was later verified when finding an

the eliminations to be justified. By this way of reducing a proof that ends with an elimination to another proof of the same conclusion, the conclusion of the elimination becomes warranted, provided of course that this other proof is valid. Thus, proofs that end with eliminations are valid, if the proofs that they reduce to by applying certain reductions are valid.

Thirdly, when saying that we get a valid proof of B by making the substitution just described, we are tacitly taking for granted that a valid proof from assumptions remains valid when making such substitutions.

We can in this way extract from Gentzen's example three principles about what makes something a valid proof or a *valid deduction*, as I prefer to say (since when the term proof is used, it is normally taken for granted that the reasoning is valid, a convention not strictly adhered to in my informal explanations above). The principles are formulated more precisely below, where I have adopted the terminology that a deduction is *open* when it depends on assumptions and *closed* when all assumptions are discharged or bound.

Principle I. *Introductions preserve validity: a closed deduction in canonical form is valid, if its immediate sub-deductions are.*

Principle II. *Eliminations are justified by reductions: a closed deduction not in canonical form is valid if it reduces to a valid deduction.*

Principle III. *An open deduction is valid, if all results of substituting closed valid deductions for its free (undischarged) assumptions are valid.*

Because of the fact that the premises of an introduction and the assumptions that an introduction may bind are of lower complexity than that of the conclusion, these principles can be taken as clauses of a generalized inductive definition of the notion of valid deduction, relative to a basic clause stating what is counted as valid deductions of atomic sentences. The effect of defining the notion inductively in this way is that no deduction is valid if its validity does not follow from I–III and that the converses of I–III hold true too.

When taking into account also inferences involving quantified sentences, we have to reckon with inferences that bind free individual variables: for instance, an ∀I-inference in which $\forall x\, A(x)$ is inferred from $A(a)$ is said to bind occurrences of the variable a that are free in sentences of the deduction of $A(a)$; the occurrences are said to be bound in the deduction of $\forall x\, A(x)$. A deduction is then said to be *open/closed* if it contains either/neither occurrences of unbound assumptions or/nor occurrences of unbound variables. Accordingly, in principle III the substitution referred to is also to replace all free individual variables by closed individual terms. We then arrive at a notion of validity for natural deductions in general.[7]

(Footnote 6 continued)
unpublished manuscript where Gentzen actually proved a normalization theorem for intuitionistic natural deduction with these reductions (see von Plato 2008 [25]).

[7]What was called "validity based on the introduction rules" by Prawitz (1971) [17] differs from the notion presented here in one substantial respect: in clauses corresponding to principle III, extensions of the set of valid deductions for atomic sentences were considered and it was required that substitutions preserved validity also relative to them; cf. footnote 12.

Gentzen's idea could be summarized by saying that the meaning of a sentence is determined by what counts as a canonical proof of it, which is to say among other things that non-canonical reasoning must be possible to transform to canonical form in order to be acceptable—spelled out in full, the idea is that the meaning of a sentence is determined by what is required from a valid deduction of it. Although this way of formulating Gentzen's ideas goes beyond what he said himself, the three principles of validity formulated here are implicit in the example that he gave, as has been shown above.

Closed valid deductions may be seen as representing proofs, and I shall sometimes refer to them as *Gentzen proofs*.

4 A First Comparison Between Heyting's and Gentzen's Approaches

Both Heyting and Gentzen approached questions of meaning in relation to what it is to prove something, but as seen from the above, their approaches were still very different. Gentzen was concerned with what justifies inferences and thereby with what makes something a valid form of reasoning. These concerns were absent from Heyting's explanations of mathematical propositions and assertions. The constructions that Heyting refers to in his meaning explanations, called proofs in the BHK-interpretation, are mathematical objects, naturally seen as belonging to a hierarchy of effective operations as suggested by Kreisel. They are not proofs built up from inferences. Nor does a proof in Heyting's sense, the realization of an intended construction, constitute a proof built up of inferences, although it does constitute what is required to assert the proposition in question. As was later remarked by Heyting (1958) [8], the steps taken in the realization of the intended construction, in other words, in the construction of the intended object, can be seen as corresponding to inference steps in a proof as traditionally conceived.

These differences between what I am calling BHK-proofs and Gentzen proofs do not rule out the possibility that the existence of such proofs nevertheless comes materially to the same. For instance, a BHK-proof of an implication $A \supset B$ is defined as an operation that takes a BHK-proof of A into one of B, and a closed Gentzen proof of $A \supset B$ affords similarly a construction that takes a Gentzen proof of A into one of B; the latter holds because the validity of a closed deduction of $A \supset B$ guarantees a closed valid deduction in canonical form (by principle II when seen as a clause in an inductive definition) containing a valid deduction of B from the assumption A (principle I), which gives rise to a closed valid deduction of B when a closed valid

(Footnote 7 continued)
In addition to this notion, I also defined a notion of "validity used in proofs of normalizability", similar to Martin-Löf (1971) [14] notion of computability, but as pointed out by Schroeder-Heister (2006) [22], this notion of validity is quite different and should not be counted as a semantic notion explicating Gentzen's idea of meaning, because normalizable deductions are defined outright as computable although (if open) they may not be reducible to canonical form.

deduction of A is substituted for the assumption (principle III). Such similarities may make one expect that one can construct a BHK-proof given a Gentzen proof and vice versa.

However, the ideas of Gentzen discussed above are confined to a specific formal system with particular elimination rules associated with reductions, while there is no comparable restriction of the effective operations that make up a BHK-proof. It is easily seen that for each (valid) deduction in that system there is a corresponding BHK-proof (provided that there are BHK-proofs corresponding to the deductions of atomic sentences), but the converse does not hold. For instance, there is a BHK-proof (over the set of proofs of arithmetical identities) of the conclusion obtained by an application of mathematical induction if there are BHK-proofs of the premisses, but there is no corresponding valid deduction unless we associate a reduction to applications of mathematical induction. If Gentzen proofs are to match BHK-proofs, Gentzen's ideas have first to be generalized, making them free from any particular formal system.

5 Further Development of Gentzen's Ideas

The generalization to be considered in this section will retain Gentzen's ideas of explaining the meaning of sentences in terms of certain canonical forms of reasoning and of connecting the meaning so explained with the justification of inferences. It should be mentioned however that Gentzen's and Heyting's ideas have also been developed in another way, resulting in a certain fusion of their ideas. The explanations in the BHK-interpretation may be enriched by saying à la Gentzen how proofs of sentences of various forms can be constructed. To Gentzen's introduction rules there then correspond canonical ways of forming BHK-proofs of compound sentences from BHK-proofs of the constituents, while to the elimination rules there correspond operations on BHK-proofs to BHK-proofs defined in essentially the same way as the reductions in natural deduction. These correspondences, which further develop the Curry-Howard isomorphism (Howard 1980 [9]), constitute cornerstones of Martin-Löf's type theory (see especially Martin-Löf 1984 [15, p. 24]). In the other direction, I have suggested that a legitimate inference is to be seen as involving not only a transition from assertions to assertions but also an operation on grounds for the premisses that yields a ground for the conclusion, where grounds are BHK-proofs formed in the way just described (Prawitz 2015 [21]).

In this paper, I am not concerned with such fusions of Heyting's and Gentzen's ideas, but want to compare BHK-proofs with forms of reasoning that appear as valid in accordance with Gentzen's ideas about the justification of inferences, sufficiently generalized.

In outline the general idea is this: We consider pieces of reasoning, which will be called *argument structures*, proceeding by arbitrary inferences, and possible *justifications* of these inferences in the form of a set of reductions. An argument structure

paired with a set of reductions is called an *argument*, and we define what it is for an argument to be *valid* by essentially the same three clauses that defined the notion of valid deduction. I shall develop two new notions of validity, called weak and strong validity. They are variants of notions of valid arguments that have been proposed earlier,[8] and will be shown to have distinct features that are especially important when it comes to compare valid arguments and BHK-proofs.

At the end of the paper, I reflect upon the fact that all the variants of valid arguments considered so far deviate in one important respect from the intuitions connected with Gentzen's approach as described above, and point to how the notion of justification may be developed in another way that stays closer to the original ideas.

5.1 Argument Structures

In order to extend the notion of validity defined for deductions so that it can be applied to reasoning in general that proceeds by making arbitrary inferences, I consider tree-formed arrangements of sentences of the kind employed in natural deduction, except that now the inference steps need not be instances of any fixed rules. They will be described by using common terminology from natural deduction, and are what will be called *argument structures*. A sentence standing at the top of the tree is to be seen either as an assumption or as asserted (inferred from no premisses). An occurrence of an assumption can be bound (discharged) by an inference further down in the tree. Indications of which sentences in the tree are assumptions and where they are bound (if they are bound) are to be ingredients of the argument structure.

An inference may also bind occurrences of a free variable (parameter) in sentences above the conclusion. Again it has to be marked how variables are bound by inferences. An argument structure is thus a tree of sentences with indications of these kinds, and can also be seen as a tree-formed arrangement of inferences chained to each other.

The notions of *free assumption* and *free variable*, of *open* and *closed* argument structure, and of a sentence or argument structure *depending on* a free assumption or parameter are carried over to the present context in the obvious way.

There are no restrictions on the argument structures except that an inference may not bind a variable that occurs in an assumption that remains free after the inference, that is, that the conclusion of the inference depends on (otherwise there would be a clash with the idea that an occurrence of a free assumption is free for substitution of closed argument structures, while bound variables are not free for substitution).

[8] In particular, I have in mind my original notion of valid argument (Prawitz 1973 [18]) and the variants proposed by Michael Dummett (1991) [3] and Peter Schroeder-Heister (2006) [22] after profound discussions of my notion. See further footnotes 10 and 12.

An argument structure may for instance look as follows

$$\dfrac{\begin{array}{c}\mathcal{A}_1\\ Nt\end{array} \quad \begin{array}{c}\mathcal{A}_2\\ A(0)\end{array} \quad \begin{array}{c}\overset{(1)}{[A(a)]}\\ \mathcal{A}_3(a)\\ A(s(a))\end{array}}{A(t)}\ (1)\ a$$

where the exhibited inference binds assumptions in the part $\mathcal{A}_3$ of the form $A(a)$ marked (1) as well as variables a that are free in $\mathcal{A}_3$. The inference can be seen as representing an application of mathematical induction, where N stands for 'natural number' and s is the successor operation.

We keep open what forms of sentences are used in an argument structure in order to make the notion sufficiently general. However, when making comparisons with BHK-proofs of sentences in a first order language, we restrict ourselves to such languages. It is assumed that for each form of compound sentences there are associated inferences of a certain kind called introductions, for which we retain the condition from natural deduction that for some measure of complexity, the premisses of the inference and the assumptions bound by the inference are of lower complexity than that of the conclusion. For instance, we could allow the pathological operator *tonk* proposed by Prior and associate it with the introduction rule that he proposed.

We shall say that an argument structure is *canonical* or in *canonical form* if its last inference is an introduction.

5.2 Arguments

The inferences of an argument structure that are not introductions should be justified by reductions as in natural deduction. I shall now be following Schroeder-Heister (2006) [22] partially by taking a *justification* to be simply a set of reductions[9] and a *reduction* to be a pair $(\mathcal{A}_1, \mathcal{A}_2)$ of argument structures such that $\mathcal{A}_1$ is not canonical and $\mathcal{A}_2$ ends with the same sentence as $\mathcal{A}_1$ and depends at most on what $\mathcal{A}_1$ depends on.

An *argument* is a pair $(\mathcal{A}, \mathcal{J})$, where $\mathcal{A}$ is an argument structure and $\mathcal{J}$ is a justification. An argument is said to be closed, open, or canonical (or in canonical form), if the respective attribute is applicable to the argument structure.

$\mathcal{A}_1$ is said to *reduce immediately* to $\mathcal{A}_2$ with respect to $\mathcal{J}$, if $(\mathcal{A}_1, \mathcal{A}_2)$ belongs to $\mathcal{J}$. A *reduction sequence with respect to the justification* $\mathcal{J}$ is a sequence $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n$ $(n \geq 1)$ such that for each $i < n$, either $\mathcal{A}_i$ reduces immediately to $\mathcal{A}_{i+1}$ with respect to $\mathcal{J}$ or $\mathcal{A}_{i+1}$ is obtained from $\mathcal{A}_i$ by replacing an initial part $\mathcal{A}'$ of $\mathcal{A}_i$ by an argument structure $\mathcal{A}''$ such that $\mathcal{A}'$ reduces immediately to $\mathcal{A}''$ with respect to $\mathcal{J}$. An argument structure $\mathcal{A}$ is said to *reduce* to the argument structure $\mathcal{A}^*$ *with respect to the justification* $\mathcal{J}$, if there is a reduction sequence with respect to $\mathcal{J}$ whose first element is $\mathcal{A}$ and last element is $\mathcal{A}^*$.

[9] I have dropped the requirement that the justifications should be closed under substitutions.

Justifications of deductions as described above (Sect. 3) and of argument structures as I originally defined them were effective operations assigned to inference schemata and differ in this respect from the notion that I am now adopting. The main difference is that the relation 'to reduce immediately to' becomes now one-many instead of one-one. The present notion of justification is of particular interest when we come to comparing valid arguments with BHK-proofs,[10] but as we shall see it has some unwanted consequences.

Schroeder-Heister remarks that to take justifications to be relations corresponds to the idea that there can be "alternative justifications" of the same argument structure. I think that this idea is somewhat doubtful; anyway, as we shall soon see, it can be taken in many ways.

Since a justification is just a set of reductions, it may not "really" justify the argument structure. We could say that what is called a justification is merely a proposed or possible justification, a justification candidate. What is required of a "real" justification gets expressed by the definition of what it is for an argument to be valid.

For instance, one can invent a justification of an argument structure using Prior's elimination rule for *tonk* by assigning some reductions to applications of the rule, but this will never give rise to valid arguments that make creative uses of Prior's rule.

An important example of justifications outside the standard ones for the elimination rules in natural deduction is one that can be associated with the argument structures exhibited in the preceding subsection as representing applications of mathematical induction. It consists of a pair $(\mathcal{B}_1, \mathcal{B}_2)$ where thus $\mathcal{B}_1$ is an argument structure of this form. What $\mathcal{B}_2$ is depends on the form of the first premiss of the last inference, Nt, which may be called the major premiss of the inference. If the major premiss has the form $N0$ and the conclusion accordingly has the form $A(0)$, $\mathcal{B}_2$ is to be $\mathcal{A}_2$, the argument structure for $A(0)$ that represents the induction base. If it has the form $Ns(t)$ and stands as conclusion of an inference whose premiss is Nt, the conclusion accordingly having the form $A(s(t))$, $\mathcal{B}_2$ is to be argument structure

$$\dfrac{\begin{array}{c}\mathcal{A}_1\\ Nt\end{array} \quad \begin{array}{c}\mathcal{A}_2\\ A(0)\end{array} \quad \begin{array}{c}(1)\\ [A(a)]\\ \mathcal{A}_3(a)\\ A(s(a))\end{array}}{\begin{array}{c}[A(t)]\\ \mathcal{A}_3(t)\\ A(s(t))\end{array}}\ (1)\ a$$

[10]It also offers one way to avoid a problem connected with my earlier definition of justification. When generalizing principle III in the definition of valid deduction to argument structures, the substitutions of valid argument structures $(\mathcal{A}_i, \mathcal{J}_i)$ in open arguments $(\mathcal{A}, \mathcal{J})$ had to be restricted to ones where $\mathcal{J}_i$ was consistent with $\mathcal{J}$. The restriction is unwanted and may make the notion of validity too weak (essential also when comparing with BHK-proofs). A possible alternative in order to avoid this problem while keeping the relation 'to reduce immediately to' one-one is to take reductions to be assignments of effective operations to occurrences of inferences (instead of inference schemata or inferences).

If the term t is a numeral n, the argument structure is finally transformed by successive reductions of this kind to an argument structure consisting of the induction base $\mathcal{A}_2$ followed by n applications $\mathcal{A}_3(0), \mathcal{A}_3(\underline{1}), \ldots, \mathcal{A}_3(\underline{n-1})$ of the induction step on top of each other. These reductions represent indeed the natural and commonly given justification for inferences by mathematical induction.

5.3 Validity of Arguments

We can now define what it is for an argument to be valid by adopting three principles analogous to the ones stated for valid deductions:

I. *A closed canonical argument* $(\mathcal{A}, \mathcal{J})$ *is valid, if for each immediate sub-argument structure* $\mathcal{A}^*$ of $\mathcal{A}$, *it holds that* $(\mathcal{A}^*, \mathcal{J})$ *is valid.*
II. *A closed non-canonical argument* $(\mathcal{A}, \mathcal{J})$ *is valid, if* $\mathcal{A}$ *reduces relative to* $\mathcal{J}$ *to an argument structure* $\mathcal{A}^*$ *such that* $(\mathcal{A}^*, \mathcal{J})$ *is valid.*
III. *An open argument* $(\mathcal{A}, \mathcal{J})$ *depending on the assumptions* $A_1, A_2, \ldots, A_n$ *is valid, if all its substitution instances* $(\mathcal{A}^*, \mathcal{J}^*)$ *are valid, where* $\mathcal{A}^*$ *is obtained by first substituting any closed terms for free variables in sentences of* $\mathcal{A}$, *resulting in an argument structure* $\mathcal{A}^\circ$ *depending on the assumptions* $A_1^\circ, A_2^\circ, \ldots, A_n^\circ$, *and then for any valid closed argument structures* $(\mathcal{A}_i, \mathcal{J}_i)$ *for* A_i°, $i \leq n$, *substituting* $\mathcal{A}_i$ *for* A_i° in $\mathcal{A}^\circ$, *and where* $\mathcal{J}^* = \bigcup_{i \leq n} \mathcal{J}_i \cup \mathcal{J}$.

Because of the assumed condition on the relative complexity of the ingredients of an introduction inference, the principles I-III can again be taken as clauses of a generalized inductive definition of the notion of valid argument *relative to a base* $\mathcal{B}$, which is to consist of a set of closed argument structures containing only atomic sentences. If $\mathcal{A}$ is an argument structure of $\mathcal{B}$, the argument $(\mathcal{A}, \emptyset)$, where $\emptyset$ is the empty justification, is counted as canonical and outright as valid relative to $\mathcal{B}$. A base is seen as determining the meanings of the atomic sentences. An argument that is valid relative to any base can be said to be logically valid.

If $\mathcal{A}$ is an argument structure representing mathematical induction as exhibited in Sect. 5.1, $\mathcal{J}$ is the justification associated with $\mathcal{A}$ as described in Sect. 5.2, and $\mathcal{B}$ is a base for arithmetic, say corresponding to Peano's first four axioms and the recursion schemata for addition and multiplication, then the argument $(\mathcal{A}, \mathcal{J})$ is valid relative to $\mathcal{B}$ (as was in effect first noted in a different conceptual framework by Martin-Löf (1971) [14]. This is an example of a valid argument that is not logically valid but whose validity depends on the chosen base. However, I shall often leave implicit the relativization of validity to a base.

Instead of saying that the argument $(\mathcal{A}, \mathcal{J})$ is valid it is sometimes convenient to say that the argument structure $\mathcal{A}$ is valid with respect to the justification $\mathcal{J}$. But it is argument structures paired with justifications that correspond to proofs and that will be compared to BHK-proofs.

6 Weak and Strong Validity and Their Features

6.1. As is easily seen, it comes to the same if we in clause II of the definition of validity require instead that $\mathcal{A}$ reduces relative to $\mathcal{J}$ to a canonical argument structure $\mathcal{A}^*$ that is valid with respect to $\mathcal{J}$.

An important question concerning valid arguments, especially crucial when comparing them with BHK-proofs, is whether this canonical argument $\mathcal{A}^*$ required by clause II can be found effectively.

6.1.1. If the definition of validity is read constructively, or in other words, if the existential quantifier in clause II is understood intuitionistically, the answer is of course yes, the canonical argument $\mathcal{A}^*$ can be found effectively. If so, there is also an effective operation denoted by * that is defined for every valid closed argument $(\mathcal{A}, \mathcal{J})$ and yields a canonical argument structure $\mathcal{A}^*$ such that $\mathcal{A}$ reduces to $\mathcal{A}^*$ with respect to $\mathcal{J}$ and $(\mathcal{A}^*, \mathcal{J})$ is valid.

6.1.2. Otherwise, if the definition is not taken in a constructive sense, it is not guaranteed that $\mathcal{A}^*$ can be found effectively. Even if we require of a justification that it should be possible to generate its reductions effectively, it is still not guaranteed that $\mathcal{A}^*$ can be found effectively. It is true that when we are generating the reduction sequences with respect to a justification $\mathcal{J}$ that start from a closed non-canonical argument structure $\mathcal{A}$ that is valid with respect to $\mathcal{J}$, we sooner or later hit upon a canonical argument structure $\mathcal{A}^*$ such that $(\mathcal{A}^*, \mathcal{J})$ is valid. But since validity is not a decidable property, we may not be able to tell which one(s) of the canonical structures $\mathcal{A}^*$ that we reach in this way is (are) the right one(s).

6.2. The situation was quite different when we were dealing with valid deductions based on the standard reductions in natural deduction. Given a closed valid deduction $\mathcal{A}$, a valid canonical deduction $\mathcal{A}^*$ as required by principle II can always be found effectively because of two facts: firstly, as already noted, the justifications consist of effective operations, which means that a deduction reduces immediately to at most one other deduction; and secondly, it can be shown that, regardless of the order in which the operations are applied, they will transform a closed deduction to a valid canonical one. This second feature can be called *strong* validity,[11] in analogy with how in proof theory one says that a natural deduction is strongly normalizable if all reduction sequences terminate in a normal deduction.

Similarly, we can speak of strong validity of arguments when the canonical argument $\mathcal{A}^*$ is found regardless of the order in which the reductions are taken and regardless of which reductions in $\mathcal{J}$ are employed. More precisely, a definition of an argument structure $(\mathcal{A}, \mathcal{J})$ being *strongly valid* (relative to a base $\mathcal{B}$ whose argument structures are now counted outright as strongly valid) is obtained by clauses I*-III*, where I* and III* are like I and III except that "valid" is replaced with "strongly valid" and the second clause reads:

[11] I have previously used the expression strongly valid for deductions and arguments in another way where it would be better to say strongly computable—cf. second part of footnote 7.

II* *A closed non-canonical argument* $(\mathcal{A}, \mathcal{J})$ *is strongly valid, if each reduction sequence relative to* $\mathcal{J}$ *starting from* $\mathcal{A}$ *can be prolonged to a reduction sequence that contains a canonical argument structure* $\mathcal{A}^*$ such that $(\mathcal{A}^*, \mathcal{J})$ *is strongly valid.*

Henceforth, I shall refer to the notion of validity defined by I–III as *weak validity*.

6.3. Effectiveness is restored when going from weak validity to strong validity, in spite of the justification still being a relation instead of a set of operations, provided that we require that its reductions can be generated effectively. When we generate in some arbitrarily chosen order the reduction sequences with respect to $\mathcal{J}$ that start from a closed argument structure $\mathcal{A}$ that is a strongly valid with respect to $\mathcal{J}$, the first canonical argument $\mathcal{A}^*$ that we find is guaranteed to be strongly valid with respect to $\mathcal{J}$; to verify this fact, note that reductions obviously preserve strong validity: if $\mathcal{A}$ reduces to $\mathcal{A}^*$ with respect to $\mathcal{J}$ and $(\mathcal{A}, \mathcal{J})$ is strongly valid, then so is $(\mathcal{A}^*, \mathcal{J})$.

6.3.1. That effectiveness is obtained can be seen as an aspect of the fact that strong validity requires all so-called "alternative justifications" to be "real" justifications, so to say—if a closed argument $(\mathcal{A}, \mathcal{J})$ is strongly valid and the reductions $(\mathcal{A}_1, \mathcal{A}_2)$ and $(\mathcal{A}_1, \mathcal{A}_3)$ both belong to $\mathcal{J}$, clause II* requires that regardless of which one is used in a reduction sequence, it takes $\mathcal{A}$ a step towards a valid canonical argument. Clause II, in contrast, only requires that one of the reductions does so, which means that the other reduction may lead astray and may have no significance for the validity of the argument in question.

6.3.2. An aspect of the last feature is that weak validity is obviously monotone with respect to justifications: if $(\mathcal{A}, \mathcal{J})$ is weakly valid and $\mathcal{J} \subseteq \mathcal{J}^*$, then $(\mathcal{A}^*, \mathcal{J}^*)$ is weakly valid too—whatever reductions we add to $\mathcal{J}$, the argument remains weakly valid. In contrast, strong validity is not monotone with respect to justifications—added "alternative justifications" must be "real", if validity is to be preserved.

6.3.3. Yet another aspect of essentially the same feature is that the property of an argument structure to be weakly valid with respect to some justification $\mathcal{J}$ is indeed a very weak property. In fact, there is a justification $\mathcal{J}$ for a given language such that any non-canonical argument structure $\mathcal{A}$ for a sentence A in that language is weakly valid with respect to $\mathcal{J}$, provided only that there exists a weakly valid closed argument $(\mathcal{A}^*, \mathcal{J}^*)$ in that language for $\mathcal{A}$. We can simply choose as $\mathcal{J}$ the universal set of reductions in that language, call it $\mathcal{UR}$. Since $\mathcal{J}^* \subseteq \mathcal{UR}$, the argument $(\mathcal{A}^*, \mathcal{UR})$ is weakly valid by the monotonicity of weak validity, and since $\mathcal{A}$ reduces to $\mathcal{A}^*$ with respect to $\mathcal{UR}$, $(\mathcal{A}, \mathcal{J})$ is weakly valid too in virtue of clause II.

It must be said that this argument $(\mathcal{A}, \mathcal{UR})$ may be quite far from an intuitively valid argument for A—the inferences in $\mathcal{A}$ may lack any significance for the validity of the argument, and the only relevant property of $\mathcal{UR}$ for the validity is that the reduction $(\mathcal{A}, \mathcal{A}^*)$ is an element and that $\mathcal{J}^*$ is included.

6.4. It should be noted that strong validity does not entail weak validity; a strongly valid argument for an implication $A \supset B$ is also weakly valid if A does not contain implication, but as soon as implication becomes nested in the antecedent, this may cease to hold because of the third clause in the definitions of validity.

The features discussed here of the two variants of validity are essential when we are to compare the valid argument with the BHK-proofs, as will be seen in the next section.[12]

7 Mappings of Valid Arguments on BHK-Proofs and Vice Versa

After having now made Gentzen's approach free from ties to a specific formal system, we return to the question whether the two approaches come to the same thing extensionally. Let us assume that $\mathcal{P}$ is a set of BHK-proofs of atomic sentences, that $\mathcal{B}$ is a base of valid arguments for atomic sentences, and that they have been mapped on each other. We shall try to extend these mappings to compound sentences.

In other words, we shall try to define one mapping called

$\mathcal{P}roof$ which applied to a valid closed argument relative to $\mathcal{B}$ for a sentence A gives as value a BHK-proof of A over $\mathcal{P}$,

and a mapping in the other direction called

$\mathcal{A}rg$ which applied to a BHK-proof over $\mathcal{P}$ of a sentence A gives as value a valid closed argument relative to $\mathcal{B}$ for A,

assuming as an induction assumption that we have been able to define such effective mappings for all sentences of complexity less than that of A.

If α is a BHK-proof of a sentence A, $\mathcal{A}rg(\alpha)$ has to be a pair, which will be written $(\mathcal{A}rg_1(\alpha), \mathcal{A}rg_2(\alpha))$; thus, $\mathcal{A}rg_1(\alpha)$ is an argument structure for A and $\mathcal{A}rg_2(\alpha)$ is a justification.

I restrict myself to the cases when A is an implication or a universal quantification, and shall consider in parallel the problems that arise for different variants of validity of arguments.

---

[12] As to the other variants of validity mentioned in footnote 8, Dummett defines validity directly for argument structures, thus leaving the justifications implicit. I have commented on this difference elsewhere [20], but then overlooked one important consequence of it, which is now taken up in footnote 14.

Schroeder-Heister's notion of validity differs from weak validity as defined here by following my previous definition of valid deduction as regards extensions of the given base $\mathcal{B}$ (see footnote 7). We get this notion by requiring in clause III that also for every extension $\mathcal{B}^*$ of $\mathcal{B}$, it holds that all substitution instances $(\mathcal{A}^*, \mathcal{J}^*)$ are valid relative to $\mathcal{B}^*$ where $\mathcal{A}^*$ is obtained by substituting for $\mathcal{A}_i^\circ$ a closed argument structure $\mathcal{A}_i$ such that $(\mathcal{A}_i, \mathcal{J}_i)$ is valid relative to $\mathcal{B}^*$.

To consider extensions of the given base in this way is natural when a base is seen as representing a state of knowledge, but is in conflict with the view adopted here that a base is to be understood as giving the meanings of the atomic sentences. For instance, the argument representing reasoning by mathematical induction presented in Sect. 5.3 ceases to be valid relative to the arithmetical base $\mathcal{B}$ if we require in clause III that validity be monotone with respect to the base.

Concerning my original notion of validity see remarks made in the text and in footnote 10.

7.1 Extending the Mapping $\mathcal{P}roof$ to Arguments for A

7.1.1. Consider first the case when A is an implication $B \supset C$. $\mathcal{P}roof$ is then to be defined for any valid closed argument $(\mathcal{A}, \mathcal{J})$ for A, which is done by saying that $\mathcal{P}roof(\mathcal{A}, \mathcal{J})$ is to be the operation α defined for BHK-proofs β of B such that

$$\alpha(\beta) = \mathcal{P}roof\begin{pmatrix} \mathcal{A}rg_1(\beta) & \\ [B] & \\ \mathcal{A}^\circ & , \ \mathcal{J} \cup \mathcal{A}rg_2(\beta) \\ C & \end{pmatrix} \tag{a}$$

I have to explain what operation $\mathcal{A}^\circ$ is and show—under the assumptions that $(\mathcal{A}, \mathcal{J})$ is a valid closed argument for A and that β is a BHK-proof of B and the induction assumption—that:

(i) the operation $\mathcal{A}^\circ$ is an effective procedure for finding an argument structure for C, and
(ii) the pair to which $\mathcal{P}roof$ is applied above in (a) is effectively obtained from $(\mathcal{A}, \mathcal{J})$ and β, and is a valid closed argument for C.

It then follows by the induction assumption that $\mathcal{P}roof$ is defined for this argument and that $\alpha(\beta)$ as defined in (a) is a BHK-proof of C, which means that the operation α is a BHK-proof of A.

If $\mathcal{A}$ is in canonical form, that is, has the form

$$\dfrac{\begin{matrix}(1)\\ [B]\\ \mathcal{A}'\\ C\end{matrix}}{B \supset C}\ (1)$$

we let $\mathcal{A}^\circ$ be the immediate sub-structure $\mathcal{A}'$ of $\mathcal{A}$, which is an argument structure for C.[13]

If $\mathcal{A}$ is not in canonical form, we want $\mathcal{A}^\circ$ to be the immediate sub-structure of a closed canonical argument structure $\mathcal{A}^*$ to which $\mathcal{A}$ reduces with respect to $\mathcal{J}$ and that is valid with respect to $\mathcal{J}$. Now it becomes important what kind of validity we are dealing with. If the argument $(\mathcal{A}, \mathcal{J})$ is strongly valid, then as noted in Sect. 6.3, there is an effective procedure for finding such an argument structure $\mathcal{A}^*$ that is strongly valid with respect to $\mathcal{J}$: Generating the reduction sequences with respect to $\mathcal{J}$ that

[13]Note the difference between the two notations below, commonly used in natural deduction:

$$\begin{matrix}\mathcal{A}\\ A\end{matrix} \qquad \dfrac{\mathcal{A}}{A}$$

The left one stands for the same argument structure as $\mathcal{A}$ and is used to indicate that the last sentence of $\mathcal{A}$ is A. The right one stands for an argument structure formed by putting A under the structure $\mathcal{A}$ (and it is left open what the last sentence of $\mathcal{A}$ is).

start from $\mathcal{A}$ in some arbitrarily chosen order, we take the first canonical argument structure $\mathcal{A}^*$ that we find. We then let $\mathcal{A}^\circ$ be its immediate sub-structure; that is, $\mathcal{A}^\circ$ is again $\mathcal{A}'$ if $\mathcal{A}^*$ has the form shown above.

Note that if $(\mathcal{A}, \mathcal{J})$ is weakly valid, the procedure described above may result in an argument structure such that $\mathcal{A}^*$ is not weakly valid with respect to $\mathcal{J}$, and that if $(\mathcal{A}, \mathcal{J})$ is neither strongly nor weakly valid, the procedure may not give any result at all. But when $(\mathcal{A}, \mathcal{J})$ is strongly valid and closed, the operation $\mathcal{A}^*$ is defined and is an effective procedure. Hence, $\mathcal{A}^\circ$ is an effective procedure for finding an argument structure for C.

If $(\mathcal{A}, \mathcal{J})$ is weakly valid and this is taken in a constructive sense, then as already noted (Sect. 6.1.1), there is an effective procedure * defined for all weakly valid closed arguments $(\mathcal{A}, \mathcal{J})$ which yields an argument structure $\mathcal{A}^*$ such that $\mathcal{A}$ reduces to $\mathcal{A}^*$ with respect to $\mathcal{J}$ and $\mathcal{A}^*$ is weakly valid with respect to $\mathcal{J}$. Letting $\mathcal{A}^\circ$ be the immediate substructure of $\mathcal{A}^*$, we have again explained the operation $\mathcal{A}^\circ$ as an effective procedure for finding an argument structure for C.

Task (i) has thus been carried out for strong validity and for weak validity read constructively, but not for weak validity read non-constructively. In the two successful cases, task (ii) is now easy. That the pair to which $\mathcal{P}roof$ is applied in (a) is effectively obtained follows from the induction assumption and the effectiveness of the operation $\mathcal{A}^\circ$. The demonstration of the fact that the pair is a strongly or weakly valid closed argument for C follows the same pattern for the two cases of validity, so we may let valid mean either weakly or strongly valid: That $(\mathcal{A}^\circ, \mathcal{J})$ is a valid argument for C follows from the validity of $(\mathcal{A}, \mathcal{J})$ or of $(\mathcal{A}^*, \mathcal{J})$, as the case may be. By the induction assumption $(\mathcal{A}rg_1(\beta), \mathcal{A}rg_2(\beta))$ is a valid argument for B, and from these two facts it follows by clause III* or III that the argument to which $\mathcal{P}roof$ is applied in (a) is a closed valid argument for C, as was to be shown.

7.1.2. Let now A be the sentence $\forall x\, B(x)$, and let $(\mathcal{A}, \mathcal{J})$ be a closed argument for $\forall x\, B(x)$ that is strongly valid or is weakly valid taken in a constructive sense.

As in Sect. 2, it is assumed that the elements in the individual domain D have canonical names. I apply the conventions explained there, and define $\mathcal{P}roof(\mathcal{A}, \mathcal{J})$ to be the operation α defined for the elements e in the individual domain D such that

$$\alpha(e) = \mathcal{P}roof(\mathcal{A}^\circ(e), \mathcal{J}) \tag{b}$$

The operation $\mathcal{A}^\circ$ is explained analogously to how it was explained in the preceding case. Thus, if $\mathcal{A}$ is in canonical form, $\mathcal{A}$ has the form

$$\begin{array}{c} \mathcal{A}'(a) \\ B(a) \\ \hline \forall x\, B(x) \end{array}$$

and we let $\mathcal{A}^\circ$ be $\mathcal{A}'(a)$, the immediate sub-structure of $\mathcal{A}$. $\mathcal{A}^\circ(e)$ is then the result of substituting for a in $\mathcal{A}^\circ(a)$ the canonical name for e.

If $\mathcal{A}$ is not in canonical form, we find effectively as in the preceding case a closed canonical argument structure $\mathcal{A}^*$ to which $\mathcal{A}$ reduces with respect to $\mathcal{J}$ such that $(\mathcal{A}^*, \mathcal{J})$ has the same kind of validity as $(\mathcal{A}, \mathcal{J})$. We let then $\mathcal{A}^\circ(a)$ be the immediate substructure of $\mathcal{A}^*$ and $\mathcal{A}^\circ(e)$ the result of substituting for a in $\mathcal{A}^\circ(a)$ the canonical name for e.

Since by clauses I* and III* or by clauses I and III $(\mathcal{A}^*(e), \mathcal{J})$ is a closed valid argument for $B(e)$, validity taken in one of the two forms here considered, it follows by the induction assumption in question, that *Proof* is defined for this argument and that $\alpha(e)$ as defined in (b) is a BHK-proof of $B(e)$. Thus, α is a BHK-proof of A.

7.2 Extending the Mapping $\mathcal{A}rg$ to BHK-Proofs of A

7.2.1. Now I first consider the easiest case when A is a universal sentence $\forall x\, B(x)$. Let α be a BHK-proof of A. I define $\mathcal{A}rg(\alpha) = (\mathcal{A}rg_1(\alpha), \mathcal{A}rg_2(\alpha))$ as follows:

$$\mathcal{A}rg_1(\alpha) = \frac{\overline{B(a)}}{\forall x\, B(x)} \qquad \mathcal{A}rg_2(\alpha) = \bigcup_t \mathcal{A}rg_2(\alpha^*(t)) \cup \{(\overline{B(t)}, \mathcal{A}rg_1(\alpha^*(t)))\}_t$$

The line above the top sentence $B(a)$ in the argument structure that $\mathcal{A}rg_1(\alpha)$ assumes as value is meant to indicate that $B(a)$ is not an assumption but is inferred from zero premisses; thus, the parameter a does not occur in any assumption that the sentence at the bottom depends on, and it becomes therefore bound by the $\forall$I-inference as usual.

For the argument structure $\mathcal{A}rg_1(\alpha)$ to be valid with respect to a justification $\mathcal{J}$, it is necessary and sufficient that $\mathcal{J}$ contains a reduction such that any instance of the argument structure $\overline{B(a)}$ reduces with respect to $\mathcal{J}$ to an argument structure $\mathcal{A}$ that is valid with respect to $\mathcal{J}$. The problem is that it is not sufficient to find, for each closed term t, appropriate reductions for $\overline{B(t)}$.[14] Instead we must find a set $\mathcal{J}$ of reductions such that it can be shown that, for each term t, $\mathcal{J}$ contains appropriate reductions. I succeed in showing this only for the case of weak validity. The set $\mathcal{A}rg_2(\alpha)$ defined above will be shown to be such a justification in that case. The same result could be obtained more easily by choosing the universal set of reductions for the language in question, but it may be of some interest to see that this smaller set will do.

For the understanding of the definition of $\mathcal{A}rg_2(\alpha)$, recall that α^* is the effective operation assumed in Sect. 2 to be possible to obtain effectively from α such that for each closed term t, $\alpha^*(t)$ is a BHK-proof of $B(t)$. I also want to make clear that $\mathcal{A}rg_2(\alpha)$ is the union of two sets (i) and (ii) where (i) is the union of all sets $\mathcal{A}rg_2(\alpha^*(t))$ for closed terms t and (ii) is the set of all pairs $(\overline{B(t)}, \mathcal{A}rg_1(\alpha^*(t)))$ where t is a closed term. By the induction assumption, $\mathcal{A}rg_1(\alpha^*(t))$ and $\mathcal{A}rg_2(\alpha^*(t))$ are both defined.

[14] This is all that is required by Dummett's notion of valid argument structure, which means that his notion is quite obviously extensionally equivalent to the notion of BHK-proof.

In order to show that $(\mathcal{A}rg_1(\alpha), \mathcal{A}rg_2(\alpha))$ is a weakly valid argument for $\forall x\, B(x)$, we have to show in view of principles I and III and since $\mathcal{A}rg_1(\alpha)$ is a closed argument structure for $\forall x\, B(x)$ in canonical form that the argument $(\overline{B(t)}, \mathcal{A}rg_2(\alpha))$ is weakly valid for each closed term t. To this end we must show in view of principle II that $\overline{B(t)}$ for each closed term t reduces with respect to $\mathcal{A}rg_2(\alpha)$ to an argument structure $\mathcal{A}$ such that $(\mathcal{A}, \mathcal{A}rg_2(\alpha))$ is weakly valid.

We shall now verify that for each closed term t, $\mathcal{A}rg_1(\alpha^*(t))$ is such an argument structure $\mathcal{A}$. Firstly note that it has been arranged so that $\overline{B(t)}$ reduces to $\mathcal{A}rg_1(\alpha^*(t))$ with respect to $\mathcal{A}rg_2(\alpha)$ for each closed term t by the defining $\mathcal{A}rg_2(\alpha)$ as a union of two sets (i) and (ii) where (ii) is the set of all pairs $(\overline{B(t)}, \mathcal{A}rg_1(\alpha^*(t)))$ for closed terms t. Secondly, note that by the induction assumption, for each closed term t, $(\mathcal{A}rg_1(\alpha^*(t)), \mathcal{A}rg_2(\alpha^*(t)))$ is a closed weakly valid argument for $B(t)$, since $\alpha^*(t)$ is a BHK-proof of $B(t)$. Thirdly, we recognize that from the last fact follows the wanted result that $(\mathcal{A}rg_1(\alpha^*(t)), \mathcal{A}rg_2(\alpha))$ is weakly valid, because $\mathcal{A}rg_2(\alpha^*(t))$ is a subset of $\mathcal{A}rg_2(\alpha)$ (in virtue of being a subset of the set (i) described above) and weak validity is monotone with respect to justifications, as remarked in Sect. 6.3.2.

As seen the monotonicity of weak validity with respect to justifications is used in establishing this mapping, and therefore a similar demonstration does not go through for strong validity, not being monotone with respect to justifications.

7.2.2. Let now A be an implication $B \supset C$ and let α be a BHK-proof of $B \supset C$. The construction of $\mathcal{A}rg(\alpha)$ is similar to the preceding case. Clearly, $\mathcal{A}rg_1(\alpha)$ is to be the canonical argument structure

$$\dfrac{\dfrac{\overset{(1)}{[B]}}{C}}{B \supset C}\ (1)$$

It is weakly valid with respect to $\mathcal{A}rg_2(\alpha)$, if and only if, for each weakly valid, closed argument $(\mathcal{A}, \mathcal{J})$ for B, the argument structure

$$\dfrac{\begin{array}{c}\mathcal{A}\\ B\end{array}}{C} \qquad \text{(c)}$$

reduces with respect to $\mathcal{J} \cup \mathcal{A}rg_2(\alpha)$ to an argument structure $\mathcal{A}^\circ$ such that $(\mathcal{A}^\circ, \mathcal{J} \cup \mathcal{A}rg_2(\alpha))$ is weakly valid (as is seen by applying clauses I, III, and II in this order). To guarantee that there is such an $\mathcal{A}^\circ$ for each weakly valid closed argument $(\mathcal{A}, \mathcal{J})$, I define

$$\begin{aligned}\mathcal{A}rg_2(\alpha) = \bigcup\{&\mathcal{A}rg_2(\alpha(\mathcal{P}roof(\mathcal{A}, \mathcal{J}))) : (\mathcal{A}, \mathcal{J}) \text{ is a weakly valid closed argument for } B\}\\ &\cup \{(\mathcal{A}^*, \mathcal{A}^{**}) : \text{ there is weakly valid closed argument } (\mathcal{A}, \mathcal{J}) \text{ for } B\\ &\quad \text{such that } \mathcal{A}^* \text{ is of the form (c) and } \mathcal{A}^{**} \text{ is } \mathcal{A}rg_1(\alpha(\mathcal{P}roof(\mathcal{A}, \mathcal{J})))\}\end{aligned}$$

Assume now that $(\mathcal{A}, \mathcal{J})$ is a closed argument for B that is weakly valid. We shall verify that $\mathcal{A}rg_1(\alpha(\mathcal{P}roof(\mathcal{A}, \mathcal{J})))$ is the wanted $\mathcal{A}^\circ$. Firstly, note that the argument structure (c) reduces with respect to $\mathcal{J} \cup \mathcal{A}rg_2(\alpha)$ to $\mathcal{A}rg_1(\alpha(\mathcal{P}roof(\mathcal{A}, \mathcal{J})))$ in virtue of the fact that the pair ((c), $\mathcal{A}rg_1(\alpha(\mathcal{P}roof(\mathcal{A}, \mathcal{J}))))$ is a member of the second set in the union that by definition constitutes $\mathcal{A}rg_2(\alpha)$. Secondly, we note that by the induction assumption, $\mathcal{P}roof(\mathcal{A}, \mathcal{J})$ is a BHK-proof of B. Hence $\alpha(\mathcal{P}roof(\mathcal{A}, \mathcal{J}))$ is a BHK-proof of C. Therefore, by the induction assumption in the other direction,

$$(\mathcal{A}rg_1(\alpha(\mathcal{P}roof(\mathcal{A}, \mathcal{J}))), \mathcal{A}rg_2(\alpha(\mathcal{P}roof(\mathcal{A}, \mathcal{J})))) \tag{d}$$

is a weakly valid argument for A. Thirdly, we recognize that from the weak validity of the argument (d) follows the wanted result that the argument $(\mathcal{A}rg_1(\alpha(\mathcal{P}roof(\mathcal{A}, \mathcal{J}))), \mathcal{J} \cup \mathcal{A}rg_2(\alpha))$ is weakly valid, because $\mathcal{A}rg_2(\varphi(\mathcal{P}roof(\mathcal{A}, \mathcal{J})))$ is a subset of $\mathcal{A}rg_2(\varphi)$ (in virtue of being a subset of the first set of the union that constitutes $\mathcal{A}rg_2(\varphi)$ by definition) and weak validity is monotone with respect to justifications.

The demonstrations in 7.2.1 and 7.2.2 have been entirely constructive and thus show that the result that $\mathcal{A}rg(\alpha)$ is a closed weakly valid argument for A when α is a BHK-proof of A holds even when the notion of weak validity is understood constructively.

8 Concluding Remarks

8.1. We have thus shown that the notion of a weak valid argument taken constructively is extensionally equivalent with the notion of a BHK-proof.

When weak validity is taken non-constructively, I have not been able to construct a BHK-proof of A from a weakly valid argument for A, but only in the other direction a weakly valid argument for A from a BHK-proof of A, given the induction assumption.

In contrast, from a strongly valid argument for A, I have constructed a BHK-proof of A, given the induction assumption and the assumption that the reductions can be generated effectively, but have not been able to construct in the other direction a strongly valid argument for A from a BHK-proof of A.

Since the mentioned constructions depend on the assumption that there are mappings in both directions for sub-sentences, nothing has been established about the relations between on the one hand BHK-proofs and on the other hand arguments that are weakly valid understood in a non-constructive sense or are strongly valid.

8.2. As has been seen above, when the notion of valid deduction is generalized to the notion of valid argument, the justifications come to play the major role and the inferences of the argument structures a correspondingly minor role. Some of the intuitions behind the notion of valid deduction are lost in this way. It would therefore be interesting to investigate a more restricted notion of reductions than the one used here in connection with arguments.

The standard reductions in natural deduction are all transformations of a given deduction by two kinds of very simple effective operations, possibly combined with each other. One kind consists of operations φ such that $\varphi(\mathcal{D})$ is a sub-deduction of $\mathcal{D}$. The other kind consists of operations φ such that $\varphi(\mathcal{D})$ is the result of substituting in $\mathcal{D}$ an individual term occurring in a sentence of $\mathcal{D}$ for a free variable occurring in a sentence of $\mathcal{D}$ or substituting in a sub-deduction of $\mathcal{D}$ for a free assumption (in that sub-deduction) another sub-deduction of $\mathcal{D}$. Also the reduction associated with mathematical induction (Sect. 5.2) is a transformation built up of these two kinds of operations.

By applying operations of these two kinds to a deduction or an argument structure one obtains an argument structure that is contained in the given deduction or argument structure; in case substitutions have been carried out, we should perhaps say that the result is implicitly contained. A reduction of this kind associated to an inference constitutes a justification of the inference in a much stronger sense than the reductions that have been considered in connection with argument structures: Given that the arguments for the premisses are acceptable, there is an acceptable argument for the conclusion, because an argument for the conclusion is already contained, at least implicitly, in the arguments for the premisses taken together. This is actually the kind of justification of Gentzen's elimination rules that I have labelled the inversion principle, using a term from Lorenzen, and have presented as the intuition behind the normalization theorem for natural deductions [16].

An argument structure that is valid with respect to a justification that assigns such operations to occurrences of inferences would in itself have an epistemic force. Perhaps one could say that the function of the justifications would then be to verify that they have such a force, whereas valid arguments as they have been defined here often get their entire epistemic force from the justifications.

A notion of valid argument based on justifications of this kind would be a quite different concept from the variants of valid argument that have been dealt with in this paper. It would also be different from the notion of BHK-proof, it seems.

References

1. Carnap, R.: Logische Syntax der Sprache. Springer, Wien (1934)
2. Dummett, M.: The philosophical basis of intuitionistic logic. In: Rose, H.E., et al. (eds.) Logic Colloquium '73, pp. 5–40. Amsterdam, North-Holland (1975)
3. Dummett, M.: The Logical Basis of Metaphysics. Duckworth, London (1991)
4. Gentzen, G.: Untersuchungen über das logische Schließen. Mathematische Zeitschrift **39**, 176–210, 405–431 (1934–1935)
5. Heyting, A.: Sur la logique intuitionniste. Académie Royale de Belgique, Bulletin de la Classe des Sciences **16**, 957–963 (1930)
6. Heyting, A.: Die intuitionistische Grundlegung der Mathematik. Erkenntnis **2**, 106–115 (1931)

7. Heyting, A.: Mathematische Grundlagenforschung, Intuitionismus, Beweistheorie. Springer, Berlin (1934)
8. Heyting, A.: Intuitionism in mathematics. In: Klibansky, R. (ed.) Philosophy in the Mid-Century, pp. 101–115. La Nuova Italia, Florence (1958)
9. Howard, W.: The formula-as-types notion of construction. In: Seldin, J., et al. (eds.) To H. B. Curry: Essays on Combinatory Logic, Lambda Calculus and Formalism, pp. 479-490. Academic Press, London (1980)
10. Kreisel, G.: Interpretation of analysis by means of constructive functionals of finite types. In: Heyting, A. (ed.) Constructivity of Mathematics, pp. 101–128. North-Holland, Amsterdam (1959)
11. Kreisel, G.: On weak completeness of intuitionistic predicate logic. J. Symb. Log. 27, 139-158 (1962)
12. Kreisel, G.: Foundations of intuitionistic logic. In: Nagel, E., et al. (eds.) Logic, Methodology and Philosophy of Science, pp. 198-212. Stanford University Press, Stanford (1962)
13. Kreisel, G.: Book reviews, the collected papers of Gerhard Gentzen. J. Philos. **68**, 238–265 (1971)
14. Martin-Löf, P.: Hauptsatz for the intuitionistic theory of iterated inductive definitions. In: Fenstad, J.E. (ed.) Proceedings of the Second Scandinavian Logic Symposium, pp. 179-216. North-Holland, Amsterdam (1971)
15. Martin-Löf, P.: Intuitionistic Type Theory. Bibliopolis, Napoli (1984)
16. Prawitz, D.: Natural Deduction: A Proof-Theoretic Study. Almqvist & Wicksell, Stockholm. (1965) (Republished, Dover Publications, New York (2006))
17. Prawitz, D.: Ideas and results in proof theory. In: Fenstad, J.E. (ed.) Proceedings of the Second Scandinavian Logic Symposium, pp. 235-307. North-Holland, Amsterdam (1971)
18. Prawitz, D.: Towards a foundation of general proof theory. In: Suppes, P., et al. (eds.) Logic, Methodology and Philosophy of Science IV, pp. 225–250. North-Holland, Amsterdam (1973)
19. Prawitz, D.: On the idea of a general proof theory. Synthese **27**, 63–77 (1974)
20. Prawitz, D.: Meaning approached via proofs. Synthese **148**, 507–524 (2006)
21. Prawitz, D.: Explaining deductive inference. In: Wansing, H. (ed.) Dag Prawitz on Proofs and Meaning, pp. 65–100. Springer, Cham (2015)
22. Schroeder-Heister, P.: Validity concepts in proof-theoretic semantics. Synthese **148**, 525–571 (2006)
23. Troelstra, A.S.: Aspects of constructive mathematics. In: Barwise, J. (ed.) Handbook of Mathematical Logic, pp. 973–1052. North-Holland, Amsterdam (1977)
24. Troelstra, A.S., van Dalen, D.: Constructivism in Mathematics, vol. 2. North-Holland, Amsterdam (1988)
25. von Plato, J.: Gentzen's proof of normalization for intuitionistic natural deduction. Bull. Symb. Log. 14, 240-257 (2008)

Kreisel's Theory of Constructions, the Kreisel-Goodman Paradox, and the Second Clause

Walter Dean and Hidenori Kurokawa

Abstract The goal of this paper is to consider the prospects for developing a consistent variant of the *Theory of Constructions* originally proposed by Georg Kreisel and Nicolas Goodman in light of two developments which have been traditionally associated with the theory—i.e. Kreisel's *second clause* interpretation of the intuitionistic connectives, and an antinomy about constructive provability sometimes referred to as the *Kreisel-Goodman paradox*. After discussing the formulation of the theory itself, we then discuss how it can be used to formalize the BHK interpretation in light of concerns about the impredicativity of intuitionistic implication and Kreisel's proposed amendments to overcome this. We next reconstruct Goodman's presentation of a paradox pertaining to a "naive" variant of the theory and discuss the influence this had on its subsequent reception. We conclude by considering various means of responding to this result. Contrary to the received view that the second clause interpretation itself contributes to the paradox, we argue that the inconsistency arises in virtue of an interaction between reflection and internalization principles similar to those employed in Artemov's Logic of Proofs.

Keywords BHK interpretation · Intuitionistic logic · Theory of Constructions · the Kreisel-Goodman paradox · Logic of Proofs

1 Introduction

The Brouwer-Heyting-Kolmogorov (BHK) interpretation of intuitionistic logic is traditionally characterized as a means of associating with each formula A of first-order logic a so-called *proof condition* which specifies what is required for an object

W. Dean (✉)
Department of Philosophy, University of Warwick Coventry, CV4 7AL, England, UK
e-mail: W.H.Dean@warwick.ac.uk

H. Kurokawa
Department of Information Science, Kobe University,
1-1 Rokkodai-cho, Nada-ku, Kobe 657-8501, Japan
e-mail: hidenori.kurokawa@gmail.com

T. Piecha and P. Schroeder-Heister (eds.), *Advances in Proof-Theoretic Semantics*,
Trends in Logic 43, DOI 10.1007/978-3-319-22686-6_3

to serve as a constructive proof of A in terms of its structure. An interpretation of this form was originally proposed by Heyting [19–21] and Kolmogorov [24], leading to the now familiar formulation reported in [46]:

($P_\wedge$) A proof of $A \wedge B$ consists of a proof of A and a proof of B.
($P_\vee$) A proof of $A \vee B$ consists of a proof of A or a proof of B.
($P_\rightarrow$) A proof of $A \rightarrow B$ consists of a construction which transforms any proof of A into a proof of B.
($P_\neg$) A proof of $\neg A$ consists of a construction which transforms any hypothetical proof of A into a proof of $\perp$ (a contradiction).
($P_\forall$) A proof of $\forall x\, A$ consists of a construction which transforms all c in the intended range of quantification into a proof of $A(c)$.
($P_\exists$) A proof of $\exists x\, A$ consists of an object c in the intended range of quantification together with a proof of $A(c)$.

Alongside such a formulation it is conventional to add the caveat that the notions of proof and construction alluded to in these clauses should be understood as primitives, and thus cannot be taken to correspond to derivations in any particular formal system. Rather than providing a formal semantics for intuitionistic first-order logic in a manner parallel to that provided by Tarski's definitions of truth and satisfaction for classical logic, the BHK interpretation is now often described as providing a so-called *meaning explanation* of the intuitionistic logical connectives [39]—i.e. "an account of what one knows when one understands and correctly uses the logical connectives" [47].

Despite the fact that it itself is not intended as a mathematical *interpretation* in the technical sense, the BHK interpretation has been a substantial source of work in proof theory and related disciplines which can be understood as attempting to provide a formal semantics for intuitionistic logic. Among such developments are Kleene realizability, Gödel's *Dialectica* interpretation, and Martin-Löf's Intuitionistic Type Theory [ITT]. The class of systems which we will investigate in this paper—i.e. the so-called *Theory of Constructions* which was originally developed by Georg Kreisel [25, 26], and Nicolas Goodman [16–18] in the 1960s and 1970s[1]—was also put forth in much the same spirit. For instance Kreisel originally explained the aims of the theory as follows:

> Our main purpose here is to enlarge the stock of formal rules of proof which follow directly from the meaning of the basic intuitionistic notions but not from the principles of classical mathematics so far formulated. The specific problem which we have chosen to lead us to these rules is also of independent interest: *to set up a formal system, called 'abstract theory of constructions' for the basic notions mentioned above, in terms of which formal rules of Heyting's predicate calculus can be interpreted.*

[1] As we will see below, the theories which are presented in these papers as "theories of constructions" vary in some crucial respects. Although it is thus inaccurate to speak of a *unique* formal system as corresponding to "the" Theory of Constructions, we will retain the definite article in speaking of the family of theories in question when no confusion will result.

> In other words, we give a formal semantic foundation for intuitionistic formal systems in terms of the abstract theory of constructions. This is analogous to the semantic foundation for classical systems [42] in terms of abstract set theory [25, pp. 198–199] (emphasis in the original).

The Theory of Constructions was thus unabashedly put forth as an attempt to mathematically formalize the BHK interpretation. But as we will see, there are at least two reasons to view the theory as providing a more direct analysis of the individual BHK clauses than the approaches mentioned above. First, (unlike, e.g., *Dialectica* or ITT) it treats constructive proofs explicitly as abstract objects whose properties we can reason about directly. This allows us to construct expressions which can be understood as direct translations of the BHK clauses into a language with variables which are intended to range over such proofs. Second, Goodman describes his formulation of the system as "a type- and logic-free theory directly about the rules and proofs which underlie constructive mathematics" [17, p. 101]. At least in the eyes of its originators, the Theory of Constructions thus represents an attempt to provide an account of intuitionistic validity in terms of elementary notions which (unlike, e.g., Beth or Kripke models or Kleene realizability) do not presuppose classical logic or mathematics.

But despite these far ranging ambitions, the Theory of Constructions has largely been neglected in surveys of the semantics of intuitionistic logic (e.g. [7, 46]) from the early 1980s onward. Two reasons for this appear to be as follows: (1) a "naive" form of the theory was shown by Goodman [16, 17] to be inconsistent in virtue of a "self-referential" antinomy involving constructive provability (we will see below that this is similar in form to what is now known as *Montague's paradox*); (2) it was in the context of presenting the Theory of Constructions in which Kreisel first presented a modification to the clauses ($P_{\rightarrow}$), ($P_{\neg}$) and ($P_{\forall}$) (which has come to be known as the *second clause*) which proved to be controversial and has subsequently been excised from modern expositions of the BHK interpretation.

The broad goal of the current paper will be to take some initial steps towards reevaluating the Theory of Constructions with respect to its original foundational goals. We will do so by first focusing on how the aspects of the theory just mentioned—i.e. Kreisel's second clause and the Kreisel-Goodman paradox—influenced both the original formulation of the theory as well as its subsequent reception. In Sect. 2, we will consider the features of the original formulation of the BHK interpretation which appear to have motivated Kreisel to introduce the second clause—i.e. the decidability of what we will refer to as the *proof relation* and the putative impredicativity of the clauses ($P_{\rightarrow}$), ($P_{\neg}$) and ($P_{\forall}$). In Sect. 3 we will then provide a concise account of the various formal systems considered by Kreisel and Goodman, their use in formalizing the BHK interpretation (inclusive of the second clause), and their relationship to the Kreisel-Goodman paradox. In Sect. 4 we will consider the reaction of various theorists to the Theory of Constructions and the second clause, as well as evaluating Weinstein's [49] claim that the second clause is itself to blame for the paradox. After concluding that this contention is unjustified, in Sect. 5 we will consider other poten-

tial diagnoses of the paradox, as well as discussing the prospects for formulating a version of the Theory of Constructions which addresses Kreisel and Goodman's original foundational goals.

2 Predicativity, Decidability, and the BHK Interpretation

One of Kreisel's goals in proposing the Theory of Constructions was to respond to a potential objection to the BHK interpretation which had been raised by Gödel. This problem can be understood to arise in two stages. First note that the BHK clauses initially appear to provide a characterization of the relation "*p is a proof of A*" in terms of the logical form of A, an observation which might in turn be taken to provide an implicit definition of the class of constructive proofs to which the interpretation refers. But on the other hand, note that the BHK clauses themselves cannot be taken as constituting a proper *inductive* definition of such a class in virtue of the fact that the clauses ($P_{\rightarrow}$), ($P_{\neg}$), and ($P_{\forall}$) contain quantifiers which are intended to range over the class of *all* constructive proofs, potentially inclusive of those which figure in the proof conditions of yet more complex formulas.

We will refer to this *prima facie* objection to the BHK interpretation as the *problem of impredicativity*. Gödel remarked on this aspect of the interpretation already in the following passage from a 1933 lecture in which he is attempting to compare the relative merits of Hilbert's finitism (as codified by the system he calls A) and intuitionism as foundational frameworks for formulating mathematical consistency proofs:

> So Heyting's axioms concerning absurdity and similar notions differ from the system A only by the fact that the substrate on which the consequences are carried out are proofs instead of numbers or other enumerable sets of mathematical objects. But by this very fact they do violate the principle, which I stated before, that the word "any" can be applied only to those totalities for which we have a finite procedure for generating all their elements. For the totality of all possible proofs certainly does not possess this character, and nevertheless the word "any" is applied to this totality in Heyting's axioms, as you can see from the example which I mentioned before, which reads: "Given *any* proof for a proposition p, you can construct a reductio ad absurdum for the proposition $\neg p$". Totalities whose elements cannot be generated by a well-defined procedure are in some sense vague and indefinite as to their borders. And this objection applied particularly to the totality of intuitionistic proofs because of the vagueness of the notion of constructivity [13, p. 53].

Gödel can be understood as flagging three points which have played a substantial role in guiding the subsequent understanding of the BHK interpretation: (1) a crucial difference between finitism and intuitionism is that, unlike finitists, intuitionists do not reject the meaningfulness of unrestricted quantification over a potentially infinite domain; (2) the class of constructive proofs form such a totality; but (3) this class should not be regarded as inductively generated in virtue of the occurrence of the universal quantifier over proofs in (e.g.) the clause ($P_{\neg}$).

The first point is stressed by Weinstein [49] in the course of suggesting how the Theory of Constructions might play a role in how an intuitionist ought to reply to Benacerraf's [4] dilemma in philosophy of mathematics. One horn of the dilemma alleges that a "combinatorial" theorist (i.e. one who attempts to identify truth and provability in the characteristic manner of both intuitionism and formalism) will be unable to provide a semantical account of mathematical language which is continuous with the standard referential semantics which we may wish to give for natural language as a whole. But in addition to this, Benacerraf also argues that Hilbert's development of finitism has the added disadvantage of needing to provide distinct accounts of finitary (i.e. "real") and infinitary (i.e. "ideal") mathematics.

It is in this regard that Weinstein suggests that intuitionism may have an advantage over finitism in the sense that the BHK clauses can be understood as providing a uniform semantic account applicable to both real and ideal mathematical statements. As he stresses in the following passage, however, this advantage can only be claimed if it is ensured that the proof relation is *decidable*:

> Proofs, for the intuitionist, are not to be equated with formal proofs, that is with some kind of finite quasi-perceptual objects, and, more to the point, decidable properties of proofs may involve considerations about the intuitive content of these mathematical constructions. Hence, it is precisely by admitting as meaningful the notion of a decidable property holding for arbitrary mathematical constructions that intuitionists achieve an interpretation of those sentences which are from Hilbert's point of view devoid of intuitive content. And, for intuitionists, to admit this notion as meaningful is to claim that statements asserting that decidable properties of mathematical constructions hold universally have tolerably clear proof conditions. Thus, by enlarging the contentual portion of mathematics to include universal decidable statements which are not finitary the intuitionists achieve an interpretation of mathematical statements of arbitrary logical complexity [49, p. 268].

Weinstein goes on to explain the connection between the decidability of the proof relation and the attribution of content to mathematical statements as follows:

> [I]ntuitionists identify the truth of a mathematical statement, A, with our possession of a construction, c, which is a proof of the statement A. This latter statement, that the construction c is a proof of A, involves no logical operations and is moreover the application [of] a decidable property to a given mathematical construction. Hence, this statement does not itself require a non-standard semantical interpretation and, it is hoped, can be understood along the lines of statements like "The liberty bell is made out of brass" [. . .] The idea is just that the intended intuitionistic interpretation of a mathematical language reduces the truth of any sentence of that language to the truth of an atomic sentence which is the application of a decidable predicate to a term and this latter sentence can be understood as having an ordinary referential interpretation [49, pp. 268–269].

Although we will see below that the decidability of the proof relation has occasionally been disputed, these passages make clear why it has traditionally been thought to play a crucial role in ensuring that the BHK clauses are compatible with the general goal of explaining how truth can be understood in terms of constructive provability. To see how this is related to Gödel's second and third points about how the class of constructive proofs may be characterized, note that if we assume that the proof relation itself is decidable, then the clauses ($P_{\rightarrow}$), ($P_{\neg}$), and ($P_{\forall}$) are all analogous in form to Π^0_1 statements in the language of arithmetic—i.e. they begin with an unrestricted

universal quantifier over proofs applied to a decidable matrix.[2] As such statements are not in general decidable in the technical sense of computability theory, it seems that there is reason to worry that they do not satisfy Weinstein's criteria of having "tolerably clear proof conditions" even when understood informally.

It is now only a small step which must be taken to justify the use of the term "impredicativity" to label the problem which was described by Gödel. For as Kreisel later observed

> [I]t is one of the peculiarities of constructive logic that, for some A, a *natural* formal proof of A goes *via* proofs of $A \to B$ and $(A \to B) \to A$: such a proof of A actually contains a proof of $A \to B$ [27, p. 58].

Although Kreisel formulates this point for *formal* proofs, there seems to be no *a priori* reason to suspect that the same comment should not apply to the pre-theoretical notion of constructive proof which the BHK interpretation seeks to characterize. And if this is indeed the case—i.e. that there exist formulas A which are demonstrable by proofs which may contain sub-demonstrations of formulas which contain A itself—then it seems that the quantifier over constructive proofs occurring in (e.g.) ($\mathrm{P}_{\to}$) must be understood as ranging over a totality to which it itself belongs.

A variety of other commentators have also used terms like "circular" or "impredicative" to describe either the BHK clauses or the status of implication in intuitionistic logic more generally.[3] As we will see below, it appears that Kreisel added the second clause to the formulations of ($\mathrm{P}_{\to}$), ($\mathrm{P}_{\neg}$), and ($\mathrm{P}_{\forall}$) precisely to avoid such charges and thereby also to provide a characterization of the proof relation which could plausibly be regarded as decidable. What remains to be seen is whether his attempt should be regarded as successful and also whether the various latter day critiques which have been directed towards the second clause also undermine the rationale for adopting the Theory of Constructions itself.

3 The Theory of Constructions and the Second Clause

Without further ado, we now present Kreisel's proposed modification of ($\mathrm{P}_{\to}$):

($\mathrm{P}^2_{\to}$) A proof of $A \to B$ consists of a construction that transforms any proof of A into a proof of B *together with a proof that this construction satisfies the desired property*.

The italicized material represents what is customarily referred to as the "second-clause"—i.e. the requirement that a constructive proof q of a conditional $A \to B$ is

[2] It might also be objected that the explanation of implication given by ($\mathrm{P}_{\to}$) is circular because it employs the conditional "*if* p is a proof of A, *then* $f(p)$ is a proof of B" on its righthand side. Note, however, that if it can be maintained that the proof relation is decidable, then it can also be maintained that it is permissible to interpret this conditional truth functionally.

[3] E.g. Gentzen [11, p. 167], Goodman [16, p. 7], Troelstra [45, p. 210], Dummett [7, Sect. 7.2], Fletcher [10, p. 81], and Tait [41, p. 221].

not just a construction transforming arbitrary proofs of A into proofs of B in the sense of the original clause ($P_{\rightarrow}$) but rather a pair $\langle p, q\rangle$ consisting of such a construction together with another proof p which demonstrates that q has this property. The second-clause variants are formed by adding similar clauses to ($P_{\neg}$) and ($P_{\forall}$).

Such a reformulation of BHK—which we henceforth refer to as the *BHK*^{2} *interpretation*—was stated for the first time by Kreisel [25, p. 205] and again in [26, p. 128]. In both instances, Kreisel used the formal language of the Theory of Constructions to formulate ($P^2_{\rightarrow}$). But although both of these treatments appear to have been informed by Heyting's [21] mature exposition of the original interpretation, in neither instance does Kreisel motivate the second clause directly nor does he flag that he is intending to either refine or depart from Heyting's original intentions.

These observations notwithstanding, the initial reception of the second clause appears to have been positive—e.g. second clauses are included in both Troelstra [43, p. 5] and van Dalen's [48, p. 24] surveys of intuitionistic logic (again without additional historical comment). But as we will discuss further below, by the early to mid-1980s the consensus appears to have shifted to the view that not only should the second clause not be included in the canonical formulation of BHK, but also that its very formulation rested on dubious assumptions about the nature of constructive proof.[4]

One of our goals below will be to better understand what underlies this shift in opinion about the second clause. Although subsequent commentators have typically followed Troelstra and van Dalen in formulating ($P^2_{\rightarrow}$) informally, we will suggest below that its status is bound up not only with the issues of impredicativity and decidability discussed in the prior section, but also with certain details about how ($P^2_{\rightarrow}$) should be formalized within the Theory of Constructions itself. Before turning to such considerations, it will thus be useful to consider both the formulation of the theory and how it may be used to formalize the BHK^2 interpretation.

3.1 An Overview of the Theory of Constructions

Versions of the Theory of Constructions were presented by Kreisel [25, 26], and Goodman [16–18]. The details of the notation and formal systems formulated in these papers differ in several respects. Our goal here will thus not be to present a systematic exposition of the different formalisms proposed by Kreisel and Goodman, nor even to provide a complete formulation of any one of them. Rather we shall simply attempt to set down some of the common characteristics of these systems with the dual goals of explaining how Kreisel and Goodman proposed to use the language of the Theory of Constructions to formalize Kreisel's reformulation of the BHK

[4]This shift in opinions is illustrated by the fact that while when Troelstra [44] originally coined the acronym "BHK", the "K" was taken to stand for Kreisel, this convention is modified by Troelstra and van Dalen [46] who take the "K" to stand for Kolmogorov.

clauses and also to be able to reconstruct as closely as possible the reasoning of the Kreisel-Goodman paradox.

In so doing, we will adhere as closely as possible to the notation and terminology of the *unstratified* (or "naive") theory (which we will henceforth refer to as $\mathscr{T}$) which is sketched by Goodman [17] in the course of expositing the paradox. (This system should be understood in contradistinction to the *stratified* theory $\mathscr{T}^{\omega}$ which Goodman officially adopts.[5]) Before offering a formal description of $\mathscr{T}$, however, it will be useful for orientation to record several of its features which are remarked on by Sundholm [39]:

(I) The system $\mathscr{T}$ treats proofs as constructions $s, t, u, \ldots$, which themselves are understood as mathematical objects whose properties the theory attempts to axiomatize.
(II) Using the theory it is possible to define a decidable predicate $\Pi(A, s)$ with the intended interpretation "construction s proves proposition A".
(III) Statements of the latter form are themselves treated by the theory as propositions which may themselves admit to proof. In particular, it is possible within the theory to formulate statements such as $\Pi(\Pi(A, s), t)$ (i.e. "construction t proves that construction s is a proof of A").

It would appear that the ability to iterate the application of the predicate $\Pi(A, s)$ is necessary if we are to formalize clauses such as ($\mathrm{P}^2_{\rightarrow}$). But note that if this is allowed, it must also be acknowledged that the constructions must play a dual role in $\mathscr{T}$—e.g. if $\langle p, q\rangle$ is a pair satisfying the proof conditions of $A \rightarrow B$ per ($\mathrm{P}^2_{\rightarrow}$), then q is understood as a *process* (i.e. a method for transforming proofs of A into proofs of B), while q is regarded as an *object* (i.e. a completed proof that q has the required property). Sundholm [39, pp. 164–167] suggests that these two notions must be carefully distinguished if we are to develop a theory of constructions which is faithful to Heyting's original interpretation of the connectives. He also suggests (at least implicitly) that Kreisel may have conflated them in his own formulations of $\mathscr{T}$. But although this concern might be taken to call for reconsideration of the theory on historical grounds, the perspective which we will adopt here is that the specific proposals of Kreisel and Goodman are of interest in their own right.

3.1.1 The Language of $\mathscr{T}$

Described in general terms, $\mathscr{T}$ is an equational term calculus with pairing, projection, and lambda abstraction operators, application, as well as various other primitive terms

[5] Goodman's dissertation [16] provides the most comprehensive exposition of $\mathscr{T}^{\omega}$, inclusive of the interpretation of intuitionistic first-order logic, Heyting arithmetic, and accompanying consistency and faithfulness proofs. But whereas in [16] the Kreisel-Goodman paradox is presented informally, [17] contains a more detailed derivation in theory (similar or identical to what we will call $\mathscr{T}^{+}$) which is similar to the "starred" variant originally described by Kreisel [25]. We will discuss these systems in greater detail in the context of evaluating Goodman and Kreisel's response to the paradox in Sect. 5.

and predicates (which are formalized as boolean-valued terms). The terms of the theory are intended to denote "constructions" which can be understood simultaneously as either proofs or operations on proofs—i.e. what the theory seeks to axiomatize is a notion of "self-applicable" proof. The distinctive feature of all versions of the Theory of Constructions is the inclusion of a proof operator π whose intended role can be most readily described as that of axiomatically mimicking certain properties of a traditional proof predicate $\mathtt{Proof}_{\mathsf{T}}(x, y)$ for an arithmetical theory T (such as Peano or Heyting arithmetic).

More formally, the class of terms of $\mathscr{T}$ is defined by the grammar

$$t ::= x \mid \top \mid \bot \mid \langle D(tt)\rangle \mid \langle D_1(t)\rangle \mid \langle D_2(t)\rangle \mid \langle \lambda x.t\rangle \mid \langle tt\rangle \mid \langle \pi tt\rangle$$

where $x, y, z, \ldots$ are variables, $\top$ and $\bot$ are intended to denote the truth values *true* and *false*, $D(st)$ is intended to denote the pair $\langle s, t\rangle$, $D_i(t)$ is intended to denote the first ($i = 1$) or second ($i = 2$) member of t if t is a pair and is undefined otherwise, and $\lambda x.t$ (i.e. abstraction) and st (i.e. application) are defined as usual in the untyped lambda calculus. The formulas of $\mathscr{T}$ are equations of the form $s \equiv t$. Note, however, that implicit in Goodman's [17] (and previously Kreisel's [25]) decision to base the Theory of Constructions on the *untyped* lambda calculus is that terms of the theory may be undefined. The relation $\equiv$ is thus intended to denote a notion of *intensional identity* between terms—i.e. $s \equiv t$ is intended to hold just in case s and t are both defined and reduce to the same normal form under β-conversion.

3.1.2 The Axiomatization of $\mathscr{T}$

Goodman's axiomatization of $\mathscr{T}$ is based on a single conclusion sequent calculus relative to which $\Delta \vdash_{\mathscr{T}} s \equiv t$ is assigned the intended interpretation "if all the equations in Δ hold, then $s \equiv t$". The structural rules of the system include weakening and cut. Additionally, equality axioms for $\equiv$ (e.g. $\vdash_{\mathscr{T}} s \equiv s$) as well as axioms governing the pairing operators (e.g. $\vdash_{\mathscr{T}} D_i(Ds_1s_2) \equiv s_i$) are adopted. We will assume that lambda terms are axiomatized by the formal theory $\lambda\beta$ of [22, p. 70].[6]

The most significant axioms of $\mathscr{T}$ are those pertaining to the binary operator π. Goodman [17, p. 107] describes the intended interpretation of this symbol as follows:

$$\pi st \equiv \top \text{if and only if } t \text{ is a proof that for all } x, sx \equiv \top$$

[6]The systems of [16, 17] do not officially have the abstraction operator in their language, but rather the traditional combinators S and K which may be used to mimic lambda abstracti on—e.g. in the manner described in [22, Sect. 2.2]. But as Goodman makes free use of λ-notation throughout both of his expositions (apparently via such an abbreviation), it will be here simpler to assume that the system includes $\lambda\beta$ instead of the rules which Goodman takes to axiomatize the combinators. Until Sect. 5, we will also suppress discussion of a number of other primitive notions and their corresponding axioms pertaining to the treatment of so-called "grasped domains" which are introduced in the formulation of $\mathscr{T}^{\omega}$.

Thus an equation of the form of the $\pi st \equiv \top$ is intended to express that t is a construction which serves as a proof of the fact that for all x the term sx reduces to the value $\top$.[7] One of the rules which is assumed to hold of π is intended to express that the proof relation described in Sect. 2 is decidable. This is achieved as follows:

$$\text{(DEC)}\quad \frac{\Delta, \pi uv \equiv \bot \vdash_{\mathscr{T}} s \equiv t \qquad \Delta, \pi uv \equiv \top \vdash_{\mathscr{T}} s \equiv t}{\Delta \vdash_{\mathscr{T}} s \equiv t}$$

The other principle which is assumed to hold of π is a form of *reflection principle* stating that if the proof relation holds between s and t then sx is true:

$$\text{(EXPRFN)}\quad \pi st \equiv \top \vdash_{\mathscr{T}} sx \equiv \top.$$

As both DEC and EXPRFN play a role in the derivation of the Kreisel-Goodman paradox, it will be useful to say something additional both about their motivation and also their formulation in the Theory of Constructions. As we have observed in Sect. 2, the decidability of the proof relation appears to have a strong pre-theoretical basis in the intuitionists' desire to view the BHK clauses as providing a decidable proof condition for formulas of arbitrary logical complexity. Although $\mathscr{T}$ does not contain any primitive relation symbols itself, a term α can be understood as expressing a binary relation just in case for all pairs of terms s, t, if αst is defined, then $\alpha st \equiv \top$ or $\alpha st \equiv \bot$ may be derived in the theory. The decidability of such a relation α may then be expressed by stating that αst is defined for all pairs of terms s, t—i.e. that α is *bivalent*.[8] This is what is formulated proof theoretically by the rule DEC in the case of the term π—i.e. in order to exclude the "third" case that πuv is undefined, we stipulate that it is sufficient to conclude $s \equiv t$ from Δ if this equation is derivable from both the hypotheses $\Delta, \pi uv \equiv \top$ and also $\Delta, \pi uv \equiv \bot$.

EXPRFN is a form of what we will call an *explicit reflection* principle (cf. [1])—i.e. an expression of the fact that if the proof relation holds between a constructive proof p and a formula A, then we can conclude that A is true. Kreisel [25, p. 204] remarks of such a principle that it is "obvious on the intended interpretation" of π. In the arithmetical case, we would typically express this using a conditional statement of the form $\mathtt{Proof}_{\mathsf{T}}(\overline{n}, \ulcorner\phi\urcorner) \rightarrow \phi$, all of whose instances are both valid in the standard model and provable in even weak arithmetical systems T.[9] But since the Theory of Constructions does not contain a sign for implication in its object language, this is expressed in $\mathscr{T}$ by the rule EXPRFN which allows us to conclude $sx \equiv \top$

[7] Relative to this interpretation, πst can be understood as expressing the characteristic function of the assertion that s is a proof of the universal closure of the logical formula which s interprets. In the sequel, however, s will most often be closed. And thus it will often be possible to understand πst as simply expressing that t is a proof of the formula interpreted by s.

[8] Note that by analogy with the arithmetical case, we will typically have $\mathsf{T} \vdash \mathtt{Proof}_{\mathsf{T}}(\overline{n}, \ulcorner\phi\urcorner) \vee \neg\mathtt{Proof}_{\mathsf{T}}(\overline{n}, \ulcorner\phi\urcorner)$ in virtue of the fact that $\mathtt{Proof}_{\mathsf{T}}(x, y)$ is standardly defined to be a Δ^0_1 arithmetical formula. This observation about the *derivable* properties of $\mathtt{Proof}_{\mathsf{T}}(x, y)$ appears to have been an important part of Kreisel's motivation for insisting upon the decidability of π in the Theory of Constructions—a feature which he famously justified by observing that "we recognize a proof of an assertion when we see one" [26, p. 124]. (See [39] for additional discussion of this point.).

[9] We will return in Sect. 5.4 to compare EXPRFN to the better known "implicit" reflection principle $\exists x \mathtt{Proof}_{\mathsf{T}}(x, \ulcorner\phi\urcorner) \rightarrow \phi$.

for all x from the premise $\pi st \equiv \top$. As Goodman [17, p. 106] observes, in this sense the derivability relation $\vdash_{\mathscr{T}}$ should itself be interpreted as expressing a form of intuitionistic implication.

3.1.3 Formalizing the BHK Interpretation in $\mathscr{T}$

Recall that Kreisel's original goal in introducing the Theory of Constructions was to formulate a formal system which could play a role analogous to Tarski's definition of truth for Heyting Predicate Calculus (HPC). In order to see how this might be achieved, it is useful to note that at least at an informal level, the BHK clauses can be understood as serving a role analogous to the clauses in Tarski's definition of truth in a model—i.e. that of providing a characterization of "constructive validity" relative to which it might be hoped that a logical system such as HPC could be shown to be sound and complete in the same sense that the Classical Predicate Calculus CPC is sound and complete with respect to classical validity (i.e. truth in all Tarskian models).

But before investigating how Kreisel and Goodman proposed to interpret the BHK^2 clauses in the language of $\mathscr{T}$, it is useful to first remark upon one important sense in which these clauses differ from those of Tarski. For note that on the one hand what occurs on the righthand side of one of the Tarski clauses is a *proposition* stating in the language of set theory what must be true in order for a formula $A(\overrightarrow{x})$ to be true in a model $\mathfrak{A}$ relative to an assignment v of values to variables $\overrightarrow{x}$. But what occurs on the righthand side of the BHK (and BHK^2) clauses are not propositions but rather *conditions* stating the circumstances under which a certain object is to be regarded as a proof of $A(\overrightarrow{x})$ (relative to an assignment of vales to the free variables $\overrightarrow{x}$). Thus whereas the formalization of the Tarskian satisfaction relation $\mathfrak{A} \models_v A(\overrightarrow{x})$ yields a *sentence* which can be formalized in the language of set theory, we should expect the formalization of the BHK clauses to yield a *predicate*—which Kreisel [25] symbolizes as $\Pi(A(\overrightarrow{x}), s)$—which is intended to hold of a proof s just in case it is a proof of a formula $A(\overrightarrow{x})$.

Kreisel and Goodman's formalizations of the BHK clauses thus can be understood as attempting to provide a definition of $\Pi(A(\overrightarrow{x}), s)$ which were intended to serve the role of providing a formalization of the proof relation as defined above. In order to understand the general form which their definitions took, note first that as with the analogous Tarski clauses, the BHK clauses (as well as their BHK^2 counterparts) employ logical connectives on their righthand sides—e.g. the clause ($\mathrm{P}_{\rightarrow}$) states that p is a proof of $A \rightarrow B$ just in case *for all* proofs x, *if* x is a proof of A, *then* $p(x)$ (i.e. the result of applying p to x) is a proof of B. In addition to the problem of impredicativity discussed in Sect. 2, there is also another apparent obstacle in rendering the conditional *if* … *then* appearing in this clause as a term in the "logic free" language of $\mathscr{T}$.

Kreisel and Goodman proposed to circumvent this problem by taking advantage of the following observations: (1) it is intuitionistically admissible to apply classical propositional logic to decidable statements; (2) if the truth values $\top$ and $\bot$ are taken

as abbreviating particular λ-terms, it is possible to define bivalent λ-terms $\cap_k$, $\cup_k$, and $\supset_k$ which mimic the classical truth functional connectives $\wedge$, $\vee$, and $\rightarrow$ applied to binary terms with k free variables[10]; (3) the application of these terms to terms of the form $\Pi(A(\overrightarrow{x}), s)$ will always yield a term which is defined as long as it can be ensured that $\Pi(A(\overrightarrow{x}), s)$ is itself defined so that it is bivalent.

Taking these observations into account, we can now formulate Kreisel's [25] definition of $\Pi(A, s)$ (where we assume that the free variables of A and B are contained in $\overrightarrow{x}$ of arity k) in the language of $\mathscr{T}$ as follows[11]:

$$
\begin{array}{ll}
(\mathrm{K}_\wedge) & \Pi(A \wedge B, s) := \lambda\overrightarrow{x}.(\Pi(A, D_1 s) \cap_k \Pi(B, D_2 s)) \\
(\mathrm{K}_\vee) & \Pi(A \vee B, s) := \lambda\overrightarrow{x}.(\Pi(A, D_1 s) \cup_k \Pi(B, D_2 s)) \\
(\mathrm{K}_\rightarrow) & \Pi(A \rightarrow B, s) := \lambda\overrightarrow{x}.\pi(\lambda y.(\Pi(A, y) \supset_k \Pi(B, (D_2 s)y)), D_1 s) \\
(\mathrm{K}_\neg) & \Pi(\neg A, s) \quad := \lambda\overrightarrow{x}.\pi(\lambda y.(\Pi(A, y) \supset_k \Pi(\bot, (D_2 s)y)), D_1 s) \\
(\mathrm{K}_\forall) & \Pi(\forall z A(z), s) := \lambda\overrightarrow{x}.\pi(\lambda y.\Pi(A[y/z], (D_2 s)y), D_1 s) \\
(\mathrm{K}_\exists) & \Pi(\exists z A(z), s) := \lambda\overrightarrow{x}.\Pi(A[(D_1 s)/z], D_2 s)
\end{array}
$$

Note that these clauses provide a straightforward expression of the clauses of the BHK^2 interpretation—e.g. $(\mathrm{P}^2_\rightarrow)$ is formalized by requiring that $\Pi(A \rightarrow B, s)$ holds just in case s is a pair such that $D_1 s$ is a proof that $D_2 s$ has the property of being such that if $\Pi(A, y)$, then $\Pi(B, (D_2 s)y))$. But since $(\mathrm{K}_\rightarrow)$, $(\mathrm{K}_\neg)$, and $(\mathrm{K}_\forall)$ are all of the form $\pi s t$, Kreisel's clauses can be understood as defining $\Pi(A, s)$ in terms of $\pi x y$ in such a way that the decidability of the primitive proof relation is transferred inductively to the complex proof relation.

3.1.4 Soundness, Completeness, and Internalization

The foregoing clauses can thus be understood as providing a means of interpreting the language of HPC into the language of $\mathscr{T}$ so as to provide an analysis of $\Pi(A, s)$ as characterized informally by the BHK^2 clauses. The next question we must consider is how this interpretation comports with the intuitionists' desire to identify truth and constructive provability. But needless to say, this question is complicated at least to the extent that it is traditionally maintained that "constructive provability" must be distinguished from "provable in a given formal system".

---

[10]For instance if we take $\top =_{\mathrm{df}} \lambda xy.x$ and $\bot =_{\mathrm{df}} \lambda xy.y$ (cf. [2]), then we may define $\supset_1$ to be $\lambda xyz.xzy(\lambda w.\top)z$.

[11]Goodman [16, 17] provides a related interpretation of the BHK clauses in the language of the stratified theory $\mathscr{T}^\omega$. However, relative to his interpretation, the variable y in $(\mathrm{K}_\rightarrow)$, $(\mathrm{K}_\neg)$, and $(\mathrm{K}_\forall)$ is asserted to range over proofs of a lower "level" than that of the proof $D_1 s$ (see Sect. 5.2). Kreisel and Goodman also handle the case of atomic formulas differently. On the one hand, Kreisel introduced primitive terms into the language to serve as constructions which act as the characteristic functions of non-logical predicates, which are then individually asserted to be decidable. On the other hand, Goodman considers only the language of primitive recursive arithmetic, wherein all atomic statements are equations of the form $f_1(\overrightarrow{x}) = f_2(\overrightarrow{x})$. True equations of this form are asserted to fall under the decidable equality predicate Q which he introduces as another primitive to the language of $\mathscr{T}^\omega$.

One might think that this entails that the related notion of "constructive validity" which we might hope to characterize using a system in which the BHK clauses can be interpreted must be distinguished from "valid with respect to a particular form of formal semantics".[12] Nonetheless, Kreisel and Goodman both appear to have viewed the Theory of Constructions as providing an "informally rigorous" analysis of constructive validity. In particular, both present versions of the following result for the systems described in [25, 26], and [17] (wherein $\mathscr{T}^*$ is the relevant formulation of the Theory of Constructions):

(VAL) For all formulas A in the language of HPC, $\vdash_{\mathsf{HPC}} A$ if and only if there exists a term s such that $\vdash_{\mathscr{T}^*} \Pi(A, s) \equiv \top$.

The left-to-right direction of VAL can be taken to express a form of soundness for Kreisel's interpretation of HPC into $\mathscr{T}^*$—i.e. if A is derivable from what are normally regarded as intuitionistically valid principles of reasoning, then A is indeed "constructively valid" in the sense that there is some construction which witnesses its derivability. Conversely, the right-to-left direction of VAL can be taken to express a form of completeness (also known as *faithfulness*) of the interpretation—i.e. if A is "constructively valid" in the sense that $\Pi(A, s)$ holds for some construction s, then A is in fact derivable from intuitionistically valid principles.[13]

Although both Kreisel and Goodman announced versions of these results, the situation surrounding their claims is complicated by several factors which we will not consider in detail here.[14] For what is more germane to our immediate concerns is not whether any particular variant of the Theory of Constructions satisfies VAL, but rather whether such systems satisfy what can be understood as a generalized form of soundness which we will refer to as *internalization*. Note that if we are able to demonstrate the left-to-right direction of VAL (say by induction on the length of proofs in HPC), then it also seems reasonable to suppose that we ought to be able to do this for all derivations carried out in $\mathscr{T}$ itself.[15] This would suggest that the Theory of Constructions ought to satisfy a principle of the following form:

(INT) If $\vdash_{\mathscr{T}^+} s \equiv \top$, then there exists a term c such that $\vdash_{\mathscr{T}^+} \pi s c \equiv \top$.

Here c might either be taken as a new constant or as a complex term which is built up according to the structure of the derivation of $s \equiv \top$. (Although we will return to discuss this issue in Sect. 5.5, for the moment we will assume the former interpretation

[12] For discussion of the intuitive notion of constructive validity and its relationship to various formal semantics for intuitionistic logic, see (e.g.) Scott [37], Dummett [7, chap. 5], and McCarty [32].

[13] Compare Scott [37, p. 256]: "The reason that A is *intuitionistically* (constructively, if you prefer) *valid* is that there is a specific term τ […] such that the assertion $\vdash \tau \in A$ is *provable* in the theory of constructions.".

[14] For instance, although Kreisel states versions of the completeness and faithfulness results ([25, p. 205] and [26, Sect. 2.311]), in neither case are proofs given. And although Goodman [16] contains complete proofs of both directions, the interpreting theory in his case is not $\mathscr{T}$, but rather the stratified theory $\mathscr{T}^\omega$.

[15] In fact, this is exactly how the soundness proof for HPC given by Goodman [16, Sect. 11–15] for $\mathscr{T}^\omega$ proceeds.

so as to maintain conformity with the way in which Kreisel and Goodman handle internalization.)

3.2 The Kreisel-Goodman Paradox

Although Kreisel [25] sketched a means by which one version of the Theory of Constructions could be shown to be consistent relative to Heyting arithmetic, he also observed that a carelessly formulated version of the theory (e.g. the "starred" theory of [25]) might turn out to be inconsistent. Although he does not explicitly describe what form such an inconsistency might take, in retrospect it is not difficult to see that the intended interpretation of π makes the issue of consistency of a system such as $\mathscr{T}$ or $\mathscr{T}^+$ a significant cause for concern.

To better appreciate why this is so, it is useful to begin by considering the following paradox pertaining to the notion of informal (or "absolute") provability. Suppose that we elect to express this notion by a predicate $P(x)$ of sentences. Additionally suppose that T is a mathematical theory which we have adopted for reasoning about the properties of $P(x)$ and that $\ulcorner\cdot\urcorner$ is a device which allows us to name sentences in $\mathscr{L}_{\mathsf{T}}$ (such as Gödel numbering). In order to support such a mechanism, it seems reasonable to assume that T will contain Robinson arithmetic Q (either directly or by interpretation). And from this it will follow that T will also be able to prove the existence of self-referential statements about the predicate $P(x)$ via the appropriate analog of Gödel's Diagonal Lemma.

Now consider the following two intuitively correct principles pertaining to informal provability:

(RFNP) If A is informally provable, then it is true—i.e. $P(\ulcorner A\urcorner) \to A$.

(INTP) If we can derive A, then A is informally provable—i.e. $\vdash A \;\therefore\; \vdash P(\ulcorner A\urcorner)$.

It is now easy to see that the theory T^+ obtained by adjoining all instances of RFNP to T and closing under the rule INTP is inconsistent. For by the Diagonal Lemma, let D be a sentence such that (1) $\mathsf{T}^+ \vdash D \leftrightarrow \neg P(\ulcorner D\urcorner)$. Now since (2) $\mathsf{T}^+ \vdash P(\ulcorner D\urcorner) \to D$ by RFNP, we have by (1) that (3) $\mathsf{T}^+ \vdash \neg P(\ulcorner D\urcorner)$. But again by (1), we then also have (4) $\mathsf{T}^+ \vdash D$. It thus follows by INTP that (5) $\mathsf{T}^+ \vdash P(\ulcorner D\urcorner)$, yielding a contradiction with (3).

The observation that an arithmetical theory which extends Q, derives all instances of RFNP, and is closed under INTP is inconsistent has come to be known as *Montague's paradox*.[16] Weinstein [49] subsequently suggested that the Kreisel-Goodman paradox can be understood as a translation of this result into the language of the Theory of Constructions. Goodman offers two expositions of the paradox—an informal

[16]The inconsistency of such a system appears to have first been observed by Myhill [34] in the context of an axiomatic investigation of the notion of informal provability. It was then rediscovered by Montague [33], who presents it as a simplification of the so-called *Paradox of the Knower* as originally formulated in [23]. For more on the history of these results see [5, 6].

one in [16], and a semi-formal one in a system similar to the theory $\mathscr{T}^+$ which is described in the introductory sections of [17]. We quote the former in full:

> The most natural formalization of the conception [of constructive proof] we have outlined so far is inconsistent. It suffices to construct, using π, a function f such that $f(x) = 0$ if and only if $x(x)$ is a proof that no y proves that $f(x) = 0$. Now suppose that y proves that $f(x) = 0$. Then $f(x) = 0$, and so no y proves that $f(x) = 0$. This contradiction, together with the decidability of the proof predicate, shows that no y can prove that $f(x) = 0$. Therefore there must be a function g such that, for any x, $g(x)$ proves that no y proves that $f(x) = 0$. In particular, $g(g)$ proves that no y proves that $f(g) = 0$. That is, $f(g) = 0$. Hence there is a proof that $f(g) = 0$, which is absurd [16, p. 5].

The foregoing passage provides the most complete informal description of the antinomy which subsequent authors have repeatedly associated with the Theory of Constructions. It should be borne in mind, however, that Goodman discusses the paradox *before* providing his "official" formulation of the theory $\mathscr{T}^\omega$ (which he then proceeds to show consistent in a manner we will discuss further in Sect. 5.2). The Kreisel-Goodman paradox thus should not be understood to correspond to a formal contradiction derivable within any of the variants of the Theory of Constructions which were adopted by Kreisel or Goodman themselves. Nonetheless, it will be useful for our current purposes to consider how the reasoning which Goodman describes can be mimicked in the theory $\mathscr{T}^+$ of Sect. 3.1.

As an initial step, we reconstruct the reasoning described in the prior passage in first-order logic by taking the binary predicate $R(A, p)$ to express the proof relation (i.e. "p is a proof of A"), which we will assume satisfies appropriate analogs of DEC, EXPRFN, and INT.[17] Goodman suggests that it is possible to define a function $f(x)$ (which itself should be thought of as a construction) satisfying the equation

$$(1') \ \vdash f(x) = 0 \leftrightarrow R(\forall y \neg R(f(x) = 0, y), x(x))$$

Thus the proposition expressed by $f(x) = 0$ can be understood to express something akin to what is expressed by the sentence D constructed in step (1) of the derivation of Montague's paradox—i.e. that $f(x) = 0$ is true just in case $x(x)$ is a proof that this statement itself is not provable. Next suppose that we have the following instance of the explicit reflection principle EXPRFN for $R(A, p)$

$$(2') \ \ R(f(x) = 0, y) \vdash f(x) = 0$$

But then note that by $(1')$ and modus ponens we also have

$$(2'') \ \ R(f(x) = 0, y) \vdash R(\forall y \neg R(f(x) = 0, y), x(x))$$

Thus by EXPRFN again and universal instantiation we have

$$(2''') \ \ R(f(x) = 0, y) \vdash \neg R(f(x) = 0, y)$$

[17] To simplify notation we will treat $R(A, p)$ as a two-sorted relation which holds between sentences in a first-order language and a class of terms which are understood to denote proofs. It is, nonetheless, straightforward to see that the derivation $(1')$–$(5')$ can be further formalized by treating $R(x, y)$ as a primitive formula which is adjoined to an arithmetical theory such as Q for which an appropriate Gödel numbering of sentences and proofs is available. In this case, the existence of a formula defining the function $f(x)$ in Eq. $(1')$ is guaranteed by an appropriate generalization of the Diagonal Lemma.

If we now assume that $R(A, p)$ is a decidable relation, then by an analog of the rule DEC we may conclude

$(3') \vdash \neg R(f(x) = 0, y)$

from $(2''')$. This in turn can be understood to correspond to the intermediate conclusion (3) $\neg P(\ulcorner D \urcorner)$ in the derivation of Montague's paradox.

But now note that since y was arbitrary in the foregoing reasoning, we should additionally be able to conclude by universal generalization that

$(3'') \vdash \forall y \neg R(f(x) = 0, y)$

Noting that the foregoing reasoning is also uniform in the variable x, we also ought to be able to internalize it in a manner analogous to INT. Doing so yields the existence of a function $g(x)$ such that

$(3''') \vdash R(\forall y \neg R(f(x) = 0, y), g(x))$

By substituting g for x in $(3''')$ we obtain $\vdash R(\forall y \neg R(f(g) = 0, y), g(g))$. But then again taking $x = g$ in $(1')$ and applying modus ponens yields

$(4') \vdash f(g) = 0$

which can be seen as analogous to step (4) in the derivation of Montague's paradox. Internalizing this reasoning again leads to the existence of another construction h such that

$(5') \vdash R(f(g) = 0, h)$

But now instantiating y by h in $(3'')$ finally yields $\vdash \neg R(f(g) = 0, h)$, and thus a contradiction with $(5')$.

Although we have not precisely specified the system in which the foregoing derivation is carried out, it is evident that it must satisfy a number of features. First, it must be capable of demonstrating the existence of an appropriate "self-referential" construction $f(x)$ as appears in $(1')$. Second, it must treat constructions as "self-applicable" in the sense that it makes sense to apply a construction like $f(x)$ to another construction $g(x)$. Third, the proof relation $R(A, p)$ must be understood to satisfy the analogs of EXPRFN and DEC[18] which are employed at steps $(2')$, $(2'')$, and $(3')$. Fourth, it must support the sort of first-order reasoning which stands behind the use of universal generalization and instantiation employed at steps $(3'')$, $(4')$, and $(5')$. And fifth, it must also support the use of an appropriate analog to INT applicable to reasoning mediated by all of the prior forms of reasoning about the proof relation.

Although the system $\mathscr{T}$ which we sketched in Sect. 3.1 is designed so as to satisfy the second and third of these conditions, it is not clear whether it satisfies the first, fourth, or fifth. This complicates the task of interpreting the more formal derivation of the paradox described by Goodman [17, Sect. 9] which appears to be an attempt

[18]The rule in question applied at step $(3')$ takes the form $R(A, p) \vdash \neg R(A, p) \;\therefore\; \vdash \neg R(A, p)$. Note, however, that this does not represent an additional assumption in the current setting as long as we assume that the system in which we are reasoning contains intuitionistic propositional logic. For in this case, the appeal to DEC can be replaced by the derivability of $(B \to \neg B) \to \neg B$.

to regiment the prior reasoning in a formal system similar to $\mathscr{T}$. Note, however, that although this system itself does not directly contain the Diagonal Lemma, it is still sufficient for demonstrating the existence of self-referential statements by another means.

For recall that we have defined $\mathscr{T}$ so that it includes the untyped lambda calculus in the form of the equational theory $\lambda\beta$ (see note 6). Over this theory it is possible to define so-called *fixed-point combinators*—i.e. lambda-terms Z such that for any term x, $\vdash_{\lambda\beta} Zx \equiv x(Zx)$. A well known example of such a term is the so-called *Curry combinator* $Y =_{\mathrm{df}} \lambda f.(\lambda x.f(xx))(\lambda x.f(xx))$. Goodman [17] observed that it is possible to use a similar fixed-point combinator in conjunction with the term π so as to obtain a term $t(x)$ which can be understood to express that x is not a proof of this term itself. He then proceeds to describe a derivation which can be understood as a "free-variable" variant of (1′)–(5′), in which it is again assumed that an appropriate internalization principle is available. What we present here is a simplication of this derivation which employs the combinator Y itself.

First note that although we would naturally formulate the proposition expressed by "x does not prove y" in the language of $\mathscr{T}$ as the equation $\pi yx \equiv \bot$, it can also be expressed as a term $h(y,x) =_{\mathrm{df}} \lambda y.\lambda x.(\pi yx \supset_1 \bot)$. If we now apply the Y combinator to $h(y,x)$ we get a term $Y(h(y,x))$ with only x free such that $\vdash_{\mathscr{T}} Y(h(y,x)) \equiv h(Y(h(y,x)),x)$. We may now reason in $\mathscr{T}$ as follows[19]:

(i)	$\vdash_{\mathscr{T}} Y(h(y,x)) \equiv h(Y(h(y,x)),x)$	defn. of Y
(ii)	$\pi(Y(h(y,x)))x \equiv \top \vdash_{\mathscr{T}} Y(h(y,x)) \equiv \top$	ExpRef
(iii)	$\pi(Y(h(y,x)))x \equiv \top \vdash_{\mathscr{T}} h(Y(h(y,x)),x) \equiv \top$	(i), transitivity of $\equiv$
(iv)	$\pi(Y(h(y,x)))x \equiv \top \vdash_{\mathscr{T}} (\pi(Y(h(y,x)))x \supset_1 \bot) \equiv \top$	defn. of $h(y,x)$
(v)	$\pi(Y(h(y,x)))x \equiv \top \vdash_{\mathscr{T}} \bot \equiv \top$	defn. $\supset_1$
(vi)	$\vdash_{\mathscr{T}} \pi(Y(h(y,x)))x \equiv \bot$	Dec
(vii)	$\vdash_{\mathscr{T}} (\pi(Y(h(y,x)))x \supset_1 \bot) \equiv \top$	defn. $\supset_1$
(viii)	$\vdash_{\mathscr{T}} h(Y(h(y,x)),x) \equiv \top$	defn. $h(y,x)$
(ix)	$\vdash_{\mathscr{T}} Y(h(y,x)) \equiv \top$	(i), transitivity of $\equiv$

This derivation—which up to this point may be carried out in the system $\mathscr{T}$ as presented above—can again be roughly aligned with steps (1)–(4) in the derivation of Montague's paradox—e.g. the use of ExpRfn at step (ii) in the former aligns with the use of RefP at step (2) in the latter, step (vi) of the former corresponds to step (3) in the latter, etc. In order to continue the derivation, however, we need to assume that we are working over a system $\mathscr{T}^+$ which satisfies the principle Int. We may now continue the derivation as follows[20]:

[19] At step (v) we use the rule $\Delta, \pi uv \equiv \top \vdash_{\mathscr{T}} \bot \equiv \top \;\therefore\; \Delta \vdash_{\mathscr{T}} \pi uv \equiv \bot$ which can be derived from Dec and the cut rule in $\mathscr{T}$.

[20] The step analogous to (xi) in Goodman's own presentation of the paradox is (5) on p. 108 of [17]. At this point he simply writes that the relevant internalizing term "must exist" without providing any further explanation. Note also that his system includes a substitution rule of the form $\Delta \vdash u \equiv v \;\therefore\; s \equiv s, \Delta[s/x] \vdash u[s/x] \equiv v[s/x]$ where the extra premise $s \equiv s$ serves to ensure the term s is defined. Hence to bring step xi) into better conformity with Goodman's system, we should also include axioms $c \equiv c$ for the new "internalizing constants".

(x)	$\vdash_{\mathscr{T}^+} \pi(Y(h(y,x)))c \equiv \top$	INT for some new constant c
(xi)	$\vdash_{\mathscr{T}^+} \pi(Y(h(y,x)))c \equiv \bot$	substituting c for x in vi)
(xii)	$\vdash_{\mathscr{T}^+} \top \equiv \bot$	(x), (xi), transitivity of $\equiv$

Finally, we observe that it follows that the derivability of $\top \equiv \bot$ from no premises in $\mathscr{T}^+$ entails that all equations are derivable from no premises in this system. But this is precisely how inconsistency is traditionally defined for systems based on the lambda calculus.

4 The Reception of the Theory of Constructions and the Second Clause

The foregoing derivation is carried out in the system $\mathscr{T}^+$. As we have noted, this system does not coincide with any of the variants of the Theory of Constructions explicitly adopted by Kreisel or Goodman. Nonetheless the derivation bears sufficient resemblance to that sketched by Goodman [17, pp. 107–109] so as to be a reasonable candidate for what we might call the *formalized* Kreisel-Goodman paradox. And although Goodman went on to develop $\mathscr{T}^\omega$ specifically to avoid the paradox, this initial observation about the "naive" theory we have been discussing played a substantial role in shaping subsequent opinion about the Theory of Constructions itself.

Before considering the various ways in which one might react to the paradox directly in Sect. 5, our goals in this section will be twofold. First, we will briefly describe the manner in which the conventional wisdom about the significance of the Theory of Constructions shifted during the 1970s and 1980s. Second, we will argue that several of the criticisms which have been directed against the theory appear to be based on misapprehensions about its relationship to the second clause and to the Kreisel-Goodman paradox.

4.1 Shifting Opinions

The shift in the consensus about the status of the Theory of Constructions can be readily appreciated by comparing the following passages taken respectively from the prefaces of the first (1977) and second (2000) edition of Dummett's *Elements of Intuitionism*:

> The mathematical theory of constructions is of the greatest importance for the foundations of intuitionistic logic, and it was with greatest regret that I omitted all but a mention of its existence; but it is as yet in an imperfect state, and its formulation is far too complicated to permit of a brief summary [7, p. viii].

> In the original Preface I mentioned with enthusiasm the theory of constructions inaugurated by Kreisel, aimed at supplying a canonical semantics for intuitionistic logic; unfortunately, it did not prove fruitful [7, p. iv].

Although Dummett provides no further explanation for this change of heart, his reaction echoes that of other theorists who, in the intervening years, had come to conclude that the Theory of Constructions not only did not live up to Kreisel's promise of providing a "semantical foundation" for intuitionistic logic, but was also ill-motivated because of its association with the second clause. As we are now in a good position to appreciate, however, the formulation of a theory such as $\mathscr{T}$ is independent of how (or even if) we elect to attempt to use its object language to formalize the BHK clauses. And as such, it seems that criticisms of the Theory of Constructions which are grounded in objections to the propriety of adopting the second clause are likely to be off base.

Putting this observation to the side for the moment, it is also possible to identify two broad classes of criticisms which have been targeted at the second clause itself. The first of these is that the transition from (e.g.) ($\mathrm{P}_{\rightarrow}$) to ($\mathrm{P}^2_{\rightarrow}$) either adds nothing to the original BHK interpretation or does not serve to resolve the problems which appear to have motivated Kreisel to introduce it. For instance, Girard [12] says the following:

> Since the $\rightarrow$ and $\forall$ cases were problematic (from [the …] foundational point of view), it has been proposed to add to ($\mathrm{P}_{\rightarrow}$) [. . .] the codicil "together with a proof that f has this property". Of course that settles nothing, and the Byzantine discussions about the meaning which would have to be given to this codicil—discussions without the least mathematical content—only serve to discredit an idea which, we repeat, is one of the cornerstones in Logic [12, p. 7].

Although Girard does not comment further on the claim that the second clause is "without mathematical content", several subsequent commentators appear to expand on his point that it leads to a substantial complication in how we should understand the meaning of implication. For instance Prawitz writes

> One may ask whether [what is known in understanding an implication] should not consist of a description of the procedure together with a proof that this procedure has the property required, as suggested originally by Kreisel [25]. But this would lead to an infinite regress and would defeat the whole project of a theory of meaning as discussed here [35, p. 27]

Such passages suggest that far from overcoming the apparent deficiency in the original BHK account of intuitionistic implication—i.e. that it requires that we understand what it means to quantify over *all* constructive proofs—the second clause in fact makes matters worse in the sense of introducing another kind of infinitary condition as part of its meaning.

Prawitz also does not expand on what he means by speaking of an "infinite regress". But one interpretation is that he too is making the point that in order to formulate the second clause, we must allow for the fact that it makes sense to think of the proof relation as holding between a proof p and a sentence A which may itself make reference to this relation (and thus to other proofs and formulas). If it is acknowledged that this is legitimate, then there seems to be nothing to prohibit

arbitrary iterations of the proof relation. For instance, if we continue to use $R(x, y)$ to denote this relation, then an example of the sort of "regress" Prawitz appears to have in mind might correspond to the existence of a statement A and proofs $p_1, p_2, p_3, \ldots$ such that $R(A, p_1), R(p_2, R(A, p_2)), R(p_3, R(p_2, R(A, p_2))), \ldots$ It is evident that the *syntax* of the Theory of Constructions allows us to express the existence of such a sequence in the sense that $\pi t s_1 \equiv \top, \pi(\pi t s_1)s_2 \equiv \top, \pi(\pi(\pi t s_1)s_2)s_3 \equiv \top, \ldots$ are all well-formed formulas.

One might reasonably wonder on this basis if grasping the second clause interpretation of a formula ever requires that we grasp such an infinite sequence of conditions. Beeson discusses a related point:

> Is it necessary to include [the second] clause? What does it really mean? At one extreme is the view that one should simply delete this clause: a constructive proof should contain the information a computer needs to verify the computational facts […] At the other extreme is the view that the "supplementary data" is a *proof* itself: a proof that q does indeed transform any proof of A into a proof of B. The difficulties with this view seem to be that (i) it makes the explanation of proof highly impredicative, destroying any hope of explaining proofs of complicated propositions in terms of proofs of simpler ones; and (ii) it seems to assume that "p is a proof of A" is a mathematical proposition "on the same level as" A itself, in particular, capable of being proved mathematically." [3, p. 402]

We will come back to discuss the second concern described by Beeson—i.e. that it assumes that *p is a proof of A* expresses a mathematical proposition "on the same level" as expressed by A itself—in the course of comparing the Theory of Constructions to systems like ITT (wherein *p is a proof of A* is regarded as a *judgement* as opposed to a proposition). But with regard to the first issue he raises, note that while Kreisel appears to have introduced the second clause precisely so as to avoid the form of impredicativity discussed in Sect. 2, Beeson suggests that it is this step itself which introduces impredicativity into the interpretation of intuitionistic implication.

Although Beeson also fails to expand upon the precise form this impredicativity takes, it again seems likely that what he also has in mind has something to do with the self-applicability of the proof relation. For note that not only does the formulation of the second clause require that we countenance the existence of proofs p which stand in the proof relation to statements A which may themselves refer to other particular proofs q (e.g. for A of the form $R(B, q)$), but also the case where A may contain a quantifier over all proofs (e.g. for A of the form $\forall x R(B(x), x)$), presumably inclusive of p itself.

A potentially related point about the existence of proofs with this property is made by Weinstein in the following remark about the second clause:

> If […] we suppose that universal quantifications over the universe of constructions applied to decidable properties have decidable proof conditions then we may view [($\mathrm{P}^2_{\rightarrow}$)] as providing an assignment of decidable proof conditions to each formula of the language of arithmetic […¶…] This means of securing the decidability of the proof conditions for formulas of arithmetic is not without cost. The alternative statement of the proof conditions for conditionals is self reflexive in a way that the original explanation was not. Both Kreisel and Goodman noticed that this self reflexivity leads to paradox in a theory of constructions which includes a

> reflection principle for the primitive which constructs the proof conditions for quantification over the universe of constructions applied to decidable properties [49, p. 264].

Rather than simply suggesting that the second clause is ill-motivated in virtue of leading to the sort of infinitary or impredicative proof condition mentioned by Prawitz or Beeson, Weinstein goes beyond this and suggests that it leads to a form of "self-reflexivity" which in turn is responsible for the Kreisel-Goodman paradox. It is this claim which we will focus on in the next section.

4.2 Guilt by Association?

The passages collected in the prior section make clear that not only have most commentators reacted negatively to Kreisel's proposed modifications to the clauses ($P_{\rightarrow}$), ($P_{\neg}$), and ($P_{\forall}$), but also that this reaction has contributed to their assessment of the Theory of Constructions itself. Against this backdrop, we now wish to frame two observations: (1) the second clause interpretations of the intuitionistic connectives play no role in the formulation of the Theory of Constructions itself—rather the theory merely provides a formal language in which these interpretations can be expressed; (2) the Kreisel-Goodman paradox also does not arise in virtue of assigning the connectives appearing in its premises their second clause interpretations.

The first point may be appreciated by simply recalling that variants of the Theory of Constructions like $\mathscr{T}$ are indeed "logic free" in the sense that they do not contain logical connectives such as $\rightarrow, \neg$ or $\forall$ amongst their primitive symbols. Rather such systems contain other primitives—e.g. the abstraction operator λ and the proof operator π—which Kreisel and Goodman hoped to show are sufficient for analyzing the meaning of the intuitionistic connectives. As we have seen, these analyses take the form of providing a definition of a predicate $\Pi(A, s)$ which they suggest can be understood as formalizing the second clause variants of the traditional BHK clauses.

Only once such a definition has been undertaken may we ask whether the *defined* proof relation $\Pi(A, s)$ has certain properties such as decidability. But as we are now in a good position to appreciate, such features apply to the "internal logic" of a theory which is being interpreted in a system like $\mathscr{T}$ and not to the formal properties of the Theory of Constructions itself.[21] But from this it also follows that since the second clause variants of ($P_{\rightarrow}$), ($P_{\neg}$), and ($P_{\forall}$) are conditions which we attempt to *interpret* in $\mathscr{T}$, they are no more an intrinsic feature of such a system than is the decision to interpret the natural numbers as finite von Neumann ordinals an intrinsic feature of ZF set theory.

[21] Among subsequent commentators on the Theory of Constructions, Troelstra [43] presents a version of the theory in which $\Pi(A, s)$ is itself treated as a primitive notion, whereas Sundholm [39, 40] (while clearly aware of the technical distinction between πst and $\Pi(A, s)$) continues to speak of properties like decidability as features which might be *stipulated* (rather than *proven*) to hold for $\Pi(A, s)$.

It is also evident that the second clause plays no direct role itself in the formulation of the Kreisel-Goodman paradox as discussed in Sect. 3.2. One indication of this is that although our proposed regimentation of Goodman's informal description of the paradox is conducted in first-order logic, no special treatment is accorded to the connectives $\rightarrow$, $\neg$ or $\forall$. Similarly, when we attempt to mimic this reasoning in $\mathscr{T}^+$, it is evident that the derivation of a contradiction does not require that we interpret the occurrences of the logical connectives occuring in the semi-formal version in accordance with the second clause interpretations ($K_\rightarrow$), ($K_\neg$), or ($K_\forall$).

From this it would appear to follow that Weinstein is unjustified in at least his contention that the Kreisel-Goodman paradox is directly engendered by reasoning with the intuitionistic connectives relative to their second clause interpretations. What remains to be seen, however, is whether it is possible to sustain what appears to be his more general point—i.e. that the paradox reveals that any attempt to formalize the clauses ($P^2_\rightarrow$), ($P^2_\neg$), and ($P^2_\forall$) will result in a system which is inconsistent in virtue of being "self-reflexive".

In evaluating this claim, it seems possible to interpret the relevant notion of "self-reflexivity" in one of three ways which we will respectively label "self-applicability", "self-dependency", and "self-referentiality". We have already considered a sense in which the Theory of Constructions formalizes a notion of "self-applicable" proof in the sense that it allows for iterations of the proof operation in expressions such as $\pi(\pi s t_1)t_2 \equiv \top$. But on its own, this property does not seem to lead obviously to any sort of antinomy about the proof relation $R(x, y)$. Some evidence of this is provided by the fact that it is not only consistent with familiar systems T of formal arithmetic that there exist statements A and pairs of numbers n, m such that $\mathtt{Proof}_{\mathsf{T}}(\mathtt{n}, \ulcorner \mathtt{Proof}_{\mathsf{T}}(\mathtt{m}, \ulcorner A \urcorner) \urcorner)$, but instances of such statements will typically be provable in T itself.[22]

The foregoing example pertains only to self-applicability in the general sense that the proof relation $R(x, y)$ is allowed to hold of a sentence A and a proof s in the case where A itself contains $R(B, t)$ for some sentence B and proof t. But although it would seem that this is all that is needed for the formulation of the second clause, it might also be thought that the Kreisel-Goodman paradox turns on the existence of proofs which are "self-dependent" in the sense that their definitions rely on the fact that they must be understood to already exist. An example would be witnessed by the existence of a statement D and a proof u such that $R(R(D, u), u)$, whose truth would appear to entail that u is self-dependent in the sense that the statement proven by u refers to u itself.

[22]For instance in the case where $\mathsf{T} \vdash A$, the existence of n and m such that $\mathsf{T} \vdash \mathtt{Proof}_{\mathsf{T}}(\mathtt{n}, \ulcorner \mathtt{Proof}_{\mathsf{T}}(\mathtt{m}, \ulcorner A \urcorner) \urcorner)$ is a straightforward consequence of the first and third Hilbert-Bernays derivability conditions for $\mathtt{Proof}_{\mathsf{T}}(x, y)$.

In his second exposition, Goodman appears to attribute the paradox to the existence of proofs with this property:

> There is an essential impredicativity in our definition of implication. For $[\Pi(A \to B, y)]$ involves quantification over all proofs of A, including proofs which may themselves have been built up in some way from y. Unless something is done to moderate this impredicativity, it actually leads to paradox [17, p. 107].

Goodman says this after defining $\Pi(A \to B, y)$—i.e. the proof condition for the implication $A \to B$—in the same manner as Kreisel's second clause variant ($\mathrm{K}_{\to}$). It is notable, however, that in our reconstruction of Goodman's formulation of the paradox the derivation of an inconsistency depends on our ability to construct in $\mathscr{T}$ a term $Y(h(y, x))$ which functions in a manner analogous to the *formula D* in the derivation of Montague's paradox. But although this statement is indeed self-referential in the traditional sense of being provably equivalent to its own unprovability, it does not depend on the existence of a *proof* which is self-dependent in the sense just described.[23]

Note finally that we have already seen in Sect. 2 that the concerns which Goodman raises about the impredicativity of implication appear to already arise for the original BHK clause ($\mathrm{P}_{\to}$). As we suggested there, this may indeed highlight an important conceptual problem about how the notion of constructive proof should be understood. It seems, however, that the sort of "self reflexivity" which engenders the Kreisel-Goodman paradox is more closely related to traditional forms of self-reference which figure in classical inconsistency results like Montague's paradox. And this in turn suggests that not only is the paradox not engendered by the second clause in the direct sense of requiring that we interpret the logical notions which figure in its derivation in accordance with ($\mathrm{P}^2_{\to}$), ($\mathrm{P}^2_{\neg}$), and ($\mathrm{P}^2_{\forall}$), but also that it is not engendered indirectly by introducing an impredicative element into the concept of constructive proof which was not already present.

5 Diagnosing the Paradox

Our aim in the prior section was to argue that the ultimate evaluation of both the second clause and the Theory of Constructions should be separated from the task of diagnosing and responding to the Kreisel-Goodman paradox. For not only does the adoption of the Theory of Constructions not necessitate that we interpret the intuitionistic connectives using the second clause, but also the inconsistency of the

[23]The same is also true of Goodman's own derivation of the paradox in [17] in the following exact sense. First note that Goodman's proof is based on applying a fixed-point combinator to the term $h'(y, x) = \lambda y.\lambda x.\pi(\pi yx \supset_1 \bot)(xx)$. As this term *does* contain an iterated application of π, a plausible interpretation is that it is derived from attempting to express "x is not a proof of y" within the language of $\mathscr{T}$ relative to its second-clause interpretation. However, the fixed-point which Goodman employs in his derivation is still obtained for the variable y and not x—i.e. it too can be understood as formalizing the existence of a self-referential *formula* as opposed to a self-dependent *proof*.

"naive" variant $\mathscr{T}^+$ turns on assumptions which are independent of the suitability of its language for expressing the second clause.

Once these points are acknowledged, a number of other questions naturally arise: (1) having eliminated the second clause as the direct source of the Kreisel-Goodman paradox, what other principles might be to blame? (2) was Goodman correct to conclude the most appropriate response to the paradox was to conceive of the universe of constructive proofs as stratified in the manner described by his theory $\mathscr{T}^\omega$? (3) what is the status of his [16] proofs of consistency, soundness, faithfulness, and the interpretatibility of Heyting arithmetic for $\mathscr{T}^\omega$? (4) are such results available for unstratified variants of $\mathscr{T}^\omega$? and (5) might such systems be of independent conceptual or technical interest?

A truly systematic exploration of these issues is beyond the scope of the current paper. What we hope to achieve here is the more modest goal of laying out the various principles on which the paradox appears to depend and assessing them relative to the goal of providing an "informally rigorous" account of the BHK interpretation of the sort envisioned by Kreisel and Goodman.

5.1 Self-Reference and Typing

As we have seen in Sect. 3.2, one of the principles on which the Kreisel-Goodman paradox relies is the existence of terms t containing the operator π satisfying fixed-point equations of the form $Y(t) \equiv t(Y(t))$. As we have suggested in Sect. 4.2, such terms can be understood to play a role analogous to that of self-referential sentences in traditional formulations of the semantic (or "intensional") paradoxes such as the Liar and Montague's paradox. In particular, the term $Y(h(y, x))$ can be understood to express that x is not a proof of $Y(h(y, x))$ itself.

Whereas the existence of sentences with similar intended interpretations is guaranteed in the classical setting via the arithmetization of syntax and the Diagonal Lemma, the existence of $Y(h(y, x))$ is a consequence of the existence of fixed-point combinators like Y for the system $\lambda\beta$. But although these phenomena may themselves be understood to share a common basis (cf., e.g., [2, Sect. 6.7]), the question also naturally arises why we ought to base a formulation of the Theory of Constructions on a form of the lambda calculus for which such combinators may be shown to exist.

Part of the answer to this may be understood to follow from the goal of using the language of the Theory of Constructions to formalize the clauses of the BHK (or BHK^2) interpretation. For note that it is now a familiar observation that the notions of function abstraction and application which form the basis of the lambda calculus also appear to be implicit in the BHK clauses. This point is often illustrated by pointing out that if the formulas appearing in the rules of a traditional natural deduction system for first-order intuitionistic logic are labeled with terms understood to represent their proofs, then the implication introduction rule can be understood to correspond to a form of function abstraction on proofs similar to the one which is implicit in ($\mathrm{P}_\rightarrow$).

Similarly, the implication elimination rule can be understood to correspond to a form of function application on proofs.[24]

Such observations provide a strong basis for including the lambda calculus as part of the primitive machinery in terms of which we might attempt to formalize the BHK interpretation. It would seem, however, that the interpretation itself does not nominate a *unique* form of the calculus to serve in this capacity. For as Sørensen & Urzyczyn note:

> [N]ot every lambda-term can be used as a proof notation. For instance, the self-application xx does not represent any propositional proof, no matter what the assumption annotated by x is. So before exploring the analogy between proofs and terms … we must look for the appropriate subsystem of the lambda-calculus [38, p. 56].

Such observations are often cited as the basis of the Curry-Howard isomorphism which relates logical systems with various *typed* lambda calculi (such as the simply typed Church-style system $\lambda\beta^{\rightarrow}$ of [22]). This in turn provides the basis for the interpretation of intuitionistic logic which is provided by systems such as Martin-Löf's ITT.[25]

However $\lambda\beta^{\rightarrow}$ can also be distinguished from the system $\lambda\beta$ on which we have taken $\mathscr{T}$ to be based in virtue of the fact that the latter allows not only for self-application of terms (e.g. xx), but also for the definition of fixed-point combinators like Y. The potential significance of this point with respect to the status of the Kreisel-Goodman paradox should now be clear—i.e. although it seems reasonable to base a formal theory in which we might seek to interpret the BHK clauses on *some* form of lambda calculus, not only does the informal presentation of the clauses fail to pick out a unique system, but there is also reason to suspect that $\lambda\beta$ allows for the definition of terms which are not needed for the interpretation of proofs in intuitionistic logic.

Unlike Goodman, Kreisel does not explicitly formulate a paradox as an obstacle to formulating a "naive" variant of the Theory of Constructions. It seems likely,

---

[24] For instance, by adapting the example of [38, pp. 55–56] to the notation of the semi-formal system of Sect. 3.2, we can see that the "labeled" versions of the rules $\rightarrow$-Intro and $\rightarrow$-Elim take the forms

$$\frac{\begin{array}{c}[R(A,x)]\\ \vdots \\ R(B,s_1(x))\end{array}}{R(A\rightarrow B,s_2)} \qquad \frac{R(A\rightarrow B,t_1)\quad R(A,t_2)}{R(B,t_3)}$$

where s_2 is naturally understood as having the form $\lambda x.s_2(x)$ and t_3 is naturally understood as having the form t_1t_2.

[25] Martin-Löf [30] cites the Theory of Constructions as one of several earlier systems which anticipated his development of ITT. It is indeed clear that there is an affinity between the manner in which the two systems define embeddings of intuitionistic logic into variants of the lambda calculus whose constituent clauses are intended to resemble those of the BHK interpretation. (Another historical affinity derives from the fact that Martin-Löf [29] presents ITT as a "predicative" reformulation of the system of [28] which was found to be inconsistent in virtue of Girard's paradox.) An important difference, however, is that constructive proofs are only represented indirectly in ITT as typed. But as typing judgements may not be iterated in ITT in the manner of the π operator, there is no evident manner in which the language of Martin-Löf's system can be used to express the second clause interpretations of $\rightarrow$, $\neg$, and $\forall$.

however, that he was aware of the foregoing observations. For in his first formulation of the theory [25, p. 203] Kreisel explicitly restricts lambda abstraction to the class of terms which are asserted by the axioms of the theory to be *bivalent* in the sense described above. His second formulation [26, pp. 128–129] of the theory is based on a form of typed lambda calculus similar to $\lambda\beta^{\rightarrow}$. Both approaches thus have the effect of prohibiting $Y(h(y, x))$ from being a well-formed term of the system in question. As such, Kreisel's apparent reaction to the threat of a paradox pertaining to the notion of construction can be compared both with Russell's [36] reaction to the set theoretic paradoxes and Tarski's [42] reaction to the semantic paradoxes—i.e. the existence of the offending self-referential entities (i.e. sets, formulas, or terms) are excluded on the basis of being improperly formed.

5.2 Stratification

Goodman's reaction to the paradox was guided by his view that an adequate foundation for intuitionistic logic must presuppose neither logic nor a doctrine of types. He thus proposed to retain the untyped lambda calculus as the basis of the Theory of Constructions and at the same time conceive of constructions as stratified into "levels" which he likens to set theoretic ranks.[26] Thus while we have just seen that Kreisel's reaction to the "self-referential" paradox about provability was at least superficially similar to Russell's resolution to the set theoretic paradoxes, Goodman explicitly suggests that his proposed resolution can be understood as analogous to that of Zermelo [50]:

> The set-theoretic paradoxes are resolved by observing that sets must be sets of objects already at hand. Similarly we suggest that proofs must be *about* objects already constructed. Just as in Zermelo set theory there is an implicit cumulative theory of types, so we propose to formulate a theory of constructions involving a cumulative theory of *levels*. At the bottom level we will have constructive rules operating on each other ... Given any level L, we suppose that we can extend L to a new level containing all the objects of L, all proofs about objects of L, and certain additional constructions to be described below ... We emphasize that this is not a stratification by logical type, but rather a stratification according to the subject matter of proofs [17, p. 109].

In outline, Goodman proposes to implement this proposal by defining a "stratified" version of the Theory of Constructions $\mathscr{T}^{\omega}$ with the following features: (1) the untyped lambda calculus $\lambda\beta$ is retained, as well as the possibility that terms may

[26] Goodman's other apparent reason for employing the untyped lambda calculus in formulation of $\mathscr{T}^{\omega}$ pertains to his desire to use the system for interpreting Heyting arithmetic. In particular, in order to define the natural numbers in the language of $\mathscr{T}^{\omega}$, he first uses the pairing functions to define $\mathtt{0} = \lambda x.\lambda y.x$, and $\mathtt{n} + \mathtt{1} = D\mathtt{n0}$. He then shows that it is possible to use a fixed point combinator similar to Y in order to define a decidable natural number predicate. Goodman's foundational goals are thus somewhat more ambitious than those of (e.g.) Martin-Löf [29] in the sense that he hoped to reduce not only intuitionistic logic, but also intuitionistic arithmetic to a primitive theory of constructions which does not itself contain a basic natural number type.

be undefined, identity is to be understood intensionally, etc.; (2) the notion of a so-called *grasped domain* of constructions is introduced to play the role of a *level* in the stratified hierarchy of constructions as just described[27]; (3) such levels are understood as proceeding from a *basic level* $B =_{\mathrm{df}} L_0$ and forming a hierarchy $L_0 \subseteq L_1 \subseteq L_2 \subseteq \ldots$ over which the variables of $\mathscr{T}^\omega$ are intended to range; (4) various primitive terms are introduced into the language of $\mathscr{T}^\omega$ to formalize this conception (e.g. Bx iff x is a basic level construction, Gx iff x is a grasped domain, Exy iff y is the grasped domain corresponding to the level extending x, etc.) together with axioms which ensure that they have various intended properties such as decidability; (5) the binary proof operator of πxy of the system $\mathscr{T}$ is replaced with a ternary proof operator $\pi^3 xyz$ with the intended interpretation "x is a grasped domain containing y, and z is a proof that $yw \equiv \top$ for all w in x".

Goodman's proposed resolution to the paradox [16, pp. 111–112] may be understood as turning on the following observations: (a) for each level L_n it is possible to formulate a term t_n akin to $Y(h(y, x))$ which may be interpreted as expressing its own unprovability by all constructions at level n; (b) although it is still possible to reach a conclusion analogous to (ix) in the original demonstration expressing that such a term is true (i.e. $\mathscr{T} \vdash t_n \equiv \top$), proving this statement involves reasoning with a free variable over L_n; (c) if we let c_n denote this derivation, Goodman's rules for grasped domains only allow us to show that c_n is in L_{n+1}, but not L_n; (d) as such, no contradiction arises since c_n is not in the range of the implicit universal quantifier over proofs which are asserted by $t_n \equiv \top$ to not be proofs of t_n.

Needless to say, the fact that we cannot derive a formal contradiction in $\mathscr{T}^\omega$ in this manner does not itself constitute a proof that the system is consistent. For this reason, much of [16] is taken up with providing a formal consistency proof for $\mathscr{T}^\omega$. However, the details of Goodman's proof of this are complex. And thus rather than commenting further on this feature of $\mathscr{T}^\omega$, we offer the following general observations about the role he took this theory to have in resolving the paradox.

First, note that it is evident that the transition from $\mathscr{T}$ to $\mathscr{T}^\omega$ is purchased at the cost of a substantial complication not only of the class of primitive operations and relations on constructive proofs which must be adopted (of which we have mentioned only a few), but also with respect to the axiomatic principles which must be assumed to hold of them to correctly describe the relationship between the levels in the stratified hierarchy of constructions which is the intended model of Goodman's theory. It would seem, however, that if we wish to provide an "informally rigorous" account of why $\mathscr{T}^\omega$ is indeed the appropriate formal system with which to achieve Kreisel and Goodman's goal of providing a semantic foundation for intuitionistic logic, then each of these principles must be individually justified in terms of the network of pre-theoretical notions which figure in the BHK interpretation itself. However, it is unclear whether it is possible to do so in all of the relevant cases.[28]

[27] Goodman [17, pp. 109–110] describes such a domain as the class of constructions which has been "grasped as a totality" and which is *maximal* in the sense of "including everything which is understood when its elements are understood".

[28] Especially problematic in this regard is the inclusion in $\mathscr{T}^\omega$ of a so-called *reducibility operator* F. Roughly speaking, F is supposed to achieve the role of reducing a "noncanonical" proof of an

Second, one might reasonably question the basis of Goodman's claim that the stratification of the universe of constructions is a matter of "the subject matter of proofs" as opposed to one of "logical type". For on the one hand, while the basis of Goodman's original contention that a foundation for intuitionistic logic must itself be type-free presumably derives from the observation that the notion of type does not explicitly figure in the original expositions of the BHK interpretation, it is equally evident that these expositions also do not contain any explicit reference to a stratification of constructive proofs into levels resembling set theoretic ranks.[29] And on the other hand, one consequence of Goodman's introduction of the ternary proof operator is to allow us to conclude that $\pi^3 stu \equiv \bot$ whenever it may be shown that the proof u is not in Es (i.e. the grasped domain formed by extending s). Thus although statements of the exhibited sort are still treated as syntactically well-formed, they are simply stipulated to be false whenever an appropriate containment relation fails to hold between levels and proofs. And thus although $\mathscr{T}^\omega$ does not contain the formal machinery of type judgements, the effect of typing seems to be implicitly enforced by other means.

5.3 Decidability

Although the strategies of Kreisel and Goodman may be sufficient for obtaining a consistent version of the Theory of Constructions, their approaches are not clearly grounded in considerations which follow directly from the BHK interpretation itself. As such, it seems reasonable to consider the status of the other principles which figure in the Kreisel-Goodman paradox. We will begin by considering the role of the decidability of the proof relation.

As we have seen, this is formalized within the system $\mathscr{T}$ by the rule Dec, which may in turn be understood to ensure that terms of the πst are always defined.[30] We

assertion to the objects pertaining to some level L_n in the hierarchy of constructions (i.e. one which might make reference to proofs of yet higher level) to a proof which is present at level L_{n+1}. Such an assumption plays an important instrumental role in Goodman's formulation of the clause ($\mathrm{P}^2_{\to}$) in $\mathscr{T}^\omega$ as it allows him to replace the quantifier over *all* constructive proofs with one which only ranges over the level one higher than that of the term interpreting $A \to B$. To justify this he writes "It seems to us essential to the intuitionistic position that given a fixed assertion A about a well-defined domain, there is an *a priori* upper bound to the complexity of possible proofs of A" [17, p. 111]. But as Weinstein [49] observes, it is not at all clear whether there is anything implicit in the BHK interpretation itself which justifies this assumption.

[29] This is at least true of the formulations given by Heyting [20, pp. 13–15] and Kolmogorov [24, pp. 329–330]. Martin-Löf [30, p. 128] claims that typing is already implicit in clause ($\mathrm{P}_{\to}$) if we additionally accept that every function must have a type as its domain. But it is unclear what necessitates that we adopt such an assumption.

[30] For reasons discussed in footnote 18 the same effect is also formally achieved by either reasoning about the proof relation in intuitionistic first-order logic or by adopting Kreisel's [26] proposal to base the Theory of Constructions on the calculus $\lambda\beta^{\to}$ (wherein all terms always reduce to normal form).

have also seen that the informal motivation for including such a principle derives from the desire to ensure that the relation between constructive proofs and theorems is decidable so as to in turn make available the sort of epistemic account of truth described in Sect. 2. But finally, we have seen that Kreisel introduced the second clause interpretations precisely so as to ensure that the defined proof relation $\Pi(A, s)$ introduced in Sect. 3.1 is decidable (provided that appropriate assumptions are made about the atomic case, this does indeed follow from the decidability of πst by a straightforward induction on its definition).

These considerations notwithstanding, Beeson [3, pp. 404–410] has argued against the propriety of including a rule like DEC in a version of the theory of constructions as follows: (1) he first formulates a formal inconsistency result for a system similar to $\mathscr{T}^+$; (2) he then argues that this result can be understood as a *reductio* of DEC. But since he also advocates for the inclusion of second clauses on the interpretation of $\rightarrow$, $\neg$, and $\forall$, his overall motivation for rejecting decidability appears somewhat incongruous.[31] As such, we will henceforth assume that giving up the rule DEC does not correspond to a well motivated response to the paradox.

5.4 Reflection

The explicit reflection principle EXPRFN formalizes the principle that if p is a construction proving A, then A is true. Like decidability, such a principle may plausibly be regarded as part of the intended interpretation of the proof relation. To the best of our knowledge, no one has ever argued explicitly that EXPRFN should be given up in the face of the Kreisel-Goodman paradox.[32] But although we do not wish to challenge this consensus, we will now adduce several considerations which suggest that finding an appropriate formulation of reflection in the Theory of Constructions may not be as straightforward as it might appear.

The central difficulty is most readily appreciated by again invoking the analogy between the proof relation $R(A, p)$ and the arithmetical proof predicate $\mathtt{Proof}_{\mathsf{T}}(x, y)$. If we continue to assume that the system in terms of which we reason about the former contains intuitionistic first-order logic, than one might at first think that the relevant analogs of EXPRFN would take the forms

(EXPRFNR) $R(A, p) \rightarrow A$

(EXPRFNPR$_{\mathsf{T}}$) $\mathtt{Proof}_{\mathsf{T}}(\mathtt{n}, \ulcorner\phi\urcorner) \rightarrow \phi$

[31] A similar reaction is voiced by Sundholm [40, p. 16]: "Since [the second clauses] had been introduced by Kreisel solely to guarantee that decidability, I found Beeson's theory lacking proper motivation as well as wanting in simplicity".

[32] A partial exception to this is Kreisel who, after observing that EXPRFN is "obvious on the intended interpretation" excludes this principle from his official "unstarred" theory. Although he does so on the basis of his other observation that EXPRFN is "troublesome for the consistency proof" [25, p. 204], he does not offer further non-instrumental justification for this.

Here $\mathtt{n}$ should be understood as abbreviating a numeral of the form $\mathtt{s}^n(\mathtt{0})$ for some fixed $n \in \mathbb{N}$, which may in turn be understood as the Gödel number of a proof in T. And on this model, it seems reasonable to think of p in A as abbreviating some (possibly complex) closed term in the language of $\mathscr{T}$ (or a similar theory) which is intended to denote a particular constructive proof.

Note, however, that the principle ExpRfn which is used in the derivation of the Kreisel-Goodman paradox differs from ExpRfnR and ExpRfnPr not only in that it is formulated in terms of the derivability relation $\vdash_{\mathscr{T}}$ of the Theory of Constructions, but also in that it may be used in the case where t is a *variable* of the theory.[33] But note that the free variable instances in ExpRfnR and ExpRfnPr—i.e. $R(A, x) \to A$ and $\mathtt{Proof}_{\mathsf{T}}(x, \ulcorner\phi\urcorner) \to \phi$ (where we assume $x \notin \mathrm{FV}(A)$ and $x \notin \mathrm{FV}(\phi)$)—are equivalent over intuitionistic first-order logic to the following "implicit" reflection principles:

(RfnR) $\exists x R(A, x) \to A$

(RfnPr$_{\mathsf{T}}$) $\exists x \mathtt{Proof}_{\mathsf{T}}(x, \ulcorner\phi\urcorner) \to \phi$

The contrast between ExpRfnPr$_{\mathsf{T}}$ and RfnPr is likely to be familiar: (i) all instances of ExpRfnPr are both true in the standard model of arithmetic and provable in $\mathsf{T} \supseteq \mathsf{Q}$; (ii) but while all instances of RfnPr$_{\mathsf{T}}$ are true in the standard model, in light of Löb's theorem for T, the only instances of RfnPr$_{\mathsf{T}}$ which will be provable in T (provided it is consistent) are those for which $\mathsf{T} \vdash \phi$. Moreover, although arithmetical theories $\mathsf{T} \supseteq \mathsf{Q}$ will satisfy an analog of the rule Int—i.e. if $\mathsf{T} \vdash \phi$, then $\mathsf{T} \vdash \exists x \mathtt{Proof}_{\mathsf{T}}(x, \ulcorner\phi\urcorner)$—the result of closing a theory T' which already proves all instances of RfnPr$_{\mathsf{T}'}$ will be inconsistent in light of Montague's paradox. A related observation is that not only will instances of $\exists y(\mathtt{Proof}_{\mathsf{T}}(y, \ulcorner\exists x \mathtt{Proof}_{\mathsf{T}}(x, \ulcorner\phi\urcorner) \to \phi\urcorner)$ be unprovable in T when $\mathsf{T} \nvdash \phi$, they will in fact be *false* in the standard model in light of the formalized version of Löb's theorem.

As the foregoing observations pertain to *formal* provability in the arithmetical theory T, it is not immediately clear what (if any morals) can be read off about the status of ExpRfn or RfnR on their intended interpretations.[34] What they do suggest, however, is that when the term t in ExpRfn is allowed to contain free variables, the effect of including this principle in a theory such as $\mathscr{T}$ may be closer to the effect of adding RfnR rather than ExpRfnR. For as is exemplified by the derivation of the Kreisel-Goodman paradox, the free variables of $\mathscr{T}$ (in conjunction with the relevant form of substitution principle) function very much like universally bound variables in first-order logic. And thus although the Theory of Constructions contains neither quantifiers nor implication in its object language, the instance of ExpRfn with $t = x$ can be understood as expressing *for all proofs x, if x is a proof of s, then s is true*.

[33] Moreover, inspection of the proof reveals that this is essential. For if x were not understood as free on the lefthand side of step ii), then it would be not admissible to substitute c for x at step (xi).

[34] For discussion of a related point see [31, pp. 137–138].

5.5 Internalization

The feature of the Theory of Constructions which we have yet to examine is the principle of internalization we have labeled INT. This principle has evident affinities with both the first Hilbert-Bernays condition for the arithmetical proof predicate $\mathtt{Proof}_{\mathsf{T}}(x, y)$ (i.e. if $\mathsf{T} \vdash \phi$, then $\mathsf{T} \vdash \exists x \mathtt{Proof}_{\mathsf{T}}(x, \ulcorner\phi\urcorner)$) and with the Necessitation rule of normal modal logics (i.e. if $\vdash \phi$, then $\vdash \Box A$). But such proof theoretic analogies aside, Kreisel and Goodman's motivations for including such a principle in the Theory of Constructions are at least somewhat obscure.

For instance, when rendered in the notation of the theory $\mathscr{T}$, Kreisel's original presentation of INT is as follows:

> For any sequence $\mathfrak{p}$ of sequents, $c_{\mathfrak{p}}$ is a term (if $\mathfrak{p}$ is a formal derivation in our system of $s \equiv \top$ then $c_{\mathfrak{p}}$ presents an—intuitive—proof of $s \equiv t$) [. . . ¶. . .] If $\mathfrak{p}$ is a formal derivation of $s \equiv \top$, then $\pi s c_{\mathfrak{p}} \equiv \top$ is an axiom [25, pp. 203–204].

Kreisel says nothing about how $c_{\mathfrak{p}}$ is defined relative to the derivation $\mathfrak{p}$, nor does he further elaborate on the distinction between "intuitive" proofs and formal derivations. Moreover, he does not provide any examples to justify the inclusion of an internalization principle in his system. And while Goodman provides a somewhat more straightforward presentation of internalization as a formal rule of proof, his intuitive explanation of this principle is similarly opaque.[35]

Rather than attempting to provide a direct reconstruction of Kreisel or Goodman's treatment of internalization in the Theory of Constructions, what we will now do is to present a partial reconstruction of the reasoning underlying the Kreisel-Goodman paradox using yet another system—Fitting's [9] Quantified Logic of Proofs [QLP]—for which a precise account of internalization is known to be available. QLP is an extension with first-order quantifiers over proofs of Artemov's [1] Logic of Proofs [LP], which itself may be understood as an "explicit" variant of the traditional modal logic S4 wherein instances of the operator $\Box$ are labeled with expressions similar in form to the terms of the Theory of Constructions.[36] We will present only the features of the system which are necessary to reconstruct the relevant portion of the derivation of the paradox here and refer the reader to [1, 9] for additional details.

---

[35] Goodman's formulation of the analogous rule in $\mathscr{T}^{\omega}$ [17, p. 118] is

$$\frac{\Delta, ax \equiv \top \vdash_{\mathscr{T}^{\omega}} bx \equiv \top}{\Delta, Ga \equiv \top \vdash_{\mathscr{T}^{\omega}} \pi^3 ab(\mathfrak{p}ab) \equiv \top}$$

where x is stipulated to not occur free in Δ, a or b and $\mathfrak{p}ab$ is explained as being an "infinite canonical proof of ab . . . which depends only on a and b and not on the structure of the formal proof [of $bx \equiv \top$ from $\Delta, ax \equiv \top$]" [17, p. 111]. Despite Goodman's disavowal of the relationship between $\mathfrak{p}ab$ and the relevant formal derivation in $\mathscr{T}$, we will see that it is precisely this dependency which is made explicit in the system QLP described below.

[36] Although there are many affinities between the Theory of Constructions and LP, the original inspiration for the latter is more closely related to Gödel's [15] embedding of intuitionistic propositional calculus into S4 and the "explicit" refinement thereof which he sketches in [14].

Like $\mathscr{T}$, the language of QLP contains expressions known as *proof terms* $s, t, u, \ldots$ which are intended to denote constructive proofs. These are given by the grammar

$$t := x, y, z, \ldots \mid a_i(\overrightarrow{x}) \mid \langle !t \rangle \mid \langle t \cdot t \rangle \mid \langle t + t \rangle \mid \langle (t(x)\forall x) \rangle$$

$x, y, z, \ldots$ are known as *proof variables*, and $a_1(x), a_2(x), \ldots$ as *axiom terms*, $!$, $\cdot$, $+$ and $(t(x)\forall x)$ denote *proof operations* respectively called *proof checker* (unary), *application* (binary), *sum* (binary), and *uniform verifier* (binary). Also like $\mathscr{T}$, the language of QLP contains a primitive expression intended to denote the proof relation $R(A, t)$—in particular *t is a proof of* A is expressed as $t : A$. However, unlike $\mathscr{T}$ (but like the semi-formal system of Sect. 3.2) the language of QLP contains the standard first-order connectives and quantifiers.

The axioms of QLP correspond to those of a standard Hilbert system for first-order logic (where for simplicity we regard all classical tautologies as axioms) together with the following axioms about the proof relation:

(LP1) $t : (A \to B) \to (s : A \to t \cdot s : B)$
(LP2) $t : A \to A$
(LP3) $t : A \to !t : t : A$

Among the rules of QLP are *modus ponens* and the standard formulation of the first-order universal generalization rule UG (i.e. if $\Delta \vdash_{\mathsf{QLP}} A(x)$, then $\Delta \vdash_{\mathsf{QLP}} (\forall x)A(x)$ if $x \notin \mathrm{FV}(\Delta)$). As it is a form of *modal* logic, QLP also possesses a form of the traditional Necessitation rule:

(AxNec) If B is an axiom of QLP, then $\vdash_{\mathsf{QLP}} a_B : B$ for some unstructured proof term a_B with the same free variables as B.

Note that the rule AxNec is not only similar in form to the principle Int, but can be given a justification similar to that which Kreisel gestures at above—i.e. if B is an axiom of the system, then we ought to be able to introduce a constant symbol a_B which is stipulated to bear the proof relation to B to record the thought that we regard this formula as an axiom of the system.

One of the characteristic features of both LP and QLP is that while such an internalization principle is asserted to hold for their *axioms*, it is possible to establish a parallel result for their *theorems* as a metatheorem about the system as opposed to a basic principle. In particular, we have the following:

(Lift) If $s_1 : A_1, \ldots, s_n : A_n \vdash_{\mathsf{QLP}} B$, then for some proof term t, $s_1 : A_1, \ldots, s_n : A_n \vdash_{\mathsf{QLP}} t(s_1, \ldots, s_n) : B$.

This result (which is traditionally called the *Lifting Lemma*—cf. [1, 9]) can be established by a straightforward induction on derivations. For instance, in the case of LP (which can be regarded as the quantifier-free fragment of QLP), the case where B is an axiom is handled by AxNec, and the case where B is derived from $A \to B$ and A by *modus ponens* is handled by LP1 as follows: if we assume (as induction hypotheses) that $u(\overrightarrow{x}) : A \to B$ is derivable from $\overrightarrow{s}(\overrightarrow{x}) : \Delta =_{\mathrm{df}} s_1(\overrightarrow{x}) : A_1(\overrightarrow{x}), \ldots, s_n(\overrightarrow{x}) : A_n(\overrightarrow{x})$ and $v(\overrightarrow{x}) : A$ is also derivable from

the same premises, then it follows by LP1 that $u \cdot v(\overrightarrow{x}) : B$ is also derivable from $\overrightarrow{s}(\overrightarrow{x}) : \Delta$. However, in order to extend this result to QLP, we also need to handle the case where $s_1 : A_1, \ldots, s_n : A_n \vdash_{\mathsf{QLP}} (\forall x)B(x)$ is derived from $s_1 : A_1, \ldots, s_n : A_n \vdash_{\mathsf{QLP}} B(x)$ by UG (and the appropriate free variable condition is met). This requires the adoption of an additional rule—called *explicit universal generalization* –governing the introduction of the universal verifier symbol $(\cdot\forall\cdot)$:

(EUG) If $s_1 : A_1, \ldots, s_n : A_n \vdash t(x) : B(x)$, then $s_1 : A_1, \ldots, s_n : A_n \vdash (t\forall x) : (\forall x)B(x)$, where $x \notin \mathrm{FV}(s_i : A_i)$ for $1 \leq i \leq n$.

With this machinery in place, we can now begin to record several additional observations about the role of the principle INT in the derivation of the Kreisel-Goodman paradox. Note first that whereas the terms c which are introduced by applications of INT are treated as constants in the language of $\mathscr{T}^+$, we have just seen that the terms $t(s_1, \ldots, s_n)$ which are introduced by LIFT will typically be complex functional expressions whose compositional structure represents the derivation of formula B from the premises $s_1 : A_1, \ldots, s_n : A_n$. In particular, although the derivation (i)–(xii) given in Sect. 3.2 of the Kreisel-Goodman paradox can be reconstructed (essentially) line by line in QLP, in the context of such a reconstruction, the proof term corresponding to the constant c which is introduced at step (x) will be a complex term which encodes the structure of the preceding steps (i)–(ix).

This is significant because while we have seen above that in $\mathscr{T}^+$, free variables are treated as universally bound in the derivation of Sect. 3.2), the same effect is achieved in QLP by the use of the traditional first-order quantifiers. Thus while it is the fact that variable x occurs free in the equation $Y(h(y, x)) \equiv \top$ which allows this expression to be interpreted as expressing the *unprovability* of the term $Y(h(y, x))$, the fact that a formula D has the analogous property would be expressed in QLP as $D \leftrightarrow (\forall x)\neg x : D$.[37]

In order to reach a contradiction analogous to the clash between steps (x) and (xi) in the Kreisel-Goodman paradox, an internalizing term $d(z)$ must be found such that $\vdash_{\mathsf{QLP}} d(z) : D$ and also that $\vdash_{\mathsf{QLP}} \neg d(z) : D$ (where it is assumed that $z : (D \leftrightarrow (\forall x)\neg x : D)$ in parallel to the assumption at step (i) of the original derivation).[38] However in order to construct $d(z)$ we must rely on the analog of RFNR for QLP—i.e.

(RFNQ) $(\exists x)x : A \rightarrow A$

Like $\mathscr{T}^+$, however, QLP also does not contain among its axioms an "implicit" reflection principle of this sort, but rather its "explicit" counterpart LP2. But like

[37] For as observed above, in Goodman's derivation of the paradox it is essential that we are allowed to substitute the term c for the variable x in the equation $\pi(Y(h(y, x))x \equiv \bot$ to yield $\pi(Y(h(y, x))c \equiv \bot$ (i.e. "c is a proof of the falsity of $Y(h(y, x))$"). Thus although $\mathscr{T}^+$ does not contain object language quantifiers, part of the effect of quantified reasoning is achieved by the presence of free variables and substitution in the system.

[38] Since QLP includes neither arithmetic nor the untyped lambda calculus, there is no evident means of actually proving the existence of such a z formally in the system. The relevant reconstruction of the Kreisel-Goodman paradox is hence carried out by reasoning from the assumption that $z : (D \leftrightarrow \neg(\exists x)x : D)$. See [5] for details.

ExpRfn, LP2 admits the case where t corresponds to a free variable x. And it is thus straightforward to show that RfnQ is derivable in QLP by intuitionistically valid first-order reasoning about proofs.

This, however, is not sufficient to construct the term $d(z)$ we have described above. In addition, we must show that the derivation of RfnQ we have just described can itself be internalized within QLP. This is accomplished by the following derivation:

(i)	$\vdash x : A \to A$	LP2
(ii)	$\vdash r(x) : (x : A \to A)$	AxNec
(iii)	$\vdash (r(x)\forall x) : (\forall x)(x : A \to A)$	Eug, (ii)
(iv)	$\vdash q : (\forall x)(x : A \to A) \to ((\exists x)x : A \to A)$	AxNec
(v)	$\vdash q \cdot (r(x)\forall x) : ((\exists x)x : A \to A)$	LP1, (iii), (iv)

In this derivation $r(x)$ is an axiom term internalizing the instance $x : A \to A$ of LP2, and q is an axiom term internalizing the first-order Hilbert axiom $\forall x(A(x) \to B) \to (\exists x A(x) \to B)$ where $x \notin \mathrm{FV}(B)$. The complex proof term $q \cdot (r(y)\forall y)$ then serves to internalize the relevant instance of RfnQ, which in turn must serve as a constituent in the construction of the yet more complex term $d(z)$ which figures in the derivation of the paradox.

While the existence of the internalizing constant c required in the original derivation of the Kreisel-Goodman paradox is obtained directly from the rule Int, we can now see that the term $d(z)$ required to reconstruct the reasoning of the paradox in QLP is obtained as a consequence of Lift. As we have just seen, the construction of this term depends not only on the fact that RfnQ can be derived in QLP from LP2, but also that this proof can be internalized in the system itself. In particular, since Lift differs from Int in virtue of being a metatheorem rather than a basic rule, it is also possible to inquire into the status of each of the elementary principles on which its derivability depends. And as we have observed, this requires a means of internalizing each of the basic deductive rules of QLP. If this theory is axiomatized via a Hilbert system as described here, then these correspond to the case of citing an axiom, *modus ponens*, and universal generalization. These principles are respectively internalized by AxNec, LP1, and Eug.

Upon inquiring further into the status of these principles, it is evident that LP1 can be justified on the basis of the analogy between implication elimination and function application which we have suggested is implicit in the BHK for implication. But finally taking a step towards a conceptually motivated resolution to the paradox, note that it is less clear what to say about either AxNec and (to an even greater extent) Eug. For although in the context of the Theory of Constructions it might at first seem unobjectionable to introduce a primitive constant c to record the fact that we regard a statement as a "self-evident" truth about constructive proofs (e.g. $\vdash_{\mathscr{T}} \top \equiv \top$), it is already less clear what to say about the interpretation of such a term in the case where the axiomatic principle in question contains a free variable (e.g. an instance of ExpRfn such as $\pi \bot x \vdash_{\mathscr{T}} \bot \equiv \top$).

When we move to a system like QLP wherein the sort of quantification over constructive proofs which is implicit in the use of free variables in the Theory of Constructions is made explicit, it is even less clear what to say about the justification

of the rule EUG. For it would seem that in order to be intuitively justified in concluding that a particular term $(t(x)\forall x)$ is a proof of a universally quantified statement about constructive proofs $(\forall x)A(x)$, there must be constructive justification for the fact that a proof which is uniform in x is sufficient to demonstrate that $A(x)$ holds of *all* constructive proofs simultaneously. When understood relative to the original formulation of the clause ($P_\forall$), this would appear to presuppose that we possess a means of describing the intended range of the quantifiers of a system such as QLP (or analogously for the interpretations of free variables in the Theory of Constructions).[39] And although both systems may be understood as attempting to provide a description of such a domain, what we appear to lack is an independent criterion for deciding whether they have succeeded in adequately doing so.

6 Conclusions and Further Work

In this paper we have argued for two central claims: (1) that the apparent consensus that the Kreisel-Goodman paradox is engendered by the adoption of Kreisel's second clause interpretations of $\rightarrow$, $\neg$ and $\forall$ is mistaken; and (2) that the ability of a formal system to internalize reasoning about its own proofs plays a larger role in the paradox than is customarily acknowledged. Taken in conjunction, these observations point towards the possibility of responding to the paradox by developing a system which retains as many of the features of the unstratified theory $\mathscr{T}^+$ as possible while seeking a conceptually motivated means of limiting the scope of the internalization principle INT.

The evident question is what form such a delimitation might take. Taken together with the observations we have recorded about the role of free variables and reflection principles in the paradox, one obvious proposal would be to consider subsystems of formalisms similar to QLP in which the scope of LIFT is limited by the exclusion of quantifier or substitution rules akin to EUG. Although such a proposal may be justifiable in terms of Kreisel and Goodman's original foundational goals, a variety of questions remain open: (i) is a consistency proof similar to that described by Goodman [16] available for an appropriate subsystem of $\mathscr{T}^+$? (ii) is it possible to prove the soundness and completeness of HPC in the sense of VAL for such a system? (iii) are the second clause interpretations of the intuitionistic connectives required for such a result? (iv) is it possible to formulate a version of Goodman's interpretation of Heyting arithmetic relative to the relevant system? Needless to say, these questions will have to wait for another occasion.

[39] A case in point of this was already noted by Gödel [14, p. 101] who observes that if we take $A \equiv \bot$ in the axiom LP2, then a term analogous to $(r(y)\forall y)$ in the derivation constructed above—i.e. such $\vdash_{\mathsf{QLP}} (r(y)\forall y) : (x : \bot \rightarrow \bot)$—would correspond to a consistency proof for the theory. But not only does such a proof seem too easy, it is for this reason that EUG is invalid when statements of the form $t : A$ are interpreted arithmetically as $\mathtt{Proof}_{\mathsf{T}}(\ulcorner t \urcorner, \ulcorner A \urcorner)$ (see [5] for details).

References

1. Artemov, S.N.: Explicit provability and constructive semantics. Bull. Symb. Log. **7**(1), 1–36 (2001)
2. Barendregt, H.P.: The Lambda Calculus. North Holland, Amsterdam (1984)
3. Beeson, M.: Foundations of Constructive Mathematics: Metamathematical Studies. Springer, Berlin (1985)
4. Benacerraf, P.: Mathematical truth. J. Philos. **70**(19), 661–679 (1973)
5. Dean, W.: Montague's paradox, informal provability, and explicit modal logic. Notre Dame J. Form. Log. **55**(2), 157–196 (2014)
6. Dean, W., Kurokawa, H.: The paradox of the Knower revisited. Ann. Pure Appl. Log. **165**(1), 199–224 (2014)
7. Dummett, M.: Elements of Intuitionism. Oxford University Press, Oxford (2000)
8. Feferman, S. et al. (eds.): Kurt Gödel Collected Works. Unpublished Lectures and Essays, vol. III. Oxford University Press, Oxford (1995)
9. Fitting, M.: A quantified logic of evidence. Ann. Pure Appl. Log. **152**(1–3), 67–83 (2008)
10. Fletcher, P.: Truth, Proof and Infinity: A Theory of Constructive Reasoning. Kluwer, Dordrecht (1998)
11. Gentzen, G.: The Collected Papers of Gerhard Gentzen. Studies in Logic and the Foundations of Mathematics. North-Holland, Amsterdam (1969)
12. Girard, J., Lafont, Y., Taylor, P.: Proofs and Types. Cambridge University Press, Cambridge (1989)
13. Gödel, K.: The present situation in the foundations of mathematics. In: Feferman et al. [8], pp. 36–53 (1933)
14. Gödel, K.: Lecture at Zilsel's. In: Feferman et al. [8], pp. 62–113 (1938)
15. Gödel, K.: An interpretation of the intuitionistic propositional calculus. In: Feferman, S., et al. (eds.) Kurt Gödel Collected Works, vol. I. Publications 1929–1936, pp. 301–303. Oxford University Press, Oxford (1986)
16. Goodman, N.: Intuitionistic arithmetic as a theory of constructions. Ph.D. thesis, Stanford (1968)
17. Goodman, N.: A theory of constructions equivalent to arithmetic. In: Kino, J.M.A., Vesley, R. (eds.) Intuitionism and Proof Theory, pp. 101–120. Elsevier, Amsterdam (1970)
18. Goodman, N.: The arithmetic theory of constructions. Cambridge Summer School in Mathematical Logic, pp. 274–298. Springer, Berlin (1973)
19. Heyting, A.: Die formalen Regeln der intuitionistischen Mathematik II. Sitzungsberichte der Preussischen Akademie der Wissenschaften, pp. 57–71 (1930)
20. Heyting, A.: Mathematische Grundlagenforschung: Intuitionismus. Springer, Beweistheorie (1934)
21. Heyting, A.: Intuitionism. An Introduction. North-Holland, Amsterdam (1956)
22. Hindley, J.R., Seldin, J.P.: Introduction or Combinators and Lambda Calculus, London Mathematical Society Student Texts, vol. 1. Cambridge University Press, Cambridge (1986)
23. Kaplan, D., Montague, R.: A paradox regained. Notre Dame J. Form. Log. **1**(3), 79–90 (1960)
24. Kolmogorov, A.: Zur Deutung der intuitionistischen Logik. Mathematische Zeitschrift **35**(1), 58–65 (1932)
25. Kreisel, G.: Foundations of intuitionistic logic. In: Nagel, E., Suppes, P., Tarski, A. (eds.) Logic, Methodology and Philosophy of Science, Proceedings of the 1960 International Congress, pp. 198–210. Stanford University Press, Stanford (1962)

26. Kreisel, G.: Mathematical logic. In: Saaty, T. (ed.) Lectures on Modern Mathematics, vol. III. Wiley, New York (1965)
27. Kreisel, G., Newman, M.H.A.: Luitzen Egbertus Jan Brouwer. 1881–1966. Biogr. Mem. Fellows R. Soc. **15**, 39–68 (1969)
28. Martin-Löf, P.: A theory of types. Technical report, pp. 71–3, University of Stockholm (1971)
29. Martin-Löf, P.: Intuitionistic Type Theory. Bibliopolis, Naples (1984)
30. Martin-Löf, P.: An intuitionistic theory of types. In: Sambin, G., Smith, J.M. (eds.) Twenty Five Years of Constructive Type Theory. Clarendon Press, Oxford (1998)
31. McCarty, C.: Intuitionism: an introduction to a seminar. J. Philos. Log. **12**(2), 105–149 (1983)
32. McCarty, C.: Constructive validity is nonarithmetic. J. Symb. Log. **53**(4), 1036–1041 (1988)
33. Montague, R.: Syntactical treatments of modality, with corollaries on reflexion principles and finite axiomatizability. Acta Philos. Fenn. **16**, 153–167 (1963)
34. Myhill, J.: Some remarks on the notion of proof. J. Philos. **57**, 461–471 (1960)
35. Prawitz, D.: Meaning and proofs: on the conflict between classical and intuitionistic logic. Theoria **43**(1), 2–40 (1977)
36. Russell, B.: The Principles of Mathematics. Cambridge University Press, Cambridge (1903)
37. Scott, D.: Constructive validity. Symposium on Automatic Demonstration, pp. 237–275. Springer, Berlin (1970)
38. Sørensen, M.H., Urzyczyn, P.: Lectures on the Curry-Howard Isomorphism. Elsevier Science, Philadelphia (2006)
39. Sundholm, G.: Constructions, proofs and the meaning of logical constants. J. Philos. Log. **12**(2), 151–172 (1983)
40. Sundholm, G.: Demonstrations versus proofs, being an afterword to constructions, proofs, and the meaning of the logical constants. In: der Schaar, M. (ed.) Judgement and the Epistemic Foundation of Logic, pp. 15–22. Springer, Berlin (2013)
41. Tait, W.W.: Gödel's interpretation of intuitionism. Philos. Math. **14**(2), 208–228 (2006)
42. Tarski, A.: The concept of truth in formalized languages. Logic. Semantics, Metamathematics, vol. 2, pp. 152–278. Clarendon Press, Oxford (1956)
43. Troelstra, A.S.: Principles of Intuitionism. Lecture Notes in Mathematics, vol. 95. Springer, Berlin (1969)
44. Troelstra, A.S.: Aspects of constructive mathematics. In: Barwise, J. (ed.) Handbook of Mathematical Logic, vol. 90, pp. 973–1052. Elsevier, Amsterdam (1977)
45. Troelstra, A.S.: The interplay between logic and mathematics: intuitionism. In: Agazzi, E. (ed.) Modern Logic—A Survey: Historical, Philosophical, and Mathematical Aspects of Modern Logic and Its Applications. Synthese Library, vol. 149, pp. 197–221. Reidel, Dordrecht (1980)
46. Troelstra, A.S., van Dalen, D.: Constructivism in Mathematics, An Introduction, vol. 1. North-Holland, Amsterdam (1988)
47. van Atten, M.: The development of intuitionistic logic. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (2009)
48. van Dalen, D.: Lectures on intuitionism. Cambridge Summer School in Mathematical Logic, pp. 1–94. Springer, Berlin (1973)
49. Weinstein, S.: The intended interpretation of intuitionistic logic. J. Philos. Log. **12**(2), 261–270 (1983)
50. Zermelo, E.: Über Grenzzahlen und Mengenbereiche: Neue Untersuchungen über die Grundlagen der Mengenlehre. In: Ebbinghaus, H., Kanamori, A. (eds.) Ernst Zermelo—Collected Works, pp. 390–429. Springer, Berlin (2010)

On the Paths of Categories

Kosta Došen

Abstract To determine what deductions are it does not seem sufficient to know that the premises and conclusions are propositions, or something in the field of propositions, like commands and questions. It seems equally, if not more, important to know that deductions make structures, which in mathematics we find in categories, multicategories and polycategories. It seems also important to know that deductions should be members of particular kinds of families, which is what is meant by their being in accordance with rules.

Keywords Deduction · Proposition · Command · Question · Category · Multicategory · Polycategory · Rule · Natural transformation · Proof-theoretic semantics · General proof theory · Categorial proof theory

1 Functions of Language

In a terminology like that of the old logic, the notion of deduction will be for us primarily a hypothetical and not a categorical notion. (This use of *categorical* should not be confused with *categorial*, which is found later in this paper, and which, according to the *Oxford English Dictionary* [23], means "relating to, or involving, categories"; unfortunately, in mathematical category theory *categorical* dominates in the sense of *categorial*.) The distinction between categorical and hypothetical is found when we speak about categorical and hypothetical proofs. The latter is a proof under hypotheses, while the former depends on no hypothesis. Both may involve deduction, but we will be concerned here with deduction as found in hypothetical proofs.

Schroeder-Heister (together with P. Contu in [22], Sect. 4, in [20], Sect. 3, and in [21]; see also [8]) states that the reigning semantics—both classical semantics based on model theory and constructivist proof-theoretic semantics—is based on *dogmas*, the main one of which may be formulated succinctly by saying that categorical

K. Došen (✉)
Faculty of Philosophy, University of Belgrade and Mathematical Institute, Serbian Academy of Sciences and Arts, Knez Mihailova 36, p.f. 367, 11001 Belgrade, Serbia
e-mail: kosta@mi.sanu.ac.rs

T. Piecha and P. Schroeder-Heister (eds.), *Advances in Proof-Theoretic Semantics*, Trends in Logic 43, DOI 10.1007/978-3-319-22686-6_4

notions have primacy over hypothetical notions. We conform to this dogma when we take the notion of proposition, a categorical notion, to have primacy over the notion of deduction, a hypothetical notion. We conform to the same dogma when we take that, among functions of language, asserting, which is tied to propositions, is more basic than deducing.

The question for us here should not be what function of language is the most important in general, but what function of language is the most important for logic. Even if it were the case that asserting is the most important function of language in general, it could happen that, because of the specific goals it has, logic, though it takes into account the importance of asserting, gives precedence to the function of deducing. Even if asserting is the most important function of language in general, for a specific area another function may have precedence. In the nomenclature of a science wouldn't the most important function of language consist in naming rather than in asserting?

It is questionable however that there is a most important function of language in general. Following Frege's context principle from the introduction of the *Grundlagen der Arithmetik* "never to ask for the meaning of a word in isolation, but only in the context of a proposition" [14], and following the Wittgenstein of the *Tractatus* [25], as usually understood, the most important function of language should be asserting. The belief that there is such a function and that this function is asserting was however rejected by the Wittgenstein of the *Philosophical Investigations* [26]. The later Wittgenstein said that, using his terminology, there may be language games, appropriate to particular forms of life, where various functions of language like commanding, or questioning, would have precedence over asserting, and it could not be said that these language games are less fundamental. They are not meaningful only because behind them lurks somehow the activity of asserting.

Philosophers, scientists, those living a theoretical life, were inclined since ancient times to give precedence to naming, and more recently, as it happened with Frege and Wittgenstein in the *Tractatus*, they gave precedence to asserting. (The late Frege wanted to fuse the two activities.) But does language acquire meaning primarily in theoretical life? Are not the quarters where that life is led (something like a university campus, or a leisurely residential upper-class quarter) rather lately built and not central quarters in the city language (see [26], Sect. 18)?

Even though it is not essential to agree with the later Wittgenstein on this point, it helps to do so if we want to claim without worry that in logic the most important function of language is deducing. It is strange that one has to defend nowadays this rather venerable opinion, but so many developments in the philosophy of modern logic and the philosophy of language spoke against it in the last two centuries.

Textbooks of logic in the second half of the twentieth century would often start with a definition of logic as a human endeavour concerned with deduction, and would practically not mention deduction in the remainder of the book. It is only as the century was moving to its close that natural deduction or related matters started getting ground in textbooks of logic.

2 Deductions Not Necessarily Based on Propositions

A deduction is usually taken to be a transition, a passage, from the premises to the conclusion. To simplify matters, let us assume that the premises, which are finitely many in numbers, have all been collected with the help of a connective like conjunction into a single one. Henceforth, until the last section, we will speak only about deductions that are a transition from one premise to one conclusion. Such deductions can mimic all the others.

The terms *transition* and *passage* in the preceding paragraph are far from being completely clear, and we shall return to them later, at the beginning of the third section. For the time being, let us concentrate on the premise and the conclusion. These are usually taken to be propositions, and by that is meant pieces of language that can be asserted. So it seems that our, rather common, characterization of deduction presupposes that we have propositions. Hence deducing presupposes asserting.

Could one imagine a deduction where one would pass from something that is not a proposition as a premise to something that is not a proposition as a conclusion? A deduction from a command to a command, or a deduction from a question to a question? Or, non-uniformly, a deduction from a proposition to a command, or from a command to a question? (In [24] and references therein one may find a defence of deductions where commands occur as premises and conclusions together with propositions. For the logic of questions, one may consult [16], and references therein; I don't know however of a reference dealing explicitly with deductions where questions occur as premises and conclusions, together perhaps with propositions or commands.)

Let us take a brief look at the uniform deductions, from a command to a command, or from a question to a question. Kolmogorov's contribution in [17] to the interpretation of intuitionistic logic that bears his name, besides those of Brouwer and Heyting, suggests that we should understand a deduction in constructive mathematics as taking us not from a proposition to a proposition, but from a *problem* to a *problem*. A problem however is something that does not seem to be necessarily tied with asserting. When the solution of a problem is expressed by a proposition, the statement of that problem may be a command, or a question. If the solution of our problem is "For x, y and z being respectively 3, 4 and 5 we have $x^2 + y^2 = z^2$", then the problem could be the command "Find three natural numbers x, y and z such that $x^2 + y^2 = z^2$!" or the question "Are there three natural numbers x, y and z such that $x^2 + y^2 = z^2$?".

Kolmogorov's examples of problems in [17] are expressed by commands, but it seems that they could equally well be expressed by questions. It is not however clear in these examples that the solution should be expressed by a proposition rather than by producing one or several objects, i.e. by naming them.

From $x^2+y^2 = z^2$ one can deduce $(z+x)(z-x) = y^2$. Can't we say therefore that from the command "Find three natural numbers x, y and z such that $x^2 + y^2 = z^2$!" as a premise one can deduce the command "Find three natural numbers x, y and z such that $(z + x)(z - x) = y^2$!" as a conclusion? Making one command would

yield making the other. The second command would follow from the first, it could be inferred from it. And can't we say that from the question "Are there three natural numbers x, y and z such that $x^2 + y^2 = z^2$?" as a premise one can deduce the question "Are there three natural numbers x, y and z such that $(z+x)(z-x) = y^2$?" as a conclusion? Making one question would yield making the other. The second question would follow from the first, it could be inferred from it.

To make such a deduction with commands it is not necessary to assume that the command in the premise is actually made, as in deductions with propositions it is not necessary to assume that the premise is actually asserted. Analogously, to make such a deduction with questions it is not necessary to assume that the question in the premise is actually put.

To make a deduction with propositions it does not matter whether the premise is true or not. The premise being false does not invalidate the deduction. It would invalidate it as a proof, if the deduction was proposed as a proof of the conclusion. As a deduction simpliciter, it is however perfectly legitimate with a false premise. Analogously, to make a deduction with commands it would not matter whether the premise can be fulfilled or not. The premise being impossible to fulfil would not invalidate the deduction. With propositions the deduction may serve to show that the premise is false because it yields a false conclusion, as in *reductio ad absurdum*. With commands, the deduction might serve to show that the premise cannot be fulfilled, because it yields a conclusion that cannot be fulfilled.

Commands here are assumed to have two *fulfilment* values: *can be fulfilled* and *cannot be fulfilled*, but it is not clear that therefore the logic of commands should be taken as *fulfilment-functional* and two-valued. The negation of a problem p need not be interpreted as *it is not possible to fulfil p*, but as *from the assumption that p can be fulfilled one can derive a contradiction*, which is in tune with intuitionistic logic (see [17]). The implication of that logic may be tied to deduction, and hence it would be intuitionistic.

Can the notion of deduction be widened so as to cover also non-uniform deductions involving propositions, commands and questions, like those mentioned above? Not all of these deductions need make sense. It is indeed not easy to see what would be a deduction from a command to a question. It could however again be a transition from a problem to a problem, as suggested by Kolmogorov. Deductions from propositions to commands, and vice versa, from commands to propositions, are easier to conceive, and have been examined in [24].

We shall next consider a matter that would extend even more the range of the application of the word *deduction*, and go beyond the linguistic sphere. We would thereby transcend its widest application in this sphere.

Can one make deductions involving non-linguistic entities as premises or conclusions? Could one take as a premise the perception of something small a and something big b, and deduce from that as a conclusion the proposition that a is smaller than b? Can this transition from a perception to a proposition be called a *deduction*? And can one *deduce* from a proposition a perception, not of something external, but a mental image? And can one deduce one mental image from another mental image? Why should this widening of the application of the word *deduction* to the non-linguistic

sphere represent a danger for the mathematical theory of deduction, which it is the duty of logic to formulate and investigate, and which we will consider in the next two sections?

We will not go so far as to claim that the premise and conclusion of a deduction can be anything. It seems that one could take a name as a premise or a conclusion of a deduction only in an elliptical sense. From the context one can find the proposition involving the name for which the name stands. Without that context, from a pure name, it is not clear that one could deduce anything. (Kolmogorov's solutions that are objects, which we mentioned above, could be taken as being solutions in an analogous elliptical sense.)

It does not seem unreasonable to claim however that premises and conclusions can be other things than propositions. Formulae with free variables are tied to propositions, but are not strictly speaking propositions.[1] These things, and things like commands and questions, may perhaps be tied in some way with propositions—though they are not propositions, they are somehow *in the same field*. On the other hand, the connection with propositions in the case of perceptions and mental images becomes less clear. Are they too in the field of propositions?

3 Deductions in Categories

One can surmise the following. The specificity of transitions that are deductions is not made uniquely of the things these transitions connect. Deductions are not singled out by specifying what can be premises and conclusions. Something having to do with these transitions themselves, independently of the premises and conclusions, determines that we have to do with deductions.

What could that be? What are anyway the transitions, the passages, that deductions are? Is the active, dynamic, component in the word *transition* essential?

One can next surmise the following. The specificity of transitions that are deductions does not consist in this active component. That side of the matter is psychological and is not essential from a mathematical, and logical, point of view. (The dangers of psychologism that lurk here are considered in [8].) Reified in mathematics, deductions are like arrows in a category.

A category is made of a class whose elements are called arrows, and another class whose elements are called objects, and two functions from arrows to objects—one that assigns to every arrow an object that is its source, and the other that assigns to every arrow an object that is its target. We also have operations on arrows, about which we will speak in a moment. Otherwise, arrows are not specified more closely. They can be anything, provided they have sources and targets, and the required operations. The notion of arrow is very abstract, like the notion of point or the notion of line in geometry. It is anything that satisfies the assumptions, which are very abstract too.

[1] I am grateful to Thomas Piecha for suggesting this.

In the same way, deductions have a source, called a premise, and a target, called a conclusion, and they make a structure given by operations on them. It happens that on deductions such as we have envisaged them here, with a single premise and a single conclusion, the main operations are exactly like the main operations on arrows in categories.

These are the operations that enter into the definition of a category, and they are the binary operation of composition

$$\frac{f\colon A \to B \qquad\qquad g\colon B \to C}{g \circ f\colon A \to C}$$

which in terms of deductions is a simple form of cut of sequent systems, and the nullary operations of identity arrows $\mathbf{1}_A\colon A \to A$, which as deductions are the trivial identity deductions where the premise and conclusion coincide—the primordial deductions, the axiomatic sequents. The operation of composition is partial; the target of f must be the source of g for $g \circ f$ to be defined.

For categories, one assumes associativity of composition:

$$\frac{\dfrac{f\colon A \to B \qquad g\colon B \to C}{g \circ f\colon A \to C} \qquad h\colon C \to D}{h \circ (g \circ f)\colon A \to D}$$

$$\frac{f\colon A \to B \qquad \dfrac{g\colon B \to C \qquad h\colon C \to D}{h \circ g\colon B \to D}}{(h \circ g) \circ f\colon A \to D}$$

$$h \circ (g \circ f) = (h \circ g) \circ f,$$

which makes perfect sense as an equality of deductions—it is about permuting cut with cut in sequent systems. This permuting is involved in usual cut elimination procedures (see [1], Sect. 2), and less usual ones (see [4], Chap. 1). It is however interesting in its own right, independently of these procedures. It is a perfectly natural assumption about deductions, with which they make the deepest kind of mathematical structure—a structure one finds in all categories, and in particular in the category of sets (which we will consider below).

Let us note that if, as in the Curry–Howard correspondence, one designates deductions by typed lambda terms, which is congenial with understanding proofs in the categorical, and not the hypothetical, i.e. categorial, way (see [8], Sect. 4), then composition of deductions is represented by substitution. With that, associativity of composition becomes invisible, unless one introduces, as it is sometimes done, an explicit substitution operator (see [19]). This unary operator is obtained by currying binary composition. Instead of $g \circ f$, we have something like $g\langle x, f\rangle$, which corresponds to "g where for x one substitutes f", and where $\langle x, f\rangle$ is a unary operator applied to g. Analogously, $h\langle y, g\rangle$ corresponds to "h where for y one substitutes g", and associativity of composition becomes

$$h\langle y, g\langle x, f\rangle\rangle = h\langle y, g\rangle\langle x, f\rangle.$$

It does not seem we will get closer to associativity with other notations for explicit substitution (like, for example, the notation with inverse order suggested by [13], where $g\langle x, f\rangle$ is replaced by something like $\langle f, x\rangle g$, and our equation becomes $\langle\langle f, x\rangle g, y\rangle h = \langle f, x\rangle\langle g, y\rangle h$, or a vertical notation like g^x_f, with which our equation becomes $h^y_{g^x_f} = h^{yx}_{gf}$).[2] It is improbable that one could have reached the notion of category, and realized its importance, by conceiving and representing matters pertaining to composition in that manner.

In categories one assumes moreover identity laws, i.e. laws of composing with identity arrows:

$$\frac{\mathbf{1}_A : A \to A \qquad\qquad f : A \to B}{f \circ \mathbf{1}_A : A \to B}$$

$$\frac{f : A \to B \qquad\qquad \mathbf{1}_B : B \to B}{\mathbf{1}_B \circ f : A \to B}$$

$$f \circ \mathbf{1}_A = \mathbf{1}_B \circ f = f,$$

which in terms of deductions say that composing a deduction with an identity deduction, either on the side of the premise or on the side of the conclusion, leaves the deduction unchanged. This again makes perfect sense as an equality of deductions, and is an essential ingredient of cut elimination. When the cuts have been pushed to the top of the derivation, where they are performed with axiomatic sequents, they disappear.

Lambek (see [18]) called *deductive system* the notion generalizing categories by not assuming the associativity of composition and the laws of composing with identity arrows. In [3] and [4] (Sect. 1.9) one can see how this notion of deductive system is characterized proof-theoretically by a representation result in the style of Stone, and how the notion of category is characterized proof-theoretically by a representation result in the style of Cayley.

There may be further operations on arrows, with which we enter into the field of categories with additional structure. One such operation is tied to the biendofunctor of product, which corresponds to conjunction, both in classical and intuitionistic logic. Coproduct, with which another such operation is tied, corresponds to disjunction, in both logics again. With product and coproduct we obtain equations between deductions that make perfect sense in proof theory. They stem from adjointness of functors, and are related to normalization in natural deduction and cut elimination in sequent systems (see [4], [5] and [9]). Other equations, like those involving distributivity of product over coproduct, i.e. conjunction over disjunction, may, but need not, be based on adjointness. Intuitionistic logic will differ essentially from classical logic by tying implication to adjointness (see [2] and [4]), which should not be done for

[2] Roy Dyckhoff was kind to comment upon this.

classical, material, implication (see [9], Chap. 14). We will not go further into categorial proof theory, which deals with the equations between deductions suggested by such categories with additional structure. Let it be said only that it is remarkable how equations important in mathematics in general, or in particular fields of mathematics, reemerge as perfectly sensible equations between deductions (see [7]).

What was surmised above is that the structure that deductions make with such operations is an essential ingredient in the notion of deduction. Could one go as far as to take this as the main ingredient? As in category theory, the structure of arrows would be the main thing. And, as in category theory, the arrows would be more important than the objects. With deductions the objects are the premises and conclusions, and these premises and conclusions, whatever they are precisely—propositions, commands, questions, problems, or something else—would not precede the arrows. Deducing would not be preceded by asserting, or another function of language.

When functions are reified as sets of ordered pairs, the active, dynamic, component in the notion of function is lost. This component, which comes from psychology, is also lost in the reification brought by the categorial notion of function, where a function is an arrow in a category. Categorially, functions in the category **Set**, where the objects are sets and the arrows are functions, are characterized through composition and identity functions. The same operations characterize deductions in general.

Although **Set** has the structure that deductions make, it is not natural for its objects to be called premises and conclusions, and for its arrows to be called deductions. One reason may be the nature of these objects, which are not in the field of propositions (see the end of the preceding section). Another reason might be that we have too many of these deductions. Any two objects would be connected by a deduction, except when the premise is not empty while the conclusion is empty. Deductions, in the categories where arrows may be more naturally designated by that term, are usually more discriminatory. There are more objects not connected by arrows.

The structure of deductions imitates the structure of the category **Set** even more when they involve the binary connectives of conjunction, disjunction and implication, together with the nullary connectives $\top$ and $\bot$. This structure, appropriate for intuitionistic propositional logic, imitates **Set** with the biendofunctors of product, coproduct and exponentiation (for exponentiation we have covariance in the base and contravariance in the exponent), together with the terminal and initial objects, i.e. a singleton and the empty set. Still, the arrows of the category **Set** could not be taken as deductions, but only as their model. (The question of models of deductions was discussed in [6].) This is the model that stands behind the standard proof-theoretic semantics for intuitionistic logic, which through the Curry–Howard correspondence is tied to the typed lambda calculus.

Matters become clearer when in conceiving this semantics we do not conform to the dogma mentioned at the beginning. When we look upon this semantics hypothetically, and not categorically, as in the typed lambda calculus, we will end up in the categorial setting of **Set**. With the typed lambda calculus we also end up in the sets of **Set**, but the categorial setting is hidden.

In [9] one may find a categorial setting for the proof theory of classical conjunctive-disjunctive logic different from that of **Set**, which leads to a categorial setting for the proof theory of the whole of classical propositional logic where a characterization through adjunction for classical implication is relinquished. Implication is again characterized through adjunction in the categorial setting for the proof theory of linear propositional logic without modalities, which is investigated in [10].

4 Deductions in Multicategories and Polycategories

Let us consider now the deductions where we can have more than one premise, though we still have a single conclusion. Such deductions, which correspond to Gentzen's singular sequents, with not more than one formula on the right-hand side, correspond to arrows $f : \Gamma \to A$ in Lambek's *multicategories*, sometimes called *multiarrows*, where capital Greek letters like Γ stand for finite sequences of objects. A particular kind of multicategory is an operad, where Γ in multiarrows $f : \Gamma \to A$ is a finite sequence every member of which is A. The algebraic notion of operad, which has arisen in algebraic topology, has been much investigated lately. (References concerning the notions of operad and multicategory, and a discussion of matters concerning them, may be found in [11].)

Multicategories can be mimicked by categories with additional structure like monoidal categories, which have a binary operation on objects like conjunction, enabling us to bind the objects in Γ into a single object. The particular structure of multicategories is however important and interesting, and we shall now examine one aspect of it.

In multicategories instead of composition we have *insertion* operations on multiarrows:

$$\frac{f : \Gamma \to A \qquad g : \Delta, A, \Theta \to B}{g \lhd f : \Delta, \Gamma, \Theta \to B}$$

which correspond to Gentzen's cut of singular sequents. The notation $g \lhd f$ is ambiguous, because it does not specify the cut formula. This ambiguity is remedied with the more precise notation of [11], but for the comments we will make here we can do with the less precise notation we have just introduced.

In multicategories, besides the associativity of insertion that corresponds to the associativity of composition in categories:

$$\frac{\dfrac{f : \Gamma \to A \qquad g : \Delta, A, \Theta \to B}{g \lhd f : \Delta, \Gamma, \Theta \to B} \qquad h : \Pi, B, \Sigma \to C}{h \lhd (g \lhd f) : \Pi, \Delta, \Gamma, \Theta, \Sigma \to C}$$

$$\frac{f : \Gamma \to A \qquad \dfrac{g : \Delta, A, \Theta \to B \qquad h : \Pi, B, \Sigma \to C}{h \lhd g : \Pi, \Delta, A, \Theta, \Sigma \to C}}{(h \lhd g) \lhd f : \Pi, \Delta, \Gamma, \Theta, \Sigma \to C}$$

$$h \lhd (g \lhd f) = (h \lhd g) \lhd f,$$

we have another kind of associativity of insertion, which involves also commutativity:

$$\dfrac{g\colon \Sigma \to B \qquad \dfrac{f\colon \Gamma \to A \qquad h\colon \Delta, A, \Theta, B, \Pi \to C}{h \lhd f\colon \Delta, \Gamma, \Theta, B, \Pi \to C}}{(h \lhd f) \lhd g\colon \Delta, \Gamma, \Theta, \Sigma, \Pi \to C}$$

$$\dfrac{f\colon \Gamma \to A \qquad \dfrac{g\colon \Sigma \to B \qquad h\colon \Delta, A, \Theta, B, \Pi \to C}{h \lhd g\colon \Delta, A, \Theta, \Sigma, \Pi \to C}}{(h \lhd g) \lhd f\colon \Delta, \Gamma, \Theta, \Sigma, \Pi \to C}$$

$$(h \lhd f) \lhd g = (h \lhd g) \lhd f,$$

This other associativity is interesting algebraically and combinatorially. It is also related to interesting polyhedra (see [11], Sect. 13).

In *polycategories* we have arrows $\Gamma \to \Delta$, which correspond to the plural sequents of Gentzen, where there may be several formulae on the right-hand side too. The arrows of polycategories correspond to the deductions of classical logic, which are investigated graph-theoretically in [12]. For polycategories and their operations of insertions, which correspond to Gentzen's cut for plural sequents, we have besides the associativity and the associativity involving commutativity on the left-hand side, analogous to those we have just given for multicategories, another associativity involving commutativity on the right-hand side (see [12], Propositions 2.1–3).

These plural deductions of classical logic are not very natural. They have been invented following Gentzen's suggestion, and not found by describing deduction in real life. They provide however the best means to understand classical logic proof-theoretically. They are implicit in the categorial approach to the proof theory of classical logic of [9], and in the categorial approach to the proof theory of linear logic of [10].

The remarks made here on deduction lead to the following tentative characterization of this notion. A deduction is an arrow in a category, a multicategory, or a polycategory, where the objects are more or less akin to propositions—something in the field of propositions.

5 Rules for Deductions

Nothing has been said up to now about deductions being in accordance with rules. When we deal with formal deductions, i.e. the deductions of logic, and they are conceived as arrows in categories with additional structure, which is brought by something like the functors corresponding to connectives that we mentioned in the third section, then our deductions are members of families of arrows indexed by

the objects of the category, which are usually natural transformations involving the functors we have just mentioned. Being in accordance with rules here amounts to being members of such families, and the schematic character of the rules is given by the indexing by objects of the members of our families of arrows. With this indexing, the objects that are indices serve to make the sources or targets of the arrows, i.e. the premises or conclusions of the deductions. For example, the natural transformation p^1 of the first projection for conjunction elimination with the indices being the objects A and B gives the deduction $p^1_{A,B} \colon A \wedge B \to A$.

Is this indexing necessary for the notion of rule for deduction? Can a rule correspond to a family of arrows that is a singleton, without indexing? Can one call *rule* something which covers a single deduction, with which a single deduction is in accordance? In or outside logic, is generality necessary for rules? Should a rule always cover many cases? Can a rule cover a single case?

Deductions that are not in logic may still resemble the formal deductions of logic by being in accordance with schematically given rules. Such would be the deduction from the premise "The day before yesterday was Thursday" to the conclusion "Tomorrow will be Sunday" (though it is not immediately clear how to formulate the rules in question). If they cannot be found in logic, could one find outside logic deductions that are not instances of something schematic? Shouldn't they be in accordance with rules? What would be the appropriate notion of rule there? When bereft of its psychological or sociological aspects, like compulsoriness, would this notion of rule leave something to be investigated by precise, perhaps even mathematical, means? Grammar and linguistics may give an inspiration for considering such matters, which are close to the concerns of the later Wittgenstein.

In [27] (end of Lecture XIII, Lent Term 1935) Wittgenstein taught that "a rule is something applied in many cases", but then disparaged this remark off-handedly. He considered it useless for learning how to use a rule. Why must this remark serve that purpose? In another context, where the purpose is to explain what rules are and not to teach how to use them, it may prove important to determine whether generality is necessary for rules. Wittgenstein returned to this question in [26] (Sect. 199) and in other places (for references see [15], Sect. 199, pp. 120–124), with consideration towards the generality of rules, which he put within a wider scheme.

Wittgenstein ended the lecture from which we have quoted above by a nice and enigmatic picture: "A rule is best described as being like a garden path in which you are trained to walk, and which is convenient." A path is usually something taken many times, by many people. If a rule is like a path, the deductions in accordance with the rule could perhaps be like many particular walks on this path. We will however try to consider more closely this and other matters mentioned in this section on another occasion.

Acknowledgments Work on this paper was supported by the Ministry of Education, Science and Technological Development of Serbia, while the Alexander von Humboldt Foundation has supported the presentation of a part of it at the Second Conference on Proof-Theoretic Semantics, in Tübingen, in March 2013. I am grateful to the organizers of that conference, and in particular Peter Schroeder-Heister, for their exceptional hospitality. I am also grateful to him and to Thomas Piecha for making some useful comments on this paper.

References

1. Borisavljević, M., Došen, K., Petrić, Z.: On permuting cut with contraction. Math. Struct. Comput. Sci. **10**, 99–136 (2000). http://arXiv.org
2. Došen, K.: Deductive completeness. Bull. Symb. Log. **2**(523), 243–283 (1996). For corrections see [4], Section 5.1.7, and [5]
3. Došen, K.: Deductive systems and categories. Publications de l'Institut Mathématique N.S. **64**(78), 21–35 (1998)
4. Došen, K.: Cut Elimination in Categories. Kluwer, Dordrecht (1999)
5. Došen, K.: Abstraction and application in adjunction. In: Kadelburg, Z. (ed.) Proceedings of the Tenth Congress of Yugoslav Mathematicians, pp. 33–46. Faculty of Mathematics. University of Belgrade, Belgrade (2001) Available at: http://arXiv.org
6. Došen, K.: Models of deduction. In: Kahle, R., Schroeder-Heister, P. (eds.) Proof-Theoretic Semantics, Proceedings of the Conference "Proof-Theoretic Semantics, Tübingen, 1999", Synthese, vol. 148, pp. 639–657 (2006). Available at: http://www.mi.sanu.ac.rs/kosta/publications.htm
7. Došen, K.: Algebras of deductions in category theory. In: Jokanović et al. (eds), Third Mathematical Conference of the Republic of Srpska, Proceedings, Trebinje 2013, Zbornik radova, vol. I, pp. 11–18. Univerzitet u Istočnom Sarajevu, Fakultet za proizvodnju i menadžment, Trebinje (2014). Available at: http://www.mi.sanu.ac.rs/kosta/DosenAlgebrasofDeductions.pdf; http://www.mk.rs.ba/wp-content/uploads/2015/02/TOM1-Copy.pdf, pp. 1–8
8. Došen, K.: Inferential semantics. In: Wansing, H. (ed.) Dag Prawitz on Proofs and Meaning, pp. 147–162. Springer, Cham (2015). Preprint of 2012 available at: http://www.mi.sanu.ac.rs/kosta/publications.htm
9. Došen, K., Petrić, Z.: Proof-Theoretical Coherence. KCL Publications (College Publications), London (2004). Revised version of 2007 available at: http://www.mi.sanu.ac.rs/kosta/publications.htm
10. Došen, K., Petrić, Z.: Proof-Net Categories. Polimetrica, Monza (2007). Preprint of 2005 available at: http://www.mi.sanu.ac.rs/kosta/publications.htm
11. Došen, K., Petrić, Z.: Weak cat-operads. (2010). Available as v8, the last version of the authors, at: http://arXiv.org
12. Došen, K., Petrić, Z.: Graphs of plural cuts. Theor. Comput. Sci. **484**, 41–55 (2013). Available at: http://arXiv.org
13. Dyckhoff, R., Pinto, L.: Cut-elimination and a permutation-free sequent calculus for intuitionistic logic. Studia Logica **60**, 107–118 (1998)
14. Frege, G.: Die Grundlagen der Arithmetik: Eine logisch mathematische Untersuchung über den Begriff der Zahl. Verlag von Wilhelm Koebner, Breslau (1884). English translation by Austin, J.L.: The Foundations of Arithmetic: A Logico-Mathematical Enquiry into the Concept of Number, 2nd revised edn. Blackwell, Oxford (1974)
15. Hacker, P.M.S.: Wittgenstein: Rules, Grammar and Necessity–Essays and Exegesis of §§185-242. Wiley-Blackwell, Chichester (2009). An Analytical Commentary on the Philosophical Investigations, 2nd extensively revised edn. (1st edn. of 1985 by Baker, G.P., Hacker, P.M.S.)
16. Harrah, D.: The logic of questions. In: Gabbay, D., Guenthner, F. (eds.) Handbook of Philosophical Logic, vol. 8, pp. 1–60. Kluwer, Dordrecht (2002)
17. Kolmogorov, A.N.: Zur Deutung der intuitionistischen Logik. Mathematische Zeitschrift **35**, 58-65 (1932). English translation by V.M. Volosov from a Russian translation by Uspensky, V.A.: On the interpretation of intuitionistic logic, Selected Works of Kolmogorov, A.N.: Mathematics and Mechanics, vol. 1, pp. 151–158 Kluwer, Dordrecht, (1991)

18. Lambek, J., Scott, P.J.: Introduction to Higher Order Categorical Logic. Cambridge University Press, Cambridge (1986)
19. Rose, K.H.: Explicit substitution: tutorial and survey. In: BRICS Lecture Series. University of Aarhus, Aarhus (1996)
20. Schroeder-Heister, P.: Proof-theoretic versus model-theoretic consequence. In: Peliš, M. (ed.) The Logica Yearbook 2007, pp. 187–200. Prague (2008)
21. Schroeder-Heister, P.: The categorical and the hypothetical: a critique of some fundamental assumptions of standard semantics. In: Lindström, S. et al. (eds.) The Philosophy of Logical Consequence and Inference, Proceedings of the Workshop "The Philosophy of Logical Consequence, Uppsala, 2008", Synthese, vol. 187, pp. 925–942 (2012)
22. Schroeder-Heister, P., Contu, P.: Folgerung. In: Spohn, W., et al. (eds.) Logik in der Philosophie, pp. 247–276. Synchron, Heidelberg (2005)
23. Simpson, J., Weiner, E. (eds.): The Oxford English Dictionary, 2nd edn. Oxford University Press, Oxford (1989)
24. Vranas, P.B.M.: In defense of imperative inference. J. Philos. Log. **39**, 59–71 (2010)
25. Wittgenstein, L.: Logisch-philosophische Abhandlung. Annalen der Naturphilosophie **14**, 185–262 (1921). English translation by Ogden, C.K.: Tractatus logico-philosophicus, Routledge, London (1922) new translation by Pears D.F., McGuinness B.F., Routledge, London, (1961)
26. Wittgenstein, L.: Philosophische Untersuchungen. Blackwell, Oxford (1953). English translation by Anscombe, G.E.M.: Philosophical Investigations, 4th edn. with revisions by Hacker P.M.S., Schulte J., Wiley-Blackwell, Oxford (2009)
27. Wittgenstein, L.: Wittgenstein's Lectures, Cambridge 1932-1935. In: Ambrose, A. (ed.) From the Notes of Alice Ambrose and Margaret Macdonald. Blackwell, Oxford (1979)

Some Remarks on Proof-Theoretic Semantics

Roy Dyckhoff

Abstract This is a tripartite work. The first part is a brief discussion of what it is to be a logical constant, rejecting a view that allows a particular self-referential "constant" • to be such a thing in favour of a view that leads to strong normalisation results. The second part is a commentary on the flattened version of Modus Ponens, and its relationship with rules of type theory. The third part is a commentary on work (joint with Nissim Francez) on "general elimination rules" and harmony, with a retraction of one of the main ideas of that work, i.e. the use of "flattened" general elimination rules for situations with discharge of assumptions. We begin with some general background on general elimination rules.

Keywords General elimination rules · Harmony · Strong normalisation

1 Background on General Elimination Rules

Standard natural deduction rules for **Int** (intuitionistic predicate logic) in the style of Gentzen [9] and Prawitz [24] are presumed to be familiar. The theory of cut-elimination for sequent calculus rules is very clear: whether a derivation in a sequent calculus is cut-free or not is easily defined, according to the presence or absence of instances of the *Cut* rule. For natural deduction, normality is a less clear concept: there are several inequivalent definitions (including variations such as "full normality") in the literature. For implicational logic it is easy; but rules such as the elimination rule for disjunction cause minor problems with the notion of "maximal formula occurrence" (should one include or not include the permutative conversions?), and more problems when minor premisses have vacuous discharge of assumptions.

The material was presented at the proof-theoretic semantics conference in Tübingen in March 2013; an early version was presented at the proof-theoretic semantics (Arché) workshop in St Andrews in May 2009.

R. Dyckhoff (✉)
University of St Andrews, Fife, UK
e-mail: roy.dyckhoff@st-andrews.ac.uk

T. Piecha and P. Schroeder-Heister (eds.), *Advances in Proof-Theoretic Semantics*,
Trends in Logic 43, DOI 10.1007/978-3-319-22686-6_5

One proposed solution, albeit partial, is the uniform use of *general elimination rules*, i.e. *GE-rules*. These can be motivated in terms of *Prawitz's inversion principle*[1]: "the conclusion obtained by an elimination does not state anything more than what must have already been obtained if the major premiss of the elimination was inferred by an introduction" [25, p. 246]. Normality is now the simple idea [39] that the major premiss of each elimination step should be an assumption; see also [13, 36].

The standard elimination rules for disjunction, absurdity and existential quantification are already GE rules:

$$\dfrac{A \vee B \quad \begin{array}{c}[A]\\ \vdots \\ C\end{array} \quad \begin{array}{c}[B]\\ \vdots \\ C\end{array}}{C}\ \vee E \qquad \dfrac{\bot}{C}\ \bot E \qquad \dfrac{\exists x A(x) \quad \begin{array}{c}[A(y)]\\ \vdots \\ C\end{array}}{C}\ \exists E$$

(with y fresh in $\exists E$) and the same pattern was proposed (as a GE-rule) in the early 1980 s for conjunction

$$\dfrac{A \wedge B \quad \begin{array}{c}[A, B]\\ \vdots \\ C\end{array}}{C}$$

by various authors, notably Prawitz [26, 27], Martin-Löf [15] and Schroeder-Heister [30], inspired in part by type theory (where conjunction is a special case of the Σ-type constructor, with $A \wedge B =_{def} \Sigma(A, B)$ whenever $B(x)$ is independent of x) and (perhaps) in part by linear logic [10] (where conjunction appears in two flavours: multiplicative $\otimes$ and additive &).

To this one can add GE-rules for implication[2] and universal quantification:

$$\dfrac{A \supset B \quad A \quad \begin{array}{c}[B]\\ \vdots \\ C\end{array}}{C}\ \supset GE \qquad \dfrac{\forall x.B(x) \quad t\ term \quad \begin{array}{c}[B(t)]\\ \vdots \\ C\end{array}}{C}\ \forall GE$$

Rules of the first kind are conveniently called "flattened" [29] (in comparison with Schroeder-Heister's "higher-level" rules, for which see [30, 32]). López-Escobar [13] distinguishes between the premiss A of $\supset GE$ as a "minor" premiss and that of C (assuming B) as a "transfer" premiss.[3]

One thus has a calculus of rules in natural deduction style for **Int**; such calculi, and their normalisation results, have been studied by von Plato [39], by López-Escobar

[1] See [19, 31] for discussions of this principle, including its antecedents in the work of Lorenzen.

[2] Reference [3] has an early occurrence of this.

[3] On the other hand, Francez and Dyckhoff [8] calls A the "support" and, more in line with tradition than López-Escobar [13], the remaining premiss the "minor premiss".

[13] and by Tennant [36]. With the definition (given above) that a deduction is *normal* iff the major premiss of every elimination step is an assumption, the main results are:

1. Weak Normalisation (WN): every deduction can be replaced by a normal deduction of the same conclusion from the same assumptions [13, 20, 36, 39].
2. Strong Normalisation (SN), for the implicational fragment: an obvious set of rules for reducing non-normal deductions is strongly normalising, i.e. every reduction sequence terminates [12, Sect. 6], [13, 36, 37].
3. SN, for the full language: a straightforward extension of the proof of [37] for implication[4]; also, the proofs for implication "directly carry over" [12] to a system with conjunctions and disjunctions. An argument (using the ordinary elimination rule for implication) is given in [35] for the rules for implication and existential quantification, with the virtue of illustrating in detail how to handle GE rules where the Tait–Martin-Löf method of induction on types familiar from [11] is not available. See also [13].
4. Some straightforward arguments for normalisation (by induction on the structure of the deduction) [40].
5. A 1-1 correspondence with intuitionistic sequent calculus derivations [20, 39].
6. Some interpolation properties [17].
7. Extension of the normalisation results to classical logic [41].

Despite the above results, there are some disadvantages:

1. Poor generalisation of the GE rule for implication to the type-theoretic constant Π, of which $\supset$ can be treated as a special case [15]: details below in Sect. 3.
2. Too many deductions, as in sequent calculus. Focused [aka "permutation-free"] sequent calculi [5, 6] have advantages. Sequent calculus has (for each derivable sequent) rather too many derivations, in comparison to natural deduction, since derivations often have many permutations each of which is, when translated to ordinary natural deduction, replaced by an identity of deductions. The GE-rules have the same feature, which interferes with rather than assists in root-first proof search.
3. (For some complex constants, if one adopts the methodology beyond the basic intuitionistic ones) a "disharmonious mess" [4]: details below in Sect. 4.4.
4. No SN results (yet, in general) for GE-rules for arbitrarily complex constants.

2 Is Bullet a Logical Constant?

Read [28] has, following a suggestion of Schroeder-Heister (with Appendix B of Prawitz's [26] and Ekman's Paradox[5] [7] in mind), proposed as a logical constant a nullary operator $\bullet$ (aka R, for "Russell") with the single (but impure) introduction rule

[4]Personal communication from Jan von Plato, May 2009.

[5]See [34] for a recent discussion.

$$\dfrac{\begin{array}{c}[\bullet]\\ \vdots\\ \bot\end{array}}{\bullet}\ \bullet I$$

The GE-rule justified by this (along the same lines as for implication) is then

$$\dfrac{\bullet \quad \bullet \quad \begin{array}{c}[\bot]\\ \vdots\\ C\end{array}}{C}\ \bullet GE$$

which, given the usual $\bot E$ rule and the unnecessary duplication of premisses, can be simplified to

$$\dfrac{\bullet}{C}\ \bullet E$$

So, by this $\bullet E$ rule, the premiss of the $\bullet I$ rule is deducible, hence $\bullet$ is deducible, hence $\bot$ is deducible.

There is however a weakness (other than just that it leads to inconsistency) in the alleged justification of $\bullet$ as a logical constant: it is a circularity. We follow Martin-Löf [15, 16] and Dummett [2] in accepting that we understand a proposition when we understand what it means to have a *canonical* proof of it, i.e. what forms a canonical proof can take. In the case of $\bullet$, there is a circularity: the introduction rule gives us a canonical proof only once we have a proof of $\bot$ from the assumption of $\bullet$, i.e. have a method for transforming *arbitrary* proofs of $\bullet$ into proofs of $\bot$. The reference here to "arbitrary proofs of $\bullet$" is the circularity.

There are similar ideas about type formers, and it is instructive to consider another case, an apparent circularity: the formation rule (in [15]) for the type N of natural numbers. That is a type that we understand when we know what its canonical elements are; these are 0 and, when we have an element n of N, the term $s(n)$. The reference back to "an element n of N" looks like a circularity of the same kind; but it is rather different—we don't need to grasp *all* elements of N to construct a canonical element by means of the rule, just *one* of them, namely n.

A formal treatment of this issue has long been available in the type theory literature, e.g. Mendler [18], Luo [14], *Coq* [1]. We will try to give a simplified version of the ideas. With the convention that propositions are interpreted as types (of their proofs), we take type theory as a generalisation of logic, with ideas and restrictions in the former being applicable to the latter. The simplest recursive case (N) has just been considered and the recursion explained as harmless (despite Dummett's reservations expressed as his "complexity condition" [2]). What about more general definitions?

The definition of the type N can be expressed as saying that N is the least fixed point of the operator $\Phi_N =_{def} \lambda X.(1 + X)$, i.e. $N =_{def} \mu X.(1 + X)$. Similarly, the type of lists of natural numbers is $\mu L.(1 + N \times L)$, and the type of binary trees with leaves in A and node labels in B is $\mu T.A + (T \times B \times T)$. A unary operator definition $\Phi =_{def} \lambda X. \ldots$ is said to be *positive* iff the only occurrences of the type variable X in the body ... are positive, where an occurrence of X in the expression $A \to B$ is *positive* (resp. *negative*) iff it is a positive (resp. negative) occurrence in B or a negative (resp. positive) occurrence in A; a variable occurs *positively* in itself, and occurs *positively* (resp. negatively) in $A + B$ and in $A \times B$ just where it occurs positively (resp. negatively) in A or in B. A definition of a type as the least fixed point of an operator is then *positive* iff the operator definition is positive.

Read's $\bullet$, then, is defined as $\mu X.(X \to \bot)$. This is not a positive definition; the negativity of the occurrence of X in the body $X \to \bot$ is a symptom of the circular idea that $\bullet$ can be grasped once we already have a full grasp of what the proofs of $\bullet$ might be.

In practice, a stronger requirement is imposed, that the definition be strictly positive, i.e. the only occurrences of the type variable X in the body ... are strictly positive, where an occurrence of X in the expression $A \to B$ is *strictly positive* iff it is a strictly positive occurrence in B; a variable occurs *strictly positively* in itself, and occurs *strictly positively* in $A + B$ and in $A \times B$ just where it occurs strictly positively in A or in B. A definition of a type as the least fixed point of an operator is then *strictly positive* iff the operator definition is strictly positive.

With such definitions, it can be shown that strong normalisation (of a suitable set of reductions) holds [18, Chap. 3]; similar accounts appear in [1, 14].

3 The GE-rule for Implication and the Type-Theoretic Dependent Product Type

The present author commented [3] that the general (aka "flattened" [29]) E-rule for implication didn't look promising because it didn't generalise to type theory. Here (after 27 years) are the details of this problem: Recall [15] that in the dependently-typed context

$$A \ \mathit{type}, \ B(x) \ \mathit{type} \ [x : A],$$

the rule

$$\frac{p : \Sigma(A, B) \qquad \begin{array}{c}[z : \Sigma(A, B)] \\ \vdots \\ C(z) \ \mathit{type}\end{array} \qquad \begin{array}{c}[x : A, y : B(x)] \\ \vdots \\ c(x, y) : C((x, y))\end{array}}{split(p, c) : C(p)} \ \Sigma E$$

with[6] semantics $split((a, b), c) \to c(a, b)$ is a generalisation of the rule

$$\frac{p : A \times B \quad C \ type \quad \begin{array}{c}[x : A, y : B]\\ \vdots \\ c(x, y) : C\end{array}}{split(p, c) : C} \times E.$$

Now, ordinary (but with witnesses) *Modus Ponens*

$$\frac{f : A \supset B \quad a : A}{fa : B} \supset E$$

has, in the dependently-typed context

$$A \ type, \ B(x) \ type \ [x : A],$$

the generalisation (in which $ap2(f, a)$ is often just written as $f\ a$ or $(f\ a)$):

$$\frac{f : \Pi(A, B) \quad a : A}{ap2(f, a) : B(a)} \Pi E$$

(with $\Pi(A, B)$ written as $A \supset B$ whenever $B(x)$ is independent of x); but the "flattened" GE rule

$$\frac{f : A \supset B \quad a : A \quad C \ type \quad \begin{array}{c}[y : B]\\ \vdots \\ c(y) : C\end{array}}{ap3(f, a, c) : C} \supset GE$$

with semantics $ap3(\lambda(g), a, c) \to c(g(a))$ doesn't appear to generalise:

$$\frac{f : \Pi(A, B) \quad a : A \quad \begin{array}{c}[z : \Pi(A, B)]\\ \vdots \\ C(z) \ type\end{array} \quad \begin{array}{c}[y : B(a)]\\ \vdots \\ c(y) : C(\lambda(\mathbf{?}))\end{array}}{ap3(f, a, c) : C(f)}$$

in which, note the question-mark—what should go there? In the context $y : B(a)$, the only ingredient is y, which won't do—it has the wrong type. Addition of an assumption such as $x : A$ (and making c depend on it, as in $c(x, y)$) doesn't help.

[6]The notation c is used to abbreviate $\lambda xy.c(x, y)$. Similar abbreviations are used below. $c(a, b)$ is then just c applied to a and then to b.

One solution is the system of higher-level rules of Schroeder-Heister [30]. Our own preference, to be advocated after a closer look at flattened GE-rules, is for implication (and universal quantification) to be taken as primitive, with *Modus Ponens* and the ΠE rule taken as their elimination rules, with justifications as in [15].

4 GE-Rules in General

The wide-spread idea that the "grounds for asserting a proposition" collectively form some kind of structure which can be used to construct the assumptions in the minor premiss(es)[7] of a GE-rule is attractive, as illustrated by the idea that, where two formulae A, B are used as the grounds for asserting $A \wedge B$, one may make the pair A, B the assumptions of the minor premiss of $\wedge GE$. An example of this is López-Escobar's [13], which gives I-rules and then GE-rules for implication[8] and disjunction, with the observation [13, p. 417] that:

> Had the corresponding **I**-rule had three "options with say 2, 3 and 5 premises respectively, then there would have been $2 \times 3 \times 5$ **E**-rules corresponding to that logical atom.[9] Also had there been an indirect[10] premise, say $\nabla\mathfrak{D}/\mathfrak{E}$, in one of the options then it would contribute a minor premise with conclusion $\mathfrak{E}$ and a transfer premise with discharged sentence $\mathfrak{D}$ to the appropriate[11] **E**-rule.

In practice, there is an explosion of possibilities, which we analyse in order as follows:

1. a logical constant, such as $\bot$, $\wedge$, $\vee$, $\equiv$ or $\oplus$ (exclusive or), can be introduced by zero or more rules;
2. each of these rules can have zero or more premisses, e.g. $\top I$ has zero, $\supset I$ and each $\vee I_i$ have one, $\wedge I$ has two;
3. each such premiss may discharge zero or more assumptions (as in $\supset I$);
4. each such premiss may abstract over one or more variables, as in $\forall I$;
5. and a premiss may take a term as a parameter (as in $\exists I$).

It is not suggested that this list is exhaustive: conventions such as those of substructural logic about avoiding multiple or vacuous discharge will extend it, as would recursion; but it is long enough to illustrate the explosion. The paper [8] attempted[12] to deal with all these possibilities and carry out a programme of mechanically generating GE-rules from a set of I-rules with results about harmony.

---

[7] In the sense of "all but the major premiss".

[8] He also gives primitive rules for negation; in our view this is best treated as a defined notion, since even its I-rule is impure, i.e. mentions the constant $\bot$.

[9] In our terminology, "logical constant".

[10] $\nabla\mathfrak{D}/\mathfrak{E}$ is the notation of [13] for "$\mathfrak{E}$ [assuming $\mathfrak{D}$]".

[11] Surely, in this, $\mathfrak{D}$ and $\mathfrak{E}$ are the wrong way round.

[12] In ignorance, alas, of [13].

4.1 Several I-Rules

Where a logical constant (such as $\vee$) is introduced by several alternative[13] rules, one can formulate an appropriate GE-rule as having several minor[14] premisses, one for each of the I-rules, giving a case analysis. This is very familiar from the case of $\vee$ and the usual $\vee E$ rule:

$$\frac{A \vee B \quad \begin{array}{c}[A]\\ \vdots \\ C\end{array} \quad \begin{array}{c}[B]\\ \vdots \\ C\end{array}}{C}$$

so an appropriate generalisation for $n \geq 0$ alternative I-rules is to ensure that "the GE-rule" has n minor premisses. This works well for $\bot$, with no I-rules: the $\bot E$-rule, as in [9, 24], has no minor premisses.[15]

4.2 I-Rule Has Several Premisses

Now there are two possibilities following the general idea that the conclusion of a GE-rule is arbitrary. Let us consider the intuitionistic constant $\wedge$ (with its only I-rule having two premisses) as an example. The first possibility is as illustrated earlier: the rule

$$\frac{A \wedge B \quad \begin{array}{c}[A, B]\\ \vdots \\ C\end{array}}{C}\ \wedge GE.$$

The second is to have two GE-rules:

$$\frac{A \wedge B \quad \begin{array}{c}[A]\\ \vdots \\ C\end{array}}{C}\ \wedge GE_1 \qquad \frac{A \wedge B \quad \begin{array}{c}[B]\\ \vdots \\ C\end{array}}{C}\ \wedge GE_2$$

and it is routine to show that the ordinary GE-rule for $\wedge$ is derivable in a system including these two rules, and vice-versa. Tradition goes for the first possibility; examples below show however that this doesn't always work and that the second may be required.

[13] López-Escobar [13] calls these "options".

[14] "Transfer" premises in the terminology of [13].

[15] López-Escobar [13] does it differently.

4.3 Premiss of I-Rule Discharges Some Assumptions

Natural deduction's main feature is that assumptions can be discharged, as illustrated by the I-rule for $\supset$ and the E-rule for $\vee$. This raises difficulties for the construction of the appropriate GE-rules: Prawitz [26] got it wrong (corrected in [27]), Schroeder-Heister [30] gave an answer in the form of a system of rules of higher level, allowing discharge not just of assumptions but of rules (which may themselves discharge …)—but, although much cited, use of this system seems to be modest. As already discussed, an alternative was mentioned (disparagingly) in [3] and (independently) adopted more widely by others [13, 36, 39], the "flattened" GE-rule for $\supset$ being

$$\dfrac{A \supset B \quad A \quad \begin{array}{c}[B]\\ \vdots \\ C\end{array}}{C} \supset GE$$

Let us now consider the position where two premisses discharge an assumption (just one each is enough): consider the logical constant $\equiv$ with one I-rule, namely

$$\dfrac{\begin{array}{c}[A]\\ \vdots \\ B\end{array} \quad \begin{array}{c}[B]\\ \vdots \\ A\end{array}}{A \equiv B} \equiv I$$

According to our methodology, we have two possibilities for the GE-rule; first, have the minor premiss of the rule with two assumptions B, A being discharged and some device to ensure that there are other premisses with A and B as conclusions. There seems to be no way of doing this coherently, i.e. with A somehow tied to the discharge of B and vice-versa. The alternative is to have *two* GE-rules, along the lines discussed above for $\wedge$, and these are clearly

$$\dfrac{A \equiv B \quad A \quad \begin{array}{c}[B]\\ \vdots \\ C\end{array}}{C} \equiv E_1 \qquad \dfrac{A \equiv B \quad B \quad \begin{array}{c}[A]\\ \vdots \\ C\end{array}}{C} \equiv E_2$$

by means of which it is clear that, from the assumption of $A \equiv B$, one can construct a proof of $A \equiv B$ using the introduction rule as the last step, implying the "local completeness" of this set of rules in a sense explored by Pfenning and Davies [22]:

$$\dfrac{\dfrac{A \equiv B \quad [A]^2 \quad \begin{array}{c}[B]^1\\ \vdots \\ B\end{array}}{B} \equiv E_1, 1 \quad \dfrac{A \equiv B \quad [B]^4 \quad \begin{array}{c}[A]^3\\ \vdots \\ A\end{array}}{A} \equiv E_2, 3}{A \equiv B} \equiv I, 2, 4$$

We are thus committed in general to the use of the second rather than the first possibility of GE-rules—the use of two such rules rather than one—when there are two premisses in an I-rule.

4.4 GE Harmony: A Counter-Example

Francez and the present author [8][16] developed these ideas (looking also at the analogues of universal and existential quantification) by defining the notion of "GE-harmony" (E-rules are GE-rules obtained according to a formal procedure, of which parts are as described above) and showing that it implied "local intrinsic harmony" (local soundness, i.e. reductions, and local completeness, as illustrated above for $\equiv$). The classification in [8] corresponds roughly but not exactly to the different possibilities enumerated above (1 ... 5): "non-combining" (zero or one premiss[es]) or "combining" (more than one premiss) corresponds to possibility 2; "hypothetical" (a premiss with assumptions discharge) or "categorical" (no such discharge) corresponds to possibility 3; "parametrized" (a premiss depends on a free variable) corresponds roughly to a mixture of 4 and 5; "conditional" (e.g. there is a freshness condition) corresponds roughly to 4.

Let us now consider a combination of such ideas, e.g. two I-rules each of which discharges an assumption, e.g. the pair

$$\dfrac{\begin{array}{c}[A]\\ \vdots\\ B\end{array}}{A \odot B} \qquad \dfrac{\begin{array}{c}[B]\\ \vdots\\ A\end{array}}{A \odot B}$$

What is/are the appropriate GE-rule(s)? It/they might be just

$$\dfrac{A \odot B \quad A \quad \begin{array}{c}[B]\\ \vdots\\ C\end{array}}{C}$$

but that only captures, as it were, the first of the two I-rules (and implies that $(A \odot B) \supset (A \supset B)$, surely not what should be the case); so we have to try also

$$\dfrac{A \odot B \quad B \quad \begin{array}{c}[A]\\ \vdots\\ C\end{array}}{C}$$

[16]Written in 2007–9, several years before publication.

but then these two need to be combined somehow. If into a single rule,[17] it would be something like

$$\dfrac{A \odot B \quad A \quad \begin{array}{c}[B]\\ \vdots \\ C\end{array} \quad B \quad \begin{array}{c}[A]\\ \vdots \\ C\end{array}}{C}$$

which is weird; with only the second and last of these premisses, already C can be deduced. The meaning of $A \odot B$ is thus surely not being captured, whether we go for two GE-rules or just one.

A similar example was given in 1968 by von Kutschera [38, p. 15], with two I-rules for an operator F based on the informal definition $F(A, B, C) \equiv (A \supset B) \vee (C \supset B)$ but the flattened E-rule failing to capture the definition adequately.

4.5 Another [Counter-]Example

Following Zucker and Tragesser's [42, p. 506], Olkhovikov and Schroeder-Heister [21] have given as a simpler example the ternary constant $\star$ with two introduction rules:

$$\dfrac{\begin{array}{c}[A]\\ \vdots \\ B\end{array}}{\star(A, B, C)}\ \star I_1 \qquad \dfrac{C}{\star(A, B, C)}\ \star I_2$$

and the "obvious" GE rule[18] thereby justified is:

$$\dfrac{\star(A, B, C) \quad A \quad \begin{array}{c}[B]\\ \vdots \\ D\end{array} \quad \begin{array}{c}[C]\\ \vdots \\ D\end{array}}{D}\ \star GE$$

which is clearly wrong, there being nothing to distinguish it from $\star(A, C, B)$. Their main point is to show by a semantic argument that there is no non-obvious GE rule for $\star$, thus defending the "idea of higher-level rules" [30].

[17] As the formula in [13] implies, since $1 \times 1 = 1$.

[18] Again, [13] implies there is just one GE rule.

4.6 In Other Words

The "flattening" methodology when either the constant being defined has several introduction rules or one or more of such rules have several premisses can lead

1. to a number (>1) of GE rules, none of which on its own suffices, and
2. to a "disharmonious mess", i.e. a failure to capture the correct meaning.

Already there are enough problems, before we start considering the cases where the premiss abstracts over several variables, instantiates a variable as a term or recurses on the constant being defined.

The solution of Schroeder-Heister [30] is to allow rules to discharge rules. We prefer, however, to propose instead that one should adopt the standard solution from (e.g.) *Coq* [1]: to reject the idea that the rule for handling implication (and other situations where assumptions are discharged) be treated as illustrated above and instead to take implication (and its generalisation, universal quantification), together with an inductive definition mechanism, as primitive, with traditional "special" elimination rules (e.g. *Modus Ponens*) but to allow GE rules elsewhere (e.g. for $\wedge$ and its generalisations Σ and $\exists$). This deals with $\equiv$; likewise, it deals with $\odot$ as if it were

$$\dfrac{A \odot B \qquad \begin{array}{c}[A \supset B] \\ \vdots \\ C\end{array} \qquad \begin{array}{c}[B \supset A] \\ \vdots \\ C\end{array}}{C} \quad .$$

More precisely, we note that with an introduction rule given in *Coq* by the inductive definition

```
Inductive and (A B : Prop) : Prop :=
  and_I : A -> B -> (and A B).
```

we obtain as a theorem

```
Theorem and_elim : forall A B C : Prop,
  (and A B) -> (A -> B -> C) -> C.
```

and similarly for $\odot$ we have the inductive definition

```
Inductive odot (A B : Prop) : Prop :=
| odot_I_1 : (A -> B) -> (odot A B)
| odot_I_2 : (B -> A) -> (odot A B).
```

and we can obtain as a theorem

```
Theorem odot_elim : forall A B C : Prop,
  (odot A B) -> ((A -> B) -> C) -> ((B -> A) -> C) -> C.
```

Not only can we obtain such theorems, but *Coq* will calculate them (and several variants) from the definitions automatically. Further details of this approach can be found in [23]. For example, existential quantification can be defined thus (we give also the obtained theorem representing the elimination rule):

```
Inductive ex (X:Type) (B : X -> Prop) : Prop :=
  ex_intro : forall (w:X), B w -> ex X B.
Theorem ex_elim : forall X : Type, B : X -> Prop,
  C : Prop, (ex X B) -> (forall (x:X) (B x -> C)) -> C.
```

Short shrift is given to •:

```
Inductive bullet : Prop :=
  bullet_I : ( (bullet -> False) -> bullet).
Error: Non strictly positive occurrence of"bullet" in
  "(bullet -> False) -> bullet"
```

This pushes the problem (of constructing and justifying elimination rules given a set of introduction rules, and establishing properties like harmony, local completeness and stability) elsewhere: into the same problem for a mechanism of inductive definitions and for the rules regarded as primitive: introduction and (non-general) elimination rules for implication and universal quantification. Apart from the issue of stability, we regard the latter as unproblematic, and the former as relatively straightforward (once we can base the syntax on implication and universal quantification).

To a large extent this approach may be regarded as just expressing first-order connectives using second-order logic, and not very different from Schroeder-Heister's higher-level rules. The important point is that there are difficulties (we think unsurmountable) with trying to do it all without such higher-order mechanisms.

5 Conclusion

The main conclusion is this: although the idea that the "grounds for asserting a proposition" are easily collected together as a unit is attractive, the different ways in which it can be done (disjunctive, conjunctive, with assumption discharge, with variable abstraction or parameterisation, …, recursion) generate (if the GE rules pattern is followed) many problems for the programme of mechanically generating one (or more) elimination rules for a logical constant, other than in simple cases. There are difficulties with the mechanical approach in [8]; there are similar difficulties in [13]. Without success of such a programme, it is hard to see what "GE harmony" can amount to, except as carried out in (e.g.) *Coq* [1] where strictly positive inductive type definitions lead automatically to rules for reasoning by induction and case analysis over objects of the types thus defined, and with strong normalisation results. A similar conclusion is to be found in [33].

References

1. Bertot, Y., Casteran, P.: Interactive Theorem Proving and Program Development. Springer, Heidelberg (2004)
2. Dummett, M.A.E.: The Logical Basis of Metaphysics. Duckworth, London (1991)
3. Dyckhoff, R.: Implementing a simple proof assistant. In: Derrick, J., Lewis, H.A. (eds.) Workshop on Programming for Logic Teaching, Proceedings, vol. 23.88, pp. 49–59. Centre for Theoretical Computer Science, University of Leeds (1987). http://rd.host.cs.st-andrews.ac.uk/publications/LeedsPaper.pdf
4. Dyckhoff, R.: Generalised elimination rules and harmony. In: Arché Seminar, St Andrews, 26 May 2009. http://rd.host.cs.st-andrews.ac.uk/talks/2009/GE.pdf
5. Dyckhoff, R., Pinto, L.: Cut-elimination and a permutation-free sequent calculus for intuitionistic logic. Studia Logica **60**, 107–118 (1998)
6. Dyckhoff, R., Pinto, L.: Proof search in constructive logics. In: Cooper, S.B., Truss, J.K. (eds.) Proceedings of Logic Colloquium 97, i.e. Sets and Proofs, pp. 53–65. CUP (1999)
7. Ekman, J.: Propositions in propositional logic provable only by indirect proofs. Math. Log. Q. **44**, 69–91 (1998)
8. Francez, N., Dyckhoff, R.: A note on harmony. J. Philos. Log. **41**, 613–628 (2012)
9. Gentzen, G.: Untersuchungen über das logische Schließen. Math. Zeitschrift **39**, 176–210 (1934)
10. Girard, J.Y.: Linear logic. Theor. Comput. Sci. **50**, 1–102 (1987)
11. Girard, J.Y., Lafont, Y., Taylor, P.: Proofs and Types. Cambridge University Press, Cambridge (1989)
12. Joachimski, F., Matthes, R.: Short proofs of normalization for the simply-typed lambda-calculus, permutative conversions and Gödel's T. Arch. Math. Log. **42**, 59–87 (2003)
13. López-Escobar, E.G.K.: Standardizing the N systems of Gentzen. Models, Algebras, and Proofs, pp. 411–434. Dekker, New York (1999)
14. Luo, Z.H.: Computation and Reasoning. Oxford University Press, New York (1994)
15. Martin-Löf, P.: Intuitionistic Type Theory. Bibliopolis, Naples (1986)
16. Martin-Löf, P.: On the meanings of the logical constants and the justifications of the logical laws (the Siena Lectures). Nord. J. Philos. Log. **1**, 11–60 (1996)
17. Matthes, R.: Interpolation for natural deduction with generalized eliminations. In: Kahle, R., Schroeder-Heister, P., Stärk, R. (eds.) Proof Theory in Computer Science. LNCS, vol. 2183, pp. 153–169. Springer, New York (2001)
18. Mendler, P.F. (Better known as Nax P. Mendler.): Inductive definition in type theory. Ph.D. thesis, Cornell (1987). http://www.nuprl.org/documents/Mendler/InductiveDefinition.pdf
19. Moriconi, E., Tesconi, L.: On inversion principles. Hist. Philos. Log. **29**, 103–113 (2008)
20. Negri, S., von Plato, J.: Structural Proof Theory. CUP, Cambridge (2000)
21. Olkhovikov, G.K., Schroeder-Heister, P.: On flattening elimination rules. Rev. Symb. Log. 13 pp. (2014). http://dx.doi.org/10.1017/S1755020313000385
22. Pfenning, F., Davies, R.: A judgmental reconstruction of modal logic. Math. Struct. Comput. Sci. **11**, 511–540 (2001)
23. Pierce, B., et al.: Software Foundations. http://www.cis.upenn.edu/~bcpierce/sf/Logic.html 28 July 2013. Accessed 27 Jan 2014
24. Prawitz, D.: Natural Deduction. Almqvist and Wiksell, Stockholm (1965)
25. Prawitz, D.: Ideas and results in proof theory. In: Fenstad, J.E. (ed.) Second Scandinavian Logic Symposium Proceedings, pp. 235–307. North-Holland, Amsterdam (1971)
26. Prawitz, D.: Proofs and the meaning and completeness of the logical constants. Essays on Mathematical and Philosophical Logic. Reidel, Dordrecht (1979)
27. Prawitz, D.: Beweise und die Bedeutung und Vollständigkeit der logischen Konstanten. Conceptus **16**, 31–44 (1982) (Translation and revision of [26])
28. Read, S.L.: Harmony and autonomy in classical logic. J. Philos. Log. **29**, 123–154 (2000)
29. Read, S.L.: General-elimination harmony and higher-level rules. In: Wansing, H. (ed.) Dag Prawitz on Proofs and Meaning, pp. 293–312. Springer, New York (2015)

30. Schroeder-Heister, P.: A natural extension of natural deduction. J. Symb. Log. **49**, 1284–1300 (1984)
31. Schroeder-Heister, P.: Generalized definitional reflection and the inversion principle. Log. Univ. **1**, 355–376 (2007)
32. Schroeder-Heister, P.: Generalized elimination inferences, higher-level rules, and the implications-as-rules interpretation of the sequent calculus. In: Pereira, L.C., Haeusler, E.H., de Paiva, V. (eds.) Advances in Natural Deduction: A Celebration of Dag Prawitz's Work, pp. 1–29. Springer, New York (2014)
33. Schroeder-Heister, P.: Harmony in proof-theoretic semantics: a reductive analysis. In: Wansing, H. (ed.) Dag Prawitz on Proofs and Meaning, pp. 329–358. Springer, New York (2015)
34. Schroeder-Heister, P., Tranchini, L.: Ekman's paradox (2014). Notre Dame Journal of Formal Logic (submitted)
35. Schwichtenberg, H., Wainer, S.: Proofs and Computations. Cambridge University Press, Cambridge (2012)
36. Tennant, N.: Ultimate normal forms for parallelized natural deductions. Log. J. IGPL **10**, 299–337 (2002)
37. Tesconi, L.: Strong normalization for natural deduction with general elimination rules. Ph.D. thesis, Pisa (2004)
38. von Kutschera, F.: Die Vollständigkeit des Operatorensystems für die intuitionistiche Aussagenlogik im Rahmen der Gentzensemantik. Archiv für mathematische Logik und Grundlagenforschung **11**, 3–16 (1968)
39. von Plato, J.: Natural deduction with general elimination rules. Arch. Math. Log. **40**, 541–567 (2001)
40. von Plato, J.: Explicit composition and its application in proofs of normalization (2015) (submitted)
41. von Plato, J., Siders, A.: Normal derivability in classical natural deduction. Rev. Symb. Log. **5**, 205–211 (2012)
42. Zucker, J., Tragesser, R.S.: The adequacy problem for inferential logic. J. Philos. Log. **7**, 501–516 (1978)

Categorical Harmony and Paradoxes in Proof-Theoretic Semantics

Yoshihiro Maruyama

Abstract There are two camps in the theory of meaning: the referentialist one including Davidson, and the inferentialist one including Dummett and Brandom. Proof-theoretic semantics is a semantic enterprise to articulate an inferentialist account of the meaning of logical constants and inferences within the proof-theoretic tradition of Gentzen, Prawitz, and Martin-Löf, replacing Davidson's path "from truth to meaning" by another path "from proof to meaning". The present paper aims at contributing to developments of categorical proof-theoretic semantics, proposing the principle of categorical harmony, and thereby shedding structural light on Prior's "tonk" and related paradoxical logical constants. Categorical harmony builds upon Lawvere's conception of logical constants as adjoint functors, which amount to double-line rules of certain form in inferential terms. Conceptually, categorical harmony supports the iterative conception of logic. According to categorical harmony, there are intensional degrees of paradoxicality of logical constants; in the light of the intensional distinction, Russell-type paradoxical constants are maximally paradoxical, and tonk is less paradoxical. The categorical diagnosis of the tonk problem is that tonk mixes up the binary truth and falsity constants, equating truth with falsity; hence Prior's tonk paradox is caused by equivocation, whereas Russell's paradox is not. This tells us Prior's tonk-type paradoxes can be resolved via disambiguation while Russell-type paradoxes cannot. Categorical harmony thus allows us to demarcate a border between tonk-type pseudo-paradoxes and Russell-type genuine paradoxes. I finally argue that categorical semantics based on the methods of categorical logic might even pave the way for reconciling and uniting the two camps.

Keywords Categorical proof-theoretic semantics · Categorical harmony · Iterative conception of logic · Prior's tonk · Degrees of paradoxicality of logical constants

Y. Maruyama (✉)
Quantum Group, Department of Computer Science, University of Oxford, Wolfson Building, Parks Road, Oxford OX1 3QD, UK
e-mail: maruyama@cs.ox.ac.uk

T. Piecha and P. Schroeder-Heister (eds.), *Advances in Proof-Theoretic Semantics*, Trends in Logic 43, DOI 10.1007/978-3-319-22686-6_6

1 Introduction

Broadly speaking, there are two conceptions of meaning: the referentialist one based on truth conditions as advocated by Davidson [7], and the inferentialist one based on verification or use conditions as advocated by Dummett [10] or more recent Brandom [5]. Along the latter strand of the theory of meaning, proof-theoretic semantics undertakes the enterprise of accounting for the meaning of logical constants and inferences in terms of proof rather than truth, thus replacing Davidson's path "from truth to meaning" by another Dummettian path "from proof to meaning"; the term "proof-theoretic semantics" was coined by Schroeder-Heister (for a gentle introduction, see Schroeder-Heister [21]; he also coined the term "substructural logic" with Došen). It builds upon the proof-theoretic tradition of Gentzen, Prawitz, and Martin-Löf, tightly intertwined with developments of Brouwer's intuitionism and varieties of constructive mathematics, especially the Brouwer-Heyting-Kolmogorov interpretation, and its younger relative, the Curry-Howard correspondence between logic and type theory, or rather the Curry-Howard-Lambek correspondence between logic, type theory, and category theory (see, e.g., Lambek and Scott [12]). Note however that Brouwer himself objected to the very idea of formal logic, claiming the priority of mathematics to logic (cf. Hilbert's Kantian argument in [11] concluding: "Mathematics, therefore, can never be grounded solely on logic"). "Harmony" in Dummett's terms and the justification of logical laws have been central issues in proof-theoretic semantics (see, e.g., Dummett [10] and Martin-Löf [14]).

Combining proof-theoretic semantics with category theory (see, e.g., Awodey [2]), the present paper aims at laying down a foundation for categorical proof-theoretic semantics, proposing the principle of categorical harmony, and thereby shedding structural light on Prior's "tonk" and related paradoxical logical constants. Prior's invention of a weird logical connective "tonk" in his seminal paper [18] compelled philosophical logicians to articulate the concept of logical constants, followed by developments of the notion of harmony. In a way harmony prescribes the condition of possibility for logical connectives or their defining rules to be meaning-conferring, and as such, it works as a conceptual criterion to demarcate pseudo-logical constants from genuine logical constants. Let us recall the definition of Prior's tonk. Tonk can be defined, for example, by the following rules of inference as in the system of natural deduction:

$$\frac{\xi \vdash \varphi}{\xi \vdash \varphi \,\mathrm{tonk}\, \psi}\ \text{(tonk-intro.)} \qquad \frac{\xi \vdash \varphi \,\mathrm{tonk}\, \psi}{\xi \vdash \psi}\ \text{(tonk-elim.)} \tag{1}$$

In any standard logical system (other than peculiar systems as in Cook [6]), adding tonk makes (the deductive relation of) the system trivial, and thus we presumably ought not to accept tonk as a genuine logical constant. In order to address the tonk problem, different principles of harmony have been proposed and discussed by Belnap [3], Prawitz [17], Dummett [10], and many others. From such a point of view, harmony is endorsed to ban tonk-like pathological connectives on the ground that

their defining rules violate the principle of harmony, and are not meaning-conferring as a consequence of the violation of harmony. They cannot be justified after all.

In the present paper, we revisit the tonk problem, a sort of demarcation problem in philosophy of logic, from a novel perspective based on category theory. In his seminal paper [13], Lawvere presented a category-theoretical account of logical constants in terms of adjoint functors, eventually giving rise to the entirely new discipline of categorical logic (see, e.g., Lambek and Scott [12]) including categorical proof theory. Note that "categorical logic" is sometimes called "categorial logic" in philosophy to avoid confusion with the philosopher's sense of "categorical" (like "categorical judgements"); especially, Došen, a leading researcher in the field, uses "categorial logic", even though "categorical logic" is widely used in the category theory community. Categorical logic is relevant to proof-theoretic semantics in various respects, as explicated in this paper as well. Among other things, a fundamental feature of the link between categorical logic and proof theory is that categorical adjunctions amount to bi-directional inferential rules (aka double-line rules) of certain specific form in terms of proof theory; by this link, categorical terms can be translated into proof-theoretic ones. Building upon Lawvere's understanding of logical constants, in this paper, I formulate the adjointness-based principle of categorical harmony, and compare it with other notions of harmony and related principles, including Belnap's harmony [3], Došen's idea of logical constants as punctuation marks [8, 9], and the reflection principle and definitional equations by Sambin et al. [19]. And the final aim of the present paper is to shed new light on the tonk problem from the perspective of categorical harmony.

In comparison with other concepts of harmony, there is a sharp contrast between categorical harmony and Belnap's harmony in terms of conservativity; at the same time, however, uniqueness is a compelling consequence of categorical harmony, and so both endorse uniqueness, yet for different reasons. The principle of categorical harmony looks quite similar to Došen's theory of logical constants as punctuation marks, and also to the theory of the reflection principle and definitional equations by Sambin et al. It nevertheless turns out that there are striking differences: some logical constants are definable in Sambin's or Došen's framework, but not definable according to categorical harmony, as we shall see below. It is not obvious at all whether this is an advantage of categorical harmony or not. It depends upon whether those logical constants ought to count as genuine logical constants, and thus upon our very conception of logical constants; especially, what is at stake is the logical status of substructural connectives (aka multiplicative connectives; in categorical terms, monoidal connectives).

Finally, several remarks would better be made in order to alleviate common misunderstandings on categorical semantics. The Curry-Howard correspondence is often featured with the functional programmer's dictum "propositions-as-types, proofs-as-programs". Likewise, the Curry-Howard-Lambek correspondence may be characterised by the categorical logician's dictum "propositions-as-objects, proofs-as-morphisms." One has to be careful of categorical semantics, though. For the Curry-Howard-Lambek correspondence does not hold in some well-known categorical semantics. The correspondence surely holds in the cartesian (bi)closed category

semantics for propositional intuitionistic logic, yet at the same time, it is not true at all in the topos semantics for higher-order intuitionistic logic (or intuitionistic ZF set theory), and it does not even hold in the logos (aka Heyting category) semantics for first-order intuitionistic logic. For the very facts, the latter two semantics are called proof-irrelevant: they only allow for completeness with respect to the identity of propositions, and does not yield what is called full completeness, i.e., completeness with regard to the identity of proofs. Some category theorists who do not care about the proof-relevance of semantics tend to say that the topos semantics is a generalisation of the cartesian (bi)closed category semantics; however, the claim is only justified in view of the identity of propositions, and it is indeed wrong in light of the identity of proofs, which is not accounted for in the topos semantics at all. Note that locally cartesian closed categories yield proof-relevant semantics for Martin-Löf's dependent type theory, and toposes are locally cartesian closed. If we see toposes as semantics for dependent type theory, then the topos semantics is proof-relevant; yet, this is an unusual way to interpret the term "topos semantics". The Curry-Howard-Lambek correspondence is now available for a broad variety of logical systems including substructural logics as well as intuitionistic logic, yet with the possible exception of classical logic. Although quite some efforts have been made towards semantics of proofs for classical logic, there is so far no received view on it, and there are impossibility theorems on semantics of classical proofs, including the categorical Joyal lemma. For classical linear logic, nonetheless, we have fully complete semantics in terms of so-called $*$-autotonomous categories.

The rest of the paper is organised as follows. In Sect. 2, we review Lawvere's categorical account of logical constants, and then formulate the principle of categorical harmony. There are several subtleties on how to formulate it, and naïve formulations cannot properly ban tonk-type or Russell-type logical constants. Comparison with Došen's theory is given as well. In Sect. 3, we further compare the principle of categorical harmony with Belnap's harmony conditions and Sambin's reflection principle and definitional equations. The logical status of multiplicative (monoidal) connectives is discussed as well, and three possible accounts (i.e., epistemic, informational, and physical accounts) are given. In Sect. 4, we look at tonk and other paradoxical logical constants on the basis of categorical harmony, thus exposing different degrees of paradoxicality among them. Especially, what is wrong with tonk turns out to be equivocation. The paper is then concluded with prospects on the broader significance of categorical logic in view of the theory of meaning. Little substantial knowledge on category theory is assumed throughout the paper.

2 The Principle of Categorical Harmony

In this section, we first see how logical constants can be regarded as adjoint functors, and finally lead to the principle of categorical harmony. Although I do not explain the concept of categories from the scratch, from a logical point of view, you may

conceive of a category as a sort of proof-theoretic consequence relation: suppose we have the following concepts given.

- The concept of formulae φ.
- The concept of hypothetical proofs (or deductions) from formulae φ to ψ. For any formula φ there must be an identity proof id_φ from φ to φ.
- The concept of proof-decorated relation $\vdash_p$:

$$\varphi \vdash_p \psi \tag{2}$$

  where p is a proof from φ to ψ.
- The concept of sequential composition $\circ$ of proofs: composing $\varphi \vdash_p \psi$ and $\psi \vdash_q \xi$, we obtain

$$\varphi \vdash_{q \circ p} \xi. \tag{3}$$

  If we think of a monodical category for substructural logic, we additionally have parallel composition $\otimes$:

$$\varphi \otimes \varphi' \vdash_{p \otimes p'} \psi \otimes \psi' \tag{4}$$

  where $\varphi \vdash_p \psi$ and $\varphi' \vdash_{p'} \psi'$.
- The concept of reduction of proofs such that the identity proofs may be canceled out (i.e., $id_\varphi \circ p$ equals p modulo reducibility; $q \circ id_\varphi$ equals q modulo reducibility), and proofs may locally be reduced in any order (i.e., $(p \circ q) \circ r$ equals $p \circ (q \circ r)$ modulo reducibility). Moreover, reduction must respect composition (i.e., if p equals q and r equals s modulo reducibility, $p \circ q$ must equal $r \circ s$ modulo reducibility).

The concept of a proof-theoretic consequence relation (resp. with parallel composition) thus defined is basically the same as the concept of a category (resp. monoidal category): a formula corresponds to an object in a category, and a proof to a morphism in it. To be precise, a proof-theoretic consequence relation is a way of presenting a category rather than category *per se*. In most parts of the article, however, full-fledged category theory shall not be used, for simplicity and readability, and it suffices for the reader to know some basic logic and order theory, apart from occasional exceptions.

From the perspective of Schroeder-Heister [23], categorical logic (or categorial logic to avoid confusion) places primary emphasis on hypothetical judgements, which are concerned with the question "what follows from what?", rather than categorical judgements in the philosopher's sense, which are concerned with the question "what holds on their own?". In the traditional accounts of both model-theoretic and proof-theoretic semantics, the categorical is prior to the hypothetical, and the hypothetical is reduced to the categorical via the so-called transmission view of logical consequence. These are called dogmas of standard semantics, whether model-theoretic or proof-theoretic, in Schroeder-Heister [23]. His theory takes the hypothetical to precede the categorical, and it is in good harmony with the idea of categorical logic, in which the hypothetical, i.e., the concept of morphisms, is conceptually prior to the

categorical, i.e., the concept of morphisms with their domains being a terminal object (or monoidal unit in the case of substructural logic).

Some authors have discussed how logical constants can be derived from logical consequence relations (see, e.g., Westerståhl [25]). To that end, category theory allows us to derive logical constants from abstract proof-theoretic consequence relations (i.e., categories) through the concept of adjunctions, which even give us inferential rules for the derived logical constants, and hence a proof system as a whole (in the case of intuitionistic logic, for example, the proof system thus obtained is indeed equivalent to standard ones, such as NJ and LJ). Given a category (abstract proof-theoretic consequence relation), we can always mine logical constants (if any) in the category via the generic criteria of adjunctions. In such a way, category-theoretical logic elucidates a generic link between logical constants and logical consequence, without focusing on a particular system of logic.

In the following, let us review the concept of adjoint functors in the simple case of preorders, which is basically enough for us, apart from occasional exceptions. A preorder

$$(L, \vdash_L) \tag{5}$$

consists of a set P with a reflexive and transitive relation $\vdash_L$ on L. Especially, the deductive relations of most logical systems form preorders; note that reflexivity and transitivity amount to identity and cut in logical terms. It is well known that a preorder can be seen as a category in which the number of morphisms between fixed two objects in it are at most one. Then, a functor $F : L \to L'$ between preorders L and L' is just a monotone map: i.e., $\varphi \vdash_L \psi$ implies $F(\varphi) \vdash_{L'} F(\psi)$. Now, a functor

$$F : L \to L' \tag{6}$$

is called left adjoint to

$$G : L' \to L \tag{7}$$

(or G is right adjoint to F) if and only if

$$F(\varphi) \vdash_{L'} \psi \;\Leftrightarrow\; \varphi \vdash_L G(\psi) \tag{8}$$

for any $\varphi \in L$ and $\psi \in L'$. This situation of adjunction is denoted by $F \dashv G$. Note that a left or right adjoint of a given functor does not necessarily exist.

In this formulation, it would already be evident that an adjunction $F \dashv G$ is equivalent to a sort of bi-directional inferential rule:

$$\frac{F(\varphi) \vdash_{L'} \psi}{\varphi \vdash_L G(\psi)} \tag{9}$$

where the double line means that we can infer the above "sequent" from the one below, and vice versa.

Let us look at several examples to illustrate how logical constants are characterised by adjunctions, and to articulate the inferential nature of them. Suppose that L is intuitionistic logic. We define the diagonal functor $\Delta : L \to L \times L$ by

$$\Delta(\varphi) = (\varphi, \varphi) \tag{10}$$

where $\times$ denotes binary product on categories (in this case just preorders) or in the category of (small) categories. Then, the right adjoint of Δ is conjunction $\wedge : L \times L \to L$:

$$\Delta \dashv \wedge. \tag{11}$$

The left adjoint of Δ is disjunction $\vee : L \times L \to L$:

$$\vee \dashv \Delta. \tag{12}$$

The associated bi-directional rule for $\wedge$ turns out to be the following:

$$\frac{\Delta(\varphi_1) \vdash_{L \times L} (\varphi_2, \varphi_3)}{\varphi_1 \vdash_L \varphi_2 \wedge \varphi_3} \tag{13}$$

which, by the definition of product on categories, boils down to the following familiar rule:

$$\frac{\varphi_1 \vdash_L \varphi_2 \quad \varphi_1 \vdash_L \varphi_3}{\varphi_1 \vdash_L \varphi_2 \wedge \varphi_3} \tag{14}$$

This is the inferential rule for conjunction $\wedge$ that is packed in the adjunction $\Delta \dashv \wedge$. We omit the case of $\vee$. Now, implication $\varphi \to (\text{-}) : L \to L$ for each φ is the right adjoint of $\varphi \wedge (\text{-})$, where the expressions "$\varphi \to (\text{-})$" and "$\varphi \wedge (\text{-})$" mean that φ is fixed, and (-) is an argument:

$$\varphi \wedge (\text{-}) \ \dashv \ \varphi \to (\text{-}). \tag{15}$$

The derived inferential rule is the following:

$$\frac{\varphi \wedge \xi \vdash \psi}{\xi \vdash \varphi \to \psi} \tag{16}$$

Note that we may replace $\wedge$ by comma in the format of sequent calculus. Quantifiers can be treated in a similar (but more heavily categorical) way using indexed or fibrational structures $(L_C)_{C \in \mathbf{C}}$ (intuitively, each object C in a category $\mathbf{C}$ is a collection of variables or a so-called context) rather than single L as in the above discussion (see, e.g., Pitts [16] for the case of intuitionistic logic; for a general variety of substructural logics over full Lambek calculus, categorical treatment of quantifiers is presented in Maruyama [15]). The corresponding double-line rules are the following:

$$\frac{\varphi \vdash \psi}{\varphi \vdash \forall x \psi} \qquad \frac{\exists x \varphi \vdash \psi}{\varphi \vdash \psi} \tag{17}$$

with the obvious eigenvariable conditions (which naturally emerge from a categorical setting). Categorical logicians say that $\forall$ is a right adjoint, and $\exists$ is a left adjoint of substitution (or pullback in categorical terms). Finally, truth constants $\bot$ and $\top$ are left and right adjoints of the unique operation from L to the one-element set $\{*\}$, with the following double-line rules:

$$\frac{\bot \vdash \varphi}{* \vdash *} \qquad \frac{* \vdash *}{\varphi \vdash \top} \tag{18}$$

which come down to:

$$\frac{}{\bot \vdash \varphi} \qquad \frac{}{\varphi \vdash \top} \tag{19}$$

All the double-line rules above yield a sound and complete axiomatisation of intuitionistic logic; equivalence with other standard systems can easily be verified.

Building upon these observations, we can articulate the categorical inferentialistic process of introducing a logical constant in a meaning-conferring manner:

- At the beginning, there are universally definable operations, i.e., those operations that are definable in the general language of category theory.
  - We may replace "the general language of category theory" by "the general language of monoidal category theory" if we want to account for substructural logics as well as logics with unrestricted structural rules.
  - For example, diagonal Δ above is a universally definable operation. As observed in the above case of the double-line rule for $\wedge$, the existence of Δ amounts to our meta-theoretical capacity to handle multiple sequents at once (in particular, ability to put two sequents in parallel in the case of $\wedge$).
- Logical constants are introduced step by step, by requiring the existence of right or left adjoints of existing operations, i.e., universally definable operations or already introduced logical constants.
  - In other words, we define logical constants by bi-directional inferential rules corresponding to adjunctions concerned. Thus, this may be conceived of as a special sort of inferentialistic process to confer meaning on connectives. The condition of adjointness bans non-meaning-conferring rules like tonk's (discussed later).
  - For example, conjunction and disjunction above can be introduced as adjoints of a universally definable operation (i.e., diagonal); after that, implication can be introduced as an adjoint of an existing logical constant (i.e., conjunction).
- Genuine logical constants are those introduced according to the above principle, namely the principle of categorical harmony. Others are pseudo-logical constants.

According to this view, logical constants in a logical system must be constructed step by step, from old simple to new complex ones, based upon different adjunctions. This

may be called the iterative conception of logic. The rôle of the powerset operation (or making a set of already existing sets) in the iterative conception of sets is analogous to the rôle of the operation of taking adjoint functors.

A remark on monoidal categories for substructural logics is that the language of monoidal categories is more general than the language of (plain) categories in the sense that monoidal products $\otimes$ encompass cartesian products $\times$. Formally, we have the fact that a monoidal product $\otimes$ is a cartesian product if and only if there are both diagonals $\delta : C \to C \otimes C$ and projections $p_1 : C \otimes D \to C$ and $p_2 : C \otimes D \to D$ where $\otimes$ is assumed to be symmetric. The logical counterpart of this fact is that multiplicative conjunction $\otimes$ is additive conjunction if and only if both contraction and weakening hold where exchange is assumed. Since contraction may be formulated as $\varphi \vdash \varphi \otimes \varphi$, and weakening as $\varphi \otimes \psi \vdash \varphi$ and $\varphi \otimes \psi \vdash \psi$, it is evident that diagonals correspond to contraction, and projections to weakening (this correspondence can be given a precise meaning in terms of categorical semantics).

There are different conceptions of harmony in proof-theoretic semantics, discussed by different authors. In the present article, adjointness is conceived of as a sort of proof-theoretic harmony, and it is somehow akin to Prawitz's inversion principle in that both put emphasis on (different sorts of) "invertibility" of rules; recall that an adjunction amounts to the validity of a "bi-directional" rule of certain form. Categorically speaking, adjointness exhibits harmony between two functors; logically speaking, adjointness tells us harmony between the upward and the downward rules of the induced bi-directional rule. The precise procedure of introducing logical constants according to categorical harmony has already been given above. Let us summarise the main point of categorical harmony as follows.

- A logical constant must be introduced by (the double-line rule of) an adjunction with respect to an existing operation.

As we observed above, standard logical constants can be characterised by adjunctions or adjunction-induced double-line rules. The idea of capturing logicality by double-line rules was pursued by Došen [8, 9]. It seems, however, that his focus was not on harmony, but rather on logicality only (as pointed out by Schroeder-Heister [21]), and moreover he did not really use adjointness as a criterion to ban pathological, non-meaning-conferring rules. Indeed, Bonnay and Simmenauer [4] show that Došen's theory of logicality cannot ban a weird connective "blonk"; nonetheless, the adjointness harmony of the present paper is immune to blonk, since it is not definable by an adjunction, even though it is defined by a double-line rule. The approach of this paper takes adjointness as the primary constituent of harmony, analysing issues in proof-theoretic semantics from that particular perspective. Although the double-line and adjointness approaches are quite similar at first sight, however, they are considerably different as a matter of fact, as seen in the case of Bonnay and Simmenauer's blonk. There are actually several subtleties lurking behind the formulation above:

- It turns out that definability via one adjunction is crucial, since tonk can be defined via two adjunctions.
- A logical constant must be defined as an adjoint of an existing operation, since Russell-type paradoxical constants can be defined as adjoints of themselves.

These points shall be addressed later in detail. Before getting into those issues, in the next section, we briefly compare and contrast categorical harmony with other principles.

3 Categorical Harmony in Comparison with Other Principles

Here we have a look at relationships with Belnap's harmony and the so-called reflection principle and definitional equations by Sambin and his collaborators.

The categorical approach to harmony poses several questions to Belnap's notion of harmony. As we saw above, implication $\rightarrow$ in intuitionistic logic is right adjoint to conjunction $\wedge$. Suppose that we have a logical system L with logical constants $\wedge$ and $\vee$ only, which are specified as the right and left adjoints of diagonal Δ as in the above. And suppose we want to add implication $\rightarrow$ to L. Of course, this can naturally be done by requiring the right adjoint of $\wedge$. Now, Freyd's adjoint functor theorem tells us that any right adjoint functor preserves limits (e.g., products), and any left adjoint functor preserves colimits (e.g., coproducts). This is a striking characteristic of adjoint functors. In the present case, the theorem tells us that $\wedge$ preserves $\vee$; in other words, $\wedge$ distributes over $\vee$. Thus, defining implication according to categorical harmony is not conservative over the original system L, since the bi-directional rules for $\wedge$ and $\vee$ only never imply the distributive law. Note that sequent calculus for $\wedge$ and $\vee$ allows us to derive the distributive law without any use of implication; yet the bi-directional rules alone do not imply it.

Although proponents of Belnap's harmony would regard this as a strange (and perhaps unacceptable) feature, nevertheless, this sort of non-conservativity is necessary and natural from a category-theoretical point of view. Furthermore, conservativity may be contested in some way or other. One way would be to advocate categorical harmony against Belnap's on the ground of the Quinean holistic theory of meaning, which implies that the meaning of a single logical constant in a system, in principle, can only be determined by reference to the global relationships with all the other logical constants in the whole system. If the meaning of a logical constant depends on the whole system, then adding a new logical constant may well change the meaning of old ones. Non-conservativity on logical constants is arguably a consequence of a form of holism on meaning, even though it violates Belnap's harmony condition. Anyway, we may at least say that the principle of categorical harmony, or Lawvere's idea of logical constants as adjoints, is in sharp conflict with Belnap's notion of harmony, in terms of the conservativity issue.

Another distinctive characteristic of adjoint functors is that any of a right adjoint and a left adjoint of a functor is uniquely determined (up to isomorphism). By this very fact, we are justified to define a concept via an adjunction. This actually implies that Belnap's uniqueness condition automatically holds if we define a logical constant according to the principle of categorical harmony. Thus, uniqueness is not something

postulated in the first place; rather, it is just a compelling consequence of categorical harmony. However, it is not very obvious whether this is really a good feature or not. As a matter of fact, for example, exponentials (!, ?) in linear logic do not enjoy the uniqueness property (as noted in Pitts [16]; it is essentially because there can be different (co)monad structures on a single category). At the same time, however, we could doubt that exponentials are genuine logical constants. Indeed, it is sometimes said that they were introduced by Girard himself not as proper logical constants but as a kind of device to analyse structural rules. The rôle of exponentials is to have control on resources in inference, and not to perform inference *per se* on their own. It would thus be a possible view that exponentials are a sort of "computational constants" discriminated from ordinary logical constants. This is an issue common to both categorical harmony and Belnap's harmony.

There are even more subtleties on uniqueness in categorical harmony, which involve a tension between cartesian and monoidal structures in category theory. When formulating the categorical procedure to introduce logical constants in the last section, it was remarked that we may replace the language of (plain) category theory with that of monoidal category theory if we want to treat substructural logics as well. In such a case, we first have a monoidal product $\otimes$ in our primitive language, and then require, for example, a right adjoint of $\otimes$, which functions as multiplicative implication. Since any adjoint is unique, there appears to be no room for non-uniqueness. However, the starting point $\otimes$ may not be unique if it cannot be characterised as an adjoint functor, and you can indeed find many such cases in practice. The point is that, in general, monoidal structures can only be given from "outside" categories, i.e., the same one category can have different monoidal structures on it. If we have both $\otimes$ and the corresponding implication $\rightarrow$, then $\otimes$ is a left adjoint of $\rightarrow$. However, if we do not have implication, then $\otimes$ may not be characterised as an adjoint, and thus may not be unique. This is the only room for non-uniqueness in categorical harmony, since any other logical constant must be introduced as an adjoint in the first place.

From a proof-theoretic point of view, having a monoidal structure on a category amounts to having the comma "," as a punctuation mark in the meta-language of sequent calculus. In sequent calculus, we are allowed to put sequents in parallel (otherwise we could not express quite some rules of inference), and at the same time, we are allowed to put formulae in parallel inside a sequent by means of commas. The former capacity corresponds to the categorical capacity to have cartesian products, and the latter corresponds to the capacity to have monoidal products. This seems relevant to the following question. Why can monoidal structures $\otimes$ be allowed in category theory in spite of the fact that in general they cannot be defined via universal mapping properties? To put it in terms of categorical harmony, why can monoidal structures $\otimes$ be allowed as primitive vocabularies to generate logical constants? (And why are others not allowed as primitive vocabularies?) This is a difficult question, and there would be different possible accounts of it. One answer is that there is no such reason, and $\otimes$ ought not to be accepted as primitive vocabularies in the principle of categorical harmony. Yet I would like to seek some conceptual reasons for permitting $\otimes$ as primitive vocabularies in the following.

For one thing, $\otimes$ is presumably grounded upon a sort of our epistemic capacity to put symbols in parallel (inside and outside sequents) as discussed above. The epistemic capacity may be so fundamental that it plays fundamental rôles in symbolic reasoning as well as many other cognitive practices; this will lead to a sort of epistemic account of admissibility of $\otimes$ in the principle of categorical harmony. Another "informational" account of it seems possible as well. There are three fundamental questions: What propositions hold? Why do they hold? How do they hold? The first one is about truth and falsity, the second one about proofs, and the last one about the mechanisms of proofs. An answer to the last question must presumably include an account of the way how resources or assumptions for inference are used in proofs, or how relevant inferential information is used in proofs. And $\otimes$ may be seen as a means to address that particular part of the third question. This is the informational account, which has some affinities with the view of linear logic as the logic of resources.

Yet another "physical" account may be came up with. In recent developments of categorical quantum mechanics by Abramsky and Coecke (see Abramsky [1] and references therein), the capacities to put things in parallel as well as in sequence play vital rôles in their so-called graphical calculus for quantum mechanics and computation, where parallel composition represents the composition of quantum systems (resp. processes), i.e., the tensor product of Hilbert spaces (resp. morphisms), which is crucial in quantum phenomena involving entanglement, such as the Einstein-Podolsky-Rosen paradox and the violation of the Bell inequality. In general, $\otimes$ lacks diagonals and projections, unlike cartesian $\times$, and this corresponds to the No-Cloning and No-Deleting theorems in quantum computation stating that quantum information can neither be copied nor deleted (note that diagonals $\Delta : X \to X \otimes X$ copy information X, and projections $p : X \otimes Y \to X$ delete information Y). On the other hand, classical information can be copied and deleted as you like. So, the monoidal feature of $\otimes$ witnesses a crucial border between classical and quantum information. To account for such quantum features of the microscopic world, we do need $\otimes$ in the logic of quantum mechanics, and this would justify to add $\otimes$ to primitive vocabularies.

The physical account seems relevant to the well-known question "Is logic empirical?", which was originally posed in the context of quantum logic, and has been discussed by Quine, Putnam, Dummett, and actually Kripke (see Stairs [24]). The need of multiplicative $\otimes$ in the "true" logic of quantum mechanics is quite a recent issue which has not been addressed in the philosophy community yet, and this may have some consequences to both the traditional question "Is logic empirical?" and the present question "Why are substructural logical constants are so special?", as partly argued above. A more detailed analysis of these issues will be given somewhere else.

Sambin et al. [19] present a novel method to introduce logical constants by what they call the reflection principle and definitional equalities, some of which are as follows:

- $\varphi \vee \psi \vdash \xi$ iff $\varphi \vdash \xi$ and $\psi \vdash \xi$.
- $\varphi, \psi \vdash \xi$ iff $\varphi \otimes \psi \vdash \xi$.
- $\Gamma \vdash \varphi \to \psi$ iff $\Gamma \vdash (\varphi \vdash \psi)$.

As these cases show, definitional equalities are quite similar to adjointness conditions in categorical harmony (when they are formulated as bi-directional rules), even though Sambin et al. do not mention category theory at all. Especially, in the case of additive connectives, their definitional equivalences are exactly the same as the bi-directional rules induced by the corresponding adjunctions. There are crucial differences, however. Among them, the following fact should be emphasised:

- Definitional equalities do not always imply adjointness, partly due to what they call the "visibility" condition, which requires us to restrict context formulae in sequent-style rules of inference (categorically, this amounts to restricting so-called Frobenius reciprocity conditions).
  - For example, implication is not necessarily a right adjoint of conjunction in the system of "basic logic" derived via their guiding principles.

This deviation from adjointness actually seems to be inevitable for Sambin et al., because they want to include Birkhoff-von Neumann's quantum logic with some concept of implication as a structural extension of their basic logic; however, quantum implication (if any) cannot be a right adjoint of conjunction, due to the non-distributive nature of it, which is essential in Birkhoff-von Neumann's quantum logic to account for superposition states in quantum systems.

In contrast, categorical harmony cannot allow for any sort of non-adjoint implication. Is this a good feature or not? It depends on whether such implication counts as genuine implication, and so on our very conception of logical constants. The categorical logician's answer would be no: for example, Abramsky [1] asserts that Birkhoff-von Neumann's quantum logic is considered to be "non-logic" because it does not have any adequate concept of implication (on the other hand, categorical quantum logic is said to be "hyper-logic").

Finally, it should be noted that Schroeder-Heister [21] compares the framework of Sambin et al. [19] with his framework of definitional reflection, and that Bonnay and Simmenauer [4] proposes to exploit the idea of Sambin et al. [19] in order to remedy the aforementioned defect (the "blonk" problem) of Došen's double-line approach in [8, 9].

4 Degrees of Paradoxicality of Logical Constants

In this section, we first discuss whether tonk is an adjoint functor or not, or whether tonk counts as a genuine logical constant according to categorical harmony, and we finally lead to the concept of intensional degrees of paradoxicality.

Let L be a (non-trivial) logical system with a deductive relation $\vdash_L$ admitting identity and cut. And suppose L contains truth constants $\bot$ and $\top$, which are specified by adjunction-induced rules $\bot \vdash \varphi$ and $\varphi \vdash \top$, respectively. The first straightforward observation is that, if L has tonk, then tonk has both left and right adjoints, and thus tonk is the left and right adjoint of two functors. Recall that the inferential rôle of

tonk is given by:

$$\frac{\xi \vdash \varphi}{\xi \vdash \varphi \,\mathrm{tonk}\, \psi} \qquad \frac{\xi \vdash \varphi \,\mathrm{tonk}\, \psi}{\xi \vdash \psi} \tag{20}$$

which are equivalent to the following simpler rules in the presence of identity and cut:

$$\frac{}{\varphi \vdash \varphi \,\mathrm{tonk}\, \psi} \qquad \frac{}{\varphi \,\mathrm{tonk}\, \psi \vdash \psi} \tag{21}$$

We can see tonk as a functor from $L \times L$ to L. Now, define a "truth diagonal" functor $\Delta_\top : L \to L \times L$ by

$$\Delta_\top(\varphi) := (\top, \top) \tag{22}$$

and also define a "falsity diagonal" functor $\Delta_\perp : L \to L \times L$ by

$$\Delta_\perp(\varphi) := (\perp, \perp). \tag{23}$$

We can then prove that $\Delta_\perp$ is a left adjoint of tonk, and that $\Delta_\top$ is a right adjoint of tonk. In other words, tonk is a right adjoint of $\Delta_\perp$ and a left adjoint of $\Delta_\top$; therefore, tonk is an adjoint functor in two senses (if L is already endowed with tonk).

At the same time, however, this does not mean that the principle of categorical harmony cannot exclude tonk, a pathological connective we ought not to have in a logical system. Indeed, it is a problem in the other way around: in order to define tonk in a logical system, the principle of categorical harmony requires us to add it as a right or left adjoint of some functor, or equivalently, via an adjunction-induced bi-directional rule. Thus, when one attempts to define tonk in a logical system L according to categorical harmony, the task is the following:

1. Specify a functor $F : L \to L \times L$ that has a (right or left) adjoint.
2. Prove that tonk is a (left or right) adjoint of F, or that the rules for tonk are derivable in the system L extended with the bi-directional rule that corresponds to the adjunction.

As a matter of fact, however, this turns out to be impossible.

Let us give a brief proof. Suppose for contradiction that it is possible. Then we have a functor $F : L \to L \times L$, and its right or left adjoint is tonk. Assume that tonk is a left adjoint of F, which means that F is right adjoint to tonk. It then follows that F must be truth diagonal $\Delta_\top$ as defined above. The bi-directional rule that corresponds to the adjunction tonk $\dashv F$ is actually equivalent to the following (by the property of $\Delta_\top$):

$$\frac{}{\varphi_1 \,\mathrm{tonk}\, \varphi_2 \vdash_L \psi} \tag{24}$$

But this condition is not sufficient to make the rules for tonk derivable, thus the right adjoint of F cannot be tonk, and hence a contradiction. Next, assume that tonk is a right adjoint of F, i.e., F is a left adjoint of tonk. Then, F must be falsity diagonal $\Delta_\perp$, and the rule of the adjunction $F \dashv$ tonk is equivalent to the following:

$$\overline{\varphi \vdash_L \psi_1 \text{ tonk } \psi_2} \tag{25}$$

This is not enough to derive the rules for tonk, and hence a contradiction. This completes the proof.

It has thus been shown that:

- Tonk cannot be defined as an adjoint functor (of some functor) in a logical system without tonk, even though tonk is an adjoint functor in a logical system that is already equipped with tonk.
  - This is a subtle phenomenon, and we have to be careful of what exactly the question "Is tonk an adjoint functor?" means. Due to this, naïvely formulating categorical harmony as "logical constants = adjoint functors" does not work.
- Consequently, tonk cannot be introduced in any way according to the principle of categorical harmony.

We may then conclude that tonk is a pseudo-logical constant, and the rules for tonk are not meaning-conferring, not because it is non-conservative (i.e., Belnap's harmony fails for tonk), but because it violates the principle of categorical harmony (which is able to allow for non-conservativity as discussed above). Still, it is immediate to see the following:

- Tonk can actually be defined as being right adjoint to falsity diagonal $\Delta_\perp$, and left adjoint to truth diagonal $\Delta_\top$ at once. We may say that tonk is a "doubly adjoint" functor.
- In categorical harmony, therefore, it is essential to allow for a single adjunction only rather than multiple adjunctions, which are harmful in certain cases.

We again emphasise that tonk cannot be defined in a system without tonk by a single adjunction (i.e., there is no functor F such that an adjoint of F is tonk); nevertheless tonk can be defined by two adjunctions: $\Delta_\perp \dashv \text{tonk} \dashv \Delta_\top$, i.e., $\Delta_\perp$ is left adjoint to tonk, and tonk is left adjoint to $\Delta_\top$. Note that double adjointness itself is not necessarily paradoxical.

What is then the conceptual meaning of all this? After all, what is wrong with tonk? The right adjoint t of falsity diagonal $\Delta_\perp$ may be called the binary truth constant (the ordinary truth constant $\top$ is nullary), because the double-line rule of this adjunction boils down to $\varphi \vdash_L \psi_1 \,\mathrm{t}\, \psi_2$, which means that $\psi_1 \,\mathrm{t}\, \psi_2$ is implied by any formula φ (for any ψ_1, ψ_2). Likewise, the left adjoint s of truth diagonal $\Delta_\top$ may be called the "binary falsity constant", because the double-line rule of this adjunction boils down to $\psi_1 \,\mathrm{s}\, \psi_2 \vdash_L \varphi$, which means that $\psi_1 \,\mathrm{s}\, \psi_2$ implies any φ. Now, the rôle of tonk is to make the two (binary) truth and falsity constants (t and s) collapse into the same one constant, thus leading the logical system to inconsistency (or triviality); obviously, truth and falsity cannot be the same. This confusion of truth and falsity is the problem of tonk.

To put it differently, a right adjoint of $\Delta_\perp$ and a left adjoint of $\Delta_\top$ must be different, nevertheless tonk requires the two adjoints to be the same; the one functor that are

the two adjoints at once is tonk. The problem of tonk, therefore, lies in confusing two essentially different adjoints as if they represented the same one logical constant. We may thus conclude as follows:

- The problem of tonk is the problem of equivocation. The binary truth constant and the binary falsity constant are clearly different logical constants, yet tonk mixes them up, to be absurd.

This confusion of essentially different adjoints is at the root of the paradoxicality of tonk. There is no problem at all if we add to a logical system the right adjoint of $\Delta_\perp$ and the left adjoint of $\Delta_\top$ separately, any of which is completely harmless. Unpleasant phenomena only emerge if we add the two adjoints as just a single connective, that is, we make the fallacy of equivocation.

Let us think of a slightly different sort of equivocation. As explained above, $\wedge$ is right adjoint to diagonal Δ, and $\vee$ is left adjoint to it. What if we confuse these two adjoints? By way of experiment, let us define "disconjunction" as the functor that is right adjoint to diagonal, and left adjoint to it at the same time. Of course, a logical system with disconjunction leads to inconsistency (or triviality). Needless to say, the problem of disconjunction is the problem of equivocation: conjunction and disjunction are different, yet disconjunction mixes them up.

Then, is the problem of disconjunction precisely the same as the problem of tonk? This would be extensionally true, yet intensionally false. It is true in the sense that both pseudo-logical constants fall into the fallacy of equivocation. Nonetheless, it is false in the sense that the double adjointness condition of disconjunction is stronger than the double adjointness condition of tonk.

What precisely makes the difference between tonk and disconjunction? Tonk is a right adjoint of one functor, and at the same time a left adjoint of another functor. In contrast to this, disconjunction is a right and left adjoint of just a single functor. Disconjunction is, so to say, a uniformly doubly adjoint functor, as opposed to the fact that tonk is merely a doubly adjoint functor. The difference between tonk and disconjunction thus lies in uniformity. Hence:

- On the ground that uniform double adjointness is in general stronger than double adjointness, we could say that disconjunction is more paradoxical than tonk, endorsing a stronger sort of equivocation.
- We thereby lead to the concept of intensional degrees of paradoxicality of logical constants. Degrees concerned here are degrees of uniformity of double adjointness or equivocation.

What is then the strongest degree of paradoxicality in terms of adjointness? It is self-adjointness, and it is at the source of Russell-type paradoxical constants. A self-adjoint functor is a functor that is right and left adjoint to itself. This is the strongest form of double adjointness. Now, let us think of a nullary paradoxical connective R defined by the following double-line rule (this sort of paradoxical connectives has been discussed in Schroeder-Heister [20, 22]):

$$\frac{\vdash \neg R}{\vdash R}$$

Reformulating this, we obtain the following:

$$\frac{R \vdash}{\vdash R}$$

We may consider R as a unary constant connective $\tilde{R} : L \to L$ defined by $\tilde{R}(\varphi) = R$. Then, the double-line rule above shows that R is right and left adjoint to R, and therefore the Russell-type paradoxical constant R is a self-adjoint functor.

In order to express double adjointness, we need two functors (i.e., $\Delta_\perp$ and $\Delta_\top$) in the case of tonk, one functor (i.e., Δ) in the case of disconjunction, and no functor at all in the case of paradox R. These exhibit differences in the uniformity of double adjointness. Tonk exemplifies the most general case of double adjointness and exhibits the lowest degree of uniformity. On the other hand, paradox instantiates the strongest double adjointness, and exhibits the highest degree of uniformity. Disconjunction exemplifies the only possibility in between the two.

We have thus led to three intensional degrees of paradoxicality (double adjointness $<$ uniform double adjointness $<$ self-adjointness):

	Right adjoint to	Left adjoint to
Genuine paradox R	Itself R	Itself R
Disconjunction	Diagonal Δ	Diagonal Δ
Tonk	Truth diagonal $\Delta_\top$	Falsity diagonal $\Delta_\perp$

The last two are caused by equivocation according to the categorical account of logical constants. In contrast, paradox R is not so for the reason that self-adjointness can be given by a single adjunction: if a functor is right (resp. left) adjoint to itself, it is left (resp. right) adjoint to itself. This is the reason why we call it "genuine paradox" in the table above. More conceptually speaking:

- Pseudo-paradoxes due to equivocation can be resolved by giving different names to right and left adjoints, respectively, which are indeed different logical constants, and it is natural to do so.
  - The paradoxicality of such pseudo-paradoxes is just in mixing up actually different logical constants which are harmless on their own.
- On the other hand, we cannot resolve genuine paradox in such a way: there are no multiple meanings hidden in the Russell-type paradoxical constant, and there is nothing to be decomposed in genuine paradox.
  - Genuine paradox is a truly single constant, and the paradoxicality of genuine paradox is not caused by equivocation, unlike tonk or disconjunction.

If we admit any sort of adjoint functors as logical constants, then we cannot really ban genuine paradox, which is surely an adjoint functor. A naïve formulation of Lawvere's idea of logical constants as adjoint functors, like "logical constants = adjoint functors", does not work here again (recall that we encountered another case of this in the analysis of tonk). This is the reason why we have adopted the iterative conception of logic in our formulation of categorical harmony. In that view, logical constants must be constructed step by step, from old to new ones, via adjunctions. Since genuine paradox emerges via self-adjointness, however, there is no "old" operation that is able to give rise to genuine paradox via adjunction. In this way, categorical harmony based upon the iterative conception of logic allows us to avoid genuine paradox.

5 Concluding Remarks: From Semantic Dualism to Duality

Let us finally address further potential implications of categorical logic to the theory of meaning. The dualism between the referentialist and inferentialist conceptions of meaning may be called the semantic dualism. Categorical logic may (hopefully) yield a new insight into the semantic dualism, as argued in the following.

From a categorical point of view, "duality" may be discriminated from "dualism". Dualism is a sort of dichotomy between two concepts. Duality goes beyond dualism, showing that the two concepts involved are actually two sides of the same coin, just as two categories turn out to be equivalent by taking the mirror image of each other in the theory of categorical dualities. Duality in this general sense seems to witness universal features of category theory. Indeed, the classic dualism between geometry and algebra breaks down in category theory. For example, the categorical concept of algebras of monads encompasses topological spaces in addition to algebraic structures. Category theory may be algebraic at first sight (indeed, categories are many-sorted algebras), yet it is now used to formulate geometric concepts in broad fields of geometry, ranging from algebraic and arithmetic geometry to knot theory and low-dimensional topology. It is also a vital method in representation theory and mathematical physics. Technically, there are a great number of categorical dualities between algebraic and geometric structures (e.g., the Gelfand duality and the Stone duality). It may thus be said that the concept of categories somehow captures both algebraic and geometric facets of mathematics at a deeper level, and so there is duality, rather than dualism, between algebra and geometry.

Just as in this sense category theory questions the dualism between algebra and geometry, categorical logic opaques the generally received, orthodox distinction between model theory and proof theory, and presumably even the semantic dualism above, suggesting that they are merely instances of the one concept of categorical logic. For example, the Tarski semantics and the Kripke semantics, which are two major instances of set-theoretic semantics, amount to interpreting logic in the

category of sets and the category of (pre)sheaves, respectively (from a fibrational, or in Lawvere's terms hyperdoctrinal, point of view, we conceive of topos-induced subobject fibrations rather than toposes themselves). On the other hand, proof systems or type theories give rise to what are called syntactic categories, and their proof-theoretic properties are encapsulated in those syntactic categories. For example, cartesian biclosed categories and $*$-autonomous categories give fully complete semantics of intuitionistic logic and classical linear logic, respectively, in the sense that the identity of proofs exactly corresponds to the identity of morphisms (note also that the possibility of proof normalisation is implicitly built-in to categorical semantics; if normalisation is not well behaved, syntactic categories are not well defined). There is thus no dualism between model-theoretic and proof-theoretic semantics in categorical semantics. That is, there is just the one concept of categorical semantics that can transform into either of the two semantics by choosing a suitable category (fibration, hyperdoctrine) for interpretation. Put another way, we can make a proof system out of a given structured category (which is called the internal logic of the category; some conditions are of course required to guarantee desirable properties of the proof system), and at the same time, we can also model-theoretically interpret logic in that category. This feature of categorical logic allows us to incorporate both model-theoretic and proof-theoretic aspects of logic into the one concept. In a nutshell, categorical semantics has both proof-theoretic and model-theoretic semantics inherent in it, and from this perspective, there is no dualism, but duality between proof-theoretic and model-theoretic semantics, which may be called the semantic duality.

We must, however, be careful of whether this sort of unification makes sense philosophically as well as mathematically. There may indeed be some conceptual reasons for arguing that we ought to keep model-theoretic and proof-theoretic semantics separate as usual. Yet we may at least say that categorical logic exposes some common features of the two ways of accounting for the meaning of logical constants; at a level of abstraction, model-theoretic and proof-theoretic semantics become united as particular instances of the one categorical semantics. The philosophical significance of that level of abstraction is yet to be elucidated.

Acknowledgments I would like to thank Peter Schroeder-Heister for fruitful discussions in Oxford in spring 2012, which, *inter alia*, led to the table of degrees of paradoxicality of logical constants. I am also grateful to the two reviewers for their detailed comments and suggestions for improvement. And last, I hereby acknowledge generous financial support from the following institutions: the National Institute of Informatics, the Nakajima Foundation, Wolfson College, and the University of Oxford.

References

1. Abramsky, S.: Temperley-Lieb algebra: from knot theory to logic and computation via quantum mechanics. In: Chen, G., Kauffman, L., Lomonaco, S. (eds.) Mathematics of Quantum Computing and Technology, pp. 515–558. Taylor and Francis (2007)
2. Awodey, S.: Category Theory. Oxford University Press, Oxford (2006)
3. Belnap, N.: Tonk, plonk and plink. Analysis **22**, 130–134 (1962)
4. Bonnay, D., Simmenauer, B.: Tonk strikes back. Australas. J. Log. **3**, 33–44 (2005)
5. Brandom, R.: Articulating Reasons: An Introduction to Inferentialism. Harvard University Press, Cambridge (2000)
6. Cook, R.T.: What's wrong with tonk(?). J. Philos. Log. **34**, 217–226 (2005)
7. Davidson, D.: Inquiries into Truth and Interpretation. Clarendon Press, New York (2001)
8. Došen, K.: Logical constants: an essay in proof theory. Ph.D. thesis, Oxford University (1980)
9. Došen, K.: Logical constants as punctuation marks. Notre Dame J. Form. Log. **30**, 362–381 (1989)
10. Dummett, M.: The Logical Basis of Metaphysics. Harvard University Press, Cambridge (1991)
11. Hilbert, D.: On the infinite. In: van Heijenoort, J. (ed.) From Frege to Gódel: A Source Book in Mathematical Logic, 1897–1931, pp. 369–392. Harvard University Press, Cambridge (1967)
12. Lambek, J., Scott, P.J.: Introduction to Higher-Order Categorical Logic. Cambridge University Press, Cambridge (1986)
13. Lawvere, F.W.: Adjointness in foundations. Dialectica **23**, 281–296 (1969)
14. Martin-Löf, P.: On the meanings of the logical constants and the justifications of the logical laws. Nord. J. Philos. Log. **1**, 11–60 (1996)
15. Maruyama, Y.: Full lambek hyperdoctrine: categorical semantics for first-order substructural logics. In: Kohlenbach, U., Libkin, L., de Queiroz, R. (eds.) Logic, Language, Information, and Computation. LNCS, vol. 8071, pp. 211–225. Springer, Berlin (2013)
16. Pitts, A.: Categorical logic. In: Abramsky, S., Gabbay, D.M., Maibaum, T.S.E. (eds.) Handbook of Logic in Computer Science, Chap. 2. Oxford University Press, Oxford (2000)
17. Prawitz, D.: Natural Deduction. Almqvist & Wiksell, Stockholm (1965)
18. Prior, A.N.: The runabout inference-ticket. Analysis **21**, 38–39 (1960)
19. Sambin, G., Battilotti, G., Faggian, C.: Basic logic: reflection, symmetry, visibility. J. Symb. Log. **65**, 979–1013 (2000)
20. Schroeder-Heister, P.: Definitional reflection and paradoxes, supplement to [21]. In: Stanford Encyclopedia of Philosophy (2012)
21. Schroeder-Heister, P.: Proof-theoretic semantics. In: Stanford Encyclopedia of Philosophy (2012)
22. Schroeder-Heister, P.: Proof-theoretic semantics, self-contradiction, and the format of deductive reasoning. Topoi **31**, 77–85 (2012)
23. Schroeder-Heister, P.: The categorical and the hypothetical: a critique of some fundamental assumptions of standard semantics. Synthese **187**, 925–942 (2012)
24. Stairs, A.: Could logic be empirical? The Putnam-Kripke debate. In: Chubb, J., Eskandarian, A., Harizanov, V. (eds.) Logic and Algebraic Structures in Quantum Computing and Information. Cambridge University Press, Cambridge (2015). Forthcoming
25. Westerståhl, D.: From constants to consequence, and back. Synthese **187**, 957–971 (2012)

The Paradox of Knowability from an Intuitionistic Standpoint

Gabriele Usberti

Abstract An intuitionistic solution to the Paradox of Knowability is given. It consists (i) in accepting $\alpha \to \mathrm{K}\,\alpha$, the ordinary formalization of the principle of Radical Anti-Realism (RAR) that "Every truth is known", since, intuitionistically understood, it means that proofs are epistemically transparent; and (ii) in accepting (RAR) itself, on the basis of the fact that knowledge is an intuitionistic internal truth notion. Some neo-verificationist approaches are criticized. Finally the problem of how to frame a rational discussion between Classicism and Intuitionism is briefly discussed.

Keywords Intuitionism · Knowability paradox · Anti-realistic theory of meaning · Truth notions · BHK-interpretation

1 Introduction

The Paradox of Knowability[1] is an argument that from the principle of Knowability

(K) Every truth is knowable,

accepted by anti-realists of all sorts, derives the principle of Radical Anti-Realism

(RAR) Every truth is known,

which, on the contrary, virtually no anti-realist would be prepared to accept. Contraposing, if one does not want to accept (RAR), one must reject (K) as well.[2]

[1] The paradox is usually ascribed to F. Fitch, but is due in fact to A. Church. For its history see [18].

[2] Of course, an intuitionist might also accept the negation of (K) without accepting that there are unknowable truths.

G. Usberti (✉)
Università degli Studi di Siena, DISPOC, Via Roma 56, 53100 Siena, Italy
e-mail: gabriele.usberti@unisi.it

T. Piecha and P. Schroeder-Heister (eds.), *Advances in Proof-Theoretic Semantics*,
Trends in Logic 43, DOI 10.1007/978-3-319-22686-6_7

In a nutshell, the argument has the following structure. I shall use the symbols $\&$, $+$, $\supset$, $=$, $-$, Π, Σ for the classical logical constants; and $\wedge$, $\vee$, $\rightarrow$, $\leftrightarrow$, $\neg$, $\forall$, $\exists$ for the intuitionistic ones. First, (K) and (RAR) are formalized by the two following schemas, respectively:[3]

(1) $$\alpha \supset \Diamond \mathrm{K}\,\alpha$$

(2) $$\alpha \supset \mathrm{K}\,\alpha.$$

Then replace α in (1) with the proposition "$q \;\&\; -\mathrm{K}\,q$"; you obtain the following instance:

(3) $$(q \;\&\; -\mathrm{K}\,q) \supset \Diamond \mathrm{K}(q \;\&\; -\mathrm{K}\,q),$$

from which it is not difficult to derive, by means of intuitively acceptable principles, the unacceptable (2). The principles are the following:

(a) $$\Box(\mathrm{K}(\alpha \;\&\; \beta) \supset (\mathrm{K}\,\alpha \;\&\; \mathrm{K}\,\beta)$$

(b) $$\Box(\mathrm{K}\,\alpha \supset \alpha)$$

(c) $$\Box((\alpha \;\&\; -\alpha) \supset \bot)$$

(d) $$-(\alpha \;\&\; -\beta) \supset (\alpha \supset \beta),$$

and this is the derivation:[4]

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{[q \;\&\; -\mathrm{K}\,q]^1}{\Diamond \mathrm{K}(q \;\&\; -\mathrm{K}\,q)}\ \text{by (3)}}{\Diamond(\mathrm{K}\,q \;\&\; \mathrm{K} - \mathrm{K}\,q)}\ \text{by (a)}}{\Diamond(\mathrm{K}\,q \;\&\; -\mathrm{K}\,q)}\ \text{by (b)}}{\Diamond(\bot)}\ \text{by (c)} \qquad -\Diamond(\bot)}{\bot}\ \supset\mathrm{E}}{-(q \;\&\; -\mathrm{K}\,q)}\ -\mathrm{I},\,1}{q \supset \mathrm{K}\,q}\ \text{by (d)}$$

The paradox is usually viewed as an argument against anti-realism, when this is conceived, according to a famous proposal by Michael Dummett, as a doctrine concerning meaning rather than ontology. Dummett writes:

> Realism I characterise as the belief that statements of [a certain] class possess an objective truth-value, independently of our means of knowing it: they are true or false in virtue of a reality existing independently of us. The anti-realist opposes to this the view that the statements of [that] class are to be understood only by reference to the sort of thing which we count as evidence for a statement of that class.[5]

[3] "K" should be read as "It is known that".

[4] Observe that this last step, and principle (d), are not intuitionistically valid.

[5] Reference [2], p. 146.

The fundamental opposition between realism and anti-realism concerns therefore, according to Dummett, the key notion of the theory of meaning, i.e. the notion in terms of which the meaning of the statements of the given class is to be explained: truth according to the realist, evidence according to the anti-realist. However, since Dummett holds that meaning is to be explained in any case in terms of truth-conditions, and that for the anti-realist truth can consist only in the existence of evidence, the realism/anti-realism opposition, in the final version he offers, concerns the notion of truth to be used in explaining meaning: the bivalent notion according to the realist, some non-bivalent notion according to the anti-realist.[6] The criterion of realism is therefore, in Dummett's opinion, the acceptance/refusal of the bivalence principle concerning truth.

Notice that this raises immediately a question: if we are ready to oppose the realistic, bivalent, notion of truth with other, non-bivalent, notions, we must ask at which conditions a notion can be considered as a notion of *truth*. I shall return to this question later. For the time being let me observe that the story of the semantic characterization of the realism/anti-realism debate is not finished. After Dummett it has been observed that all the anti-realistic notions of truth on the market (truth as assertibility, truth as existence of a verification, and so on) share a general characteristic: that truth is an epistemic notion, and therefore is essentially *knowable*. Knowability has therefore been identified as the essential property of truth, and anti-realism has been characterized as the view that every truth is knowable. It is precisely this feature of anti-realistic truth that the paradox is intended to hit.

In this paper I shall first argue that an intuitionistic solution to the paradox is available; then I shall examine the position of neo-verificationism concerning the paradox; finally I shall briefly consider the general question of how a rational discussion of alternative logics is possible at all.

2 An Intuitionistic Solution

The first step towards an intuitionistic solution is the remark that the formalizations of (K) and (RAR) by (1) and (2), respectively, acquire a meaning very different from the intuitive one if the logical constants occurring in them are understood according to the BHK-explanation, i.e. the explanation of their intuitionistic meaning given by Brouwer, Heyting and Kolmogorov, and that the intuitionistic formulas corresponding to (1) and (2), namely

(1′) $\alpha \to \Diamond \mathrm{K}\, \alpha$

(2′) $\alpha \to \mathrm{K}\, \alpha,$

[6] As it will become evident below, the argument stated in this paper in no way relies on Dummett's opinion (for which see for instance [4], p. XXII) that also an anti-realist should explain meaning in terms of truth-conditions.

are valid, independently from there being a Church-Fitch argument (whose last step is not intuitionistically valid).

2.1 (2′) is intuitionistically valid

Intuitionists do not agree with Dummett and other neo-verificationists that meaning is to be explained in any case in terms of truth-conditions. According to them, «The notion of truth makes no sense [...] in intuitionistic mathematics»[7]; the key notion of the theory of meaning is the notion of *proof*, and understanding α (knowing its meaning) is to be explained as being capable to recognize the proofs of α. The content of a mathematical statement α (what Frege would have called the thought expressed by α) is characterized by Heyting as the expectation of a proof of α. What a proof of α is, is explained by recursion on the logical complexity of α, under the assumption that we have an intuitive understanding of what is a proof of an atomic statement. This is the BHK-explanation. I think a revision of this explanation is necessary concerning disjunction and existential quantification.[8] According to Heyting, a proof of $\alpha \vee \beta$ is either a proof of α or a proof of β; however, even the intuitionists consider, for instance, "$\mathrm{Prime}(n) \vee \neg\,\mathrm{Prime}(n)$", where n is some very large number, as assertible even if neither "$\mathrm{Prime}(n)$" nor "$\neg\,\mathrm{Prime}(n)$" is; I propose therefore to revise the BHK-explanation in the following way: A proof of $\alpha \vee \beta$ is a procedure such that its execution yields,[9] after a finite time, either a proof of α or a proof of β.[10] An analogous modification of the clause for $\exists x\alpha$ can be similarly motivated.

Summing up, the revised version of the BHK-explanation I will make reference to is the following:

[7]Reference [12], p. 279. It should be stressed that Heyting is speaking of the realist, platonic, notion of truth. His assertion cannot therefore be understood as excluding that, within an intuitionistic framework, it is possible to define some notion that, on the one hand, can plausibly be proposed as a notion of truth, and, on the other hand, is reducible to others already present within that framework. I shall come back to this point in Sect. 2.2.

[8]For a more detailed motivation of this revision see [22], p. 42. The possibility of such a revision is explicitly envisaged, but discarded, in [3], p. 20.

[9]"Yields" is to be understood as equivalent to "is known to yield".

[10]For instance, a proof of "$\mathrm{Prime}(n) \vee \neg\,\mathrm{Prime}(n)$" is a primality test for n, i.e. an algorithm for determining whether n is prime. Such a test should not be confused with a general method consisting in applying to every number x a test for the primality of x (this general method is a proof of "$\forall x(\mathrm{Prime}(x) \vee \neg\,\mathrm{Prime}(x))$").

Definition 1

A proof of	*is*
$\alpha \wedge \beta$	a pair $\langle \pi_1, \pi_2 \rangle$, where π_1 is a proof of α and π_2 is a proof of β
$\alpha \vee \beta$	a procedure p whose execution yields, after a finite time, either a proof of α or a proof of β
$\alpha \rightarrow \beta$	a constructive function[a] f such that, for every proof π of α, $f(\pi)$ is a proof of β
$\neg\alpha$	a constructive function f such that, for every proof π of α, $f(\pi)$ is contradiction
$\forall x \alpha^{\dagger}$	a constructive function f such that, for every $c \in D$, $f(c)$ is a proof of $\alpha(\underline{c})^{\ddagger}$
$\exists x \alpha^{\dagger}$	a procedure p whose execution yields, after a finite time, a pair $\langle c, \pi \rangle$, where $c \in D$ and π is a proof of $\alpha(\underline{c})^{\ddagger}$

† where x varies on D
‡ $\underline{c}$ is a name of c
[a] Heyting speaks of "general method"; I shall use throughout "constructive function" (or briefly "function") with the same meaning.

In order to arrive at defining the meaning of (2′), we must ask how to define a proof of $\mathrm{K}\,\alpha$. There is no official intuitionistic answer, and there are several possibilities, according to the intended intuitive reading of K: "α is presently known by someone", "α is known by someone at some time", "α is presently known by the one who is considering α", and so on. I shall choose the last reading because, on the one hand, it seems to be the most congenial to intuitionistic ideas and, on the other hand, it is equivalent to the other readings for my present purposes. Here is my proposal:

Definition 2 Whenever one is presented with a proof of α, a proof of $\mathrm{K}\,\alpha$ is the observation that what one is presented with is a proof of α.

In the light of Definitions 1 and 2, (2′) expresses the expectation, for any proposition α, of a function f associating to every proof π of α the observation that π is a proof of α.

Now, a fundamental characteristic of proofs, as the intuitionists conceive them, is that «for them, *esse est concipi*», to quote Dummett's illuminating formulation.[11] In other words, a proof of α is essentially what is recognized as such by an idealized knowing subject: there is no point of view from which something can be judged *to be* a proof of α in spite of the fact that an idealized subject who is presented with it does not judge it as a proof of α, or from which something can be judged not to be a proof of α in spite of the fact that an idealized subject who is presented with it does judge it as a proof of α. I shall call this characteristic of intuitionistic proofs their *epistemic transparency*; it can be expressed in the following way:

(4) A proof of α is *epistemically transparent* if and only if a subject who is presented with it is in a position to know that it is a proof of α.

[11] «[M]athematical objects […] are mental constructions […] in the sense that, for them, *esse est concipi*.» [3, p. 7].

A little[12] reflection shows that, if proofs are transparent, the function f whose expectation is expressed by (2′) does exist: by (4), for any proof π of α, if one is presented with π, one is in a position to know that π is a proof of α; so one can associate to π the observation that π is a proof of α; and this, by Definition 2, is the proof of $\mathrm{K}\,\alpha$ required by the definition of f. Conversely, if a subject s knows, i.e. can compute, a function f associating to every proof π of α the observation that π is a proof of α, then s, when presented with a proof of α, is in a position to know that it is a proof of α. In conclusion (2′), far from saying that every intuitive truth is known, says that proofs are epistemically transparent, and is therefore obviously true; moreover, it remains true if K is read as "is known by someone at some time", since, if α is presently known by me, then α is known by someone at some time.

Williamson [25], pp. 430–1, raises the following objection to the validity of (2′). He first argues that a proof of $\alpha \to \beta$ should be conceived by intuitionists as a function f from proof-tokens to proof-tokens «that is *unitype* in the sense that if p and q are proof-tokens of the same type then so are $f(p)$ and $f(q)$.» Then, under the assumption that

> a proof of $\alpha \to \mathrm{K}\,\alpha$ is a unitype function that evidently takes any proof-token of α to a proof-token, for some time t, of the proposition that α is proved at t,

he shows that, if α has not yet been decided, the function f that associates to every proof-token of α a proof-token of the proposition that α is proved at t is not unitype: if the proof-token p is carried out at t_1 and the proof-token q is carried out at t_2, where $t_1 \neq t_2$, then $f(p) \neq f(q)$, since the proposition that α is proved at t_1 is different from the proposition that α is proved at t_2. However, the quoted assumption is by no means conceptually necessary, nor is it a consequence of the general conception of proofs of conditionals as unitype functions. If we assume that a proof of $\alpha \to \mathrm{K}\,\alpha$ is a unitype function that takes any proof-token of α to a proof-token of the proposition that α is proved (with no mention of the time at which it is proved), f is unitype.

Let us now consider (1′). It is more difficult to suggest an intuitionistic reading for it, since it is not easy to devise a clear intuitionistic sense for the possibility operator. Here is one plausible candidate:

Definition 3 A proof of $\Diamond\alpha$ is a procedure such that its execution yields,[13] after a finite time, a proof of α.

According to this explanation, (1′) expresses the expectation, for any proposition α, of a function g associating, to every proof π of α, a procedure p whose execution yields, after a finite time, the observation that what one is presented with is a proof of α. Now, if we remember that the function f whose expectation is expressed by (2′) does exist, and that f associates, to every proof π of α, directly the empirical observation that π is a proof of α, we see that also the function g exists: the procedure

[12] «To be in a position to know p, it is neither necessary to know p nor sufficient to be physically and psychologically capable of knowing p. No obstacle must block one's path to knowing p. If one is in a position to know p, and one has done what one is in a position to do to decide whether p is true, then one does know p. […] Thus being in a position to know […] is factive.» [27, p. 95].

[13] "Yields" is to be understood as equivalent to "is known to yield".

consists precisely in effecting the empirical observation. In conclusion, also between the content of (1′) and the intuitive content of (K) there is a substantial difference.

2.2 Truth Notions

The second step towards a solution consists in looking for a plausible intuitive sense of (K) and (RAR), according to which not only (K), but also (RAR), becomes acceptable. Of course, there is a sense in which (RAR) is *not* acceptable; my question is whether there is a sense in which it is. A first component of such a sense has already been made explicit: it consists in giving the logical constants their (revised) BHK-sense. The second component is of course the concept of truth, which (K) and (RAR) explicitly refer to. It is at this point that a problem mentioned above becomes relevant: at which conditions is a notion a notion *of truth*?

First, let me explain why, exactly, the question is crucial. If we read a formula of the language of classical propositional logic (CPL), it is natural and correct to read an occurrence in it of a propositional letter, say p, as "p is true"; for example, the intuitive reading of an instance of the schema $\alpha + -\alpha$ would be, "Either p is true or $-p$ is true" (which, given the definition of "p is false" as "$-p$ is true", is equivalent to "Either p is true or p is false"). This is correct because the key notion of the realistic explanation of the meaning of the logical constants is the realistic (i.e. bivalent) notion of truth; but it is no longer legitimate when we consider a formula of the language of IPL, since the key notion of the BHK-explanation is not the (bivalent) notion of truth. As a consequence, the simple occurrence of p will not be sufficient to make reference to the truth of p: in order to make reference to the truth of p it will be necessary to use a truth-predicate, or a truth-operator. Notoriously, the choice between expressing truth with a predicate or an operator has an impact on many other things, in particular on the possibility of expressing semantic paradoxes; since the questions discussed in this paper are independent of such a possibility, I shall choose the simpler alternative of expressing truth with an operator. The question arises at this point: what makes an operator a *truth* operator?

A plausible answer to this question is offered by Tarski's Convention T, in the case truth is expressed by a predicate. Tarski has proposed to consider a definition of truth as materially adequate if it entails every sentence of the form

(5) N is true if and only if t,

where N is the name of a sentence of the object language, and t is a translation of that sentence into the metalanguage. Since "materially adequate" means faithful to our intuitions about the notion of truth, we can take the validity of (5) as a criterion for a formally defined predicate to be a truth-predicate, i.e. a predicate defining a notion we are intuitively prepared to consider a notion of truth.[14] From this we may

[14] If I understand it correctly, [15], p. 148, makes essentially the same point.

easily extract an analogous condition for an operator: an operator O can be seen as a truth-operator if it is defined in such a way that it entails every sentence of the form

(6) $\mathrm{O}\,\alpha$ if and only if t,

where α is a sentence of the object language and t is a translation of that sentence into the metalanguage. Finally, if we make the further simplifying assumption that the metalanguage is an extension of the object language, (6) is equivalent to

(7) $\mathrm{O}\,\alpha$ if and only if α,

which is the usual version of what I shall call "The (T) Schema".

So, my proposal is that an operator is to be considered as a truth operator if its meaning is defined in such a way as to satisfy the (T) Schema. Before going on, let me examine an objection to this proposal raised by Dummett. In *The Logical Basis of Metaphysics* he writes:

> It is sometimes alleged that what makes a given notion a notion of truth is that it satisfies all instances of the (T) schema. This is wrong [...]. If a constructivist proposes that the only intelligible notion of truth we can have for mathematical statements is that under which they are true just in case we presently possess a proof of them, he is offering a characterisation of truth for which the (T) schema fails, since truth, so understood, does not commute with negation.[15]

Let me try to make the argument explicit. Dummett is envisaging the case of a constructivist who equates the truth of a (mathematical) statement α with the actual possession of a proof of α. The intuitionist may be seen as a case in point, and in a moment I myself shall explicitly endorse this view. At this point Dummett, assuming that a consequence of the (T) schema is that the following principle is valid:

(8) $\mathrm{T}\,\neg\alpha$ if and only if $\neg\,\mathrm{T}\,\alpha$,

remarks that (8) is invalid when truth is equated to the actual possession of a proof (since from the fact that one does not possess a proof of $\mathrm{T}\,\alpha$ it does not follow that one possesses a proof of $\mathrm{T}\,\neg\alpha$), and concludes, by contraposing, that the (T) schema is not valid. Here is the derivation of (8) from the (T) schema:

(9)
(i) $\mathrm{T}\,\neg\alpha$ iff $\neg\alpha$ [from (7), replacing α *with* $\neg\,\alpha$]
(ii) $\neg\alpha$ iff $\neg\,\mathrm{T}\,\alpha$ [from (7), by contraposition]
(iii) $\mathrm{T}\,\neg\alpha$ iff $\neg\,\mathrm{T}\,\alpha$ [from (i) and (ii), by transitivity].

It seems to me that Dummett's remark that (8) is invalid is not correct: $\neg\,\mathrm{T}\,\alpha$ does not mean that one does not possess a proof of $\mathrm{T}\,\alpha$, but that one possesses a method to transform every proof of $\mathrm{T}\,\alpha$ into a contradiction. When one possesses such a

[15] Reference [5], p. 166.

method, one has a proof of $\neg\alpha$; owing to the epistemic transparency of intuitionistic proofs, one can effect the (empirical) observation that what one has is a proof of $\neg\alpha$, and this observation is a proof of $\mathrm{K}\,\neg\alpha$, i.e. of $\mathrm{T}\,\neg\alpha$ under the present identification of truth with the actual possession of a proof. In conclusion, when the truth of a (mathematical) statement α is equated with the actual possession of a proof of α, truth does commute with intuitionistic negation.[16]

2.3 Internal and Intuitive Truth

The next question to consider is whether the validity of the (T) Schema picks out a unique notion of truth. Tarski seems to hold that it does. In [21] he expresses the conviction that the material adequacy condition imposed onto the definition of truth, is capable to select the classical Aristotelian notion of truth as correspondence. The conviction is not explicitly stated, but it can be inferred from the following facts:
(i) In section I.3 Tarski expresses an intention:

> We should like our definition to do justice to the intuitions which adhere to the *classical Aristotelian conception of truth* […] we could perhaps express this conception by means of the familiar formula:
>
> *The truth of a sentence is its agreement with (or correspondence to) reality.*[17]

(ii) In section I.4 the same intention is made precise by requiring that the definition satisfies the material adequacy condition. Hence, the material adequacy condition 'does justice' to the intuitive notion of truth as correspondence. The question whether the intuitive notion of truth as correspondence is bivalent is not explicitly addressed by Tarski; an affirmative answer from him is suggested by the fact that in [20] he derives the principle of bivalence from the (materially adequate) definition of

[16]Notice that truth commutes with intuitionistic disjunction as well, since the principle

$$(*) \qquad \mathrm{K}(\alpha \vee \beta) \rightarrow (\mathrm{K}\,\alpha \vee \mathrm{K}\,\beta)$$

is valid. A proof of $(*)$ is a function f transforming every proof of $\mathrm{K}(\alpha \vee \beta)$ into a proof of $\mathrm{K}\,\alpha \vee \mathrm{K}\,\beta$; a proof π of $\mathrm{K}(\alpha \vee \beta)$ is the observation that what one is presented with is a procedure p_1 such that its execution yields, after a finite time, either a proof of α or a proof of β. Define the following procedure p_2:

- apply p_1;
- if the outcome is a proof of α, perform the observation that what one is presented with is a proof of α, getting a proof of $\mathrm{K}\,\alpha$;
- if the outcome is a proof of β, perform the observation that what one is presented with is a proof of β, getting a proof of $\mathrm{K}\,\beta$.

Of course p_2 is a procedure such that its execution yields, after a finite time, either a proof of $\mathrm{K}\,\alpha$ or a proof of $\mathrm{K}\,\beta$; we can therefore take p_2 as $f(\pi)$.

[17]Reference [21], pp. 342–3.

truth.[18] However, the derivation crucially uses Excluded Middle, as is made clear by the following steps:

(10)
(i) $\alpha + -\alpha$ [the law of Excluded Middle]
(ii) $\mathrm{T}\,\alpha \equiv \alpha$ [the (T) Schema]
(iii) $\mathrm{T}\,\alpha + \mathrm{T} -\alpha$ [from (i) and (ii) by Replacement].

So, Tarski's conviction is correct only under the premiss that the metalanguage is associated to a metatheory whose semantics validates Excluded Middle. If this principle is not valid in the metatheory, it is possible to exhibit counterexamples to Tarski's conviction, namely it is possible to define a non-bivalent notion of truth satisfying the (T) Schema. I shall now show how.[19]

Let us adopt a metatheory in which the logical constants are read according to the (revised) BHK-explanation (which, of course, does not validate the Excluded Middle). We are looking for a materially adequate definition, i.e. such that all the equivalences

(11) $$\mathrm{T}\,\alpha \leftrightarrow \alpha$$

are logical consequences of it (where T is the intended truth operator). The definition I suggest is the following[20]:

Definition 4 $\mathrm{T}\,\alpha =_{\mathrm{def.}} \mathrm{K}\,\alpha$,

where the meaning of K is defined by Definition 2. Definition 4 is materially adequate: (2′) is valid for the reasons explained above; and the converse

(12) $$\mathrm{K}\,\alpha \rightarrow \alpha$$

is valid as well: it expresses the expectation of a function h associating, to every observation o that what one is presented with is a proof of α, a proof $h(o)$ of α, and h is warranted to exist by the factivity of proof observation.[21] In conclusion, the knowledge operator K is a truth operator, and of course this operator does not satisfy the principle of bivalence.

Concluding, it is true both (i) that the validity of the (T) schema expresses our essential intuition about the notion of truth, and (ii) that our most common intuitive notion of truth is realistic; but the reason why (ii) holds is bivalence, not the (T) schema: the validity of the (T) schema is neutral among different intuitive notions of truth. It is therefore possible, and necessary, to introduce a clear distinction between

[18] Reference [20], pp. 197–8. The principle of bivalence is called by Tarski "The principle of excluded middle".

[19] Another example is the notion of truth defined in [14].

[20] Remember that the definition is intended to apply to mathematical statements.

[21] The factivity is warranted by the assumption that proof observation can plausibly be conceived as a computational process.

the condition at which an operator is a truth operator and the condition at which an operator reflects our realistic intuitions about the notion of truth; the former consists in the validity of the (T) schema,[22] the latter may be epitomized into the slogan of truth as correspondence and consequently into the validity of the law of bivalence. A notion satisfying the former condition is capable to play (at least some of) *the roles* of the notion of truth; truth as correspondence constitutes our predominant common-sense notion of truth. Between these two extremes there is a variety of truth notions, of which knowability and existence of a verification are two instances. I shall call "internal" these theoretical notions of truth, to stress the fact that each of them is capable to play the, or at least some of the, conceptual roles of the notion of truth within the framework of the related theory of meaning and of the formal semantics that adopts it. In this terminology we can say that bivalent truth is both the predominant intuitive notion of truth and the internal notion of classical logic; and that, besides it, there are several other internal notions of truth.

At this point it should be clear that an intuitive sense of the principle (RAR), according to which it becomes acceptable, does exist: for, if the logical constants are understood according to the BHK-explanation, and truth is understood according to Definition 4, then (RAR) is a tautology, saying that every known statement is known. The intuitionistic solution of the paradox consists therefore in accepting (RAR) as obvious when the logical constants are understood intuitionistically and truth as internal.

Is the idea of equating truth to knowledge consistent? There is an argument—called by [16] "The Standard Argument"—that purports to show that it is not.[23] It consists in the following derivation of $\exists\alpha(\alpha \wedge \neg \mathrm{K}\,\alpha)$ from the assumptions $p \vee \neg p$ and $\neg \mathrm{K}\,p \wedge \neg \mathrm{K}\,\neg p$:

(13)

$$\dfrac{\begin{array}{c}\mathscr{D}\\(p \wedge \neg \mathrm{K}\,p) \vee (\neg p \wedge \neg \mathrm{K}\,\neg p)\end{array} \qquad \dfrac{[p \wedge \neg \mathrm{K}\,p]^3}{\exists\alpha(\alpha \wedge \neg \mathrm{K}\,\alpha)} \qquad \dfrac{[\neg p \wedge \neg \mathrm{K}\,\neg p]^4}{\exists\alpha(\alpha \wedge \neg \mathrm{K}\,\alpha)}}{\exists\alpha(\alpha \wedge \neg \mathrm{K}\,\alpha)}\ 3, 4$$

where $\mathscr{D}$ is:

$$\dfrac{p \vee \neg p \qquad \dfrac{\dfrac{[p]^1 \qquad \dfrac{\neg \mathrm{K}\,p \wedge \neg \mathrm{K}\,\neg p}{\neg \mathrm{K}\,p}}{p \wedge \neg \mathrm{K}\,p}}{(p \wedge \neg \mathrm{K}\,p) \vee (\neg p \wedge \neg \mathrm{K}\,\neg p)} \qquad \dfrac{\dfrac{[\neg p]^2 \qquad \dfrac{\neg \mathrm{K}\,p \wedge \neg \mathrm{K}\,\neg p}{\neg \mathrm{K}\,\neg p}}{\neg p \wedge \neg \mathrm{K}\,\neg p}}{(p \wedge \neg \mathrm{K}\,p) \vee (\neg p \wedge \neg \mathrm{K}\,\neg p)}}{(p \wedge \neg \mathrm{K}\,p) \vee (\neg p \wedge \neg \mathrm{K}\,\neg p)}\ 1, 2$$

[22]The claim that an operator O is a truth operator iff it satisfies the schema (7) should not be confounded with the minimalist claim that (7) is a good definition of the meaning of O. The former claim is perfectly compatible with the idea, embraced above, that the validity of (7) is not the definition, but the material adequacy condition of the definition, of O.

[23]See [16], p. 275.

Now, if we observe that there are statements p that the intuitionist acknowledges as being decidable (i.e. such that $p \vee \neg p$ is assertible), and that, as a matter of fact, are unknown (i.e., such that $\neg \mathrm{K}\, p \wedge \neg \mathrm{K} \neg p$ is true),[24] we obtain that $\exists \alpha (\alpha \wedge \neg \mathrm{K}\, \alpha)$ is assertible.

My answer consists in observing that the argument is valid but unsound, since $\neg \mathrm{K}\, p \wedge \neg \mathrm{K} \neg p$ is intuitionistically inconsistent. Assume that $\neg \mathrm{K}\, p \wedge \neg \mathrm{K} \neg p$ is assertible, and reason in the following way:

$$
(14) \qquad
\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{[\neg \mathrm{K}\, p \wedge \neg \mathrm{K} \neg p]^1}{\neg \mathrm{K}\, p} \quad \dfrac{[p]^2}{\mathrm{K}\, p}}{\dfrac{\bot}{\neg p}\,2} \quad \dfrac{\dfrac{\dfrac{[\neg \mathrm{K}\, p \wedge \neg \mathrm{K} \neg p]^1}{\neg \mathrm{K} \neg p} \quad \dfrac{[\neg p]^3}{\mathrm{K} \neg p}}{\bot}}{\neg\neg p}\,3}{\bot}}{\neg(\neg \mathrm{K}\, p \wedge \neg \mathrm{K} \neg p)}\,1
$$

The formula $\exists \alpha (\alpha \wedge \neg \mathrm{K}\, \alpha)$ may therefore be false. In order to show that it is actually false, let us wonder whether there could be a proof of it, i.e. a procedure p whose execution yields, after a finite time, a pair $\langle c, \pi \rangle$, where c is a proposition and π is a proof of $c \wedge \neg \mathrm{K}\, c$. A proof of $c \wedge \neg \mathrm{K}\, c$ is a pair $\langle \pi_1, \pi_2 \rangle$, where π_1 is a proof of c and π_2 is a proof of $\neg \mathrm{K}\, c$; such a pair cannot exist, on pain of contradiction: being presented with π_1, one can effect the observation that what one is presented with is a proof of c, thereby obtaining a proof π_3 of $\mathrm{K}\, c$; coupling π_3 with π_2 we obtain a proof of $\mathrm{K}\, c \wedge \neg \mathrm{K}\, c$: a contradiction; p cannot therefore exist. In conclusion, the intuitionist cannot assert $\exists \alpha (\alpha \wedge \neg \mathrm{K}\, \alpha)$, and the idea of statements that, being unknown, are not yet true nor false is not inconsistent.

The intuitionistic inconsistency of $\neg \mathrm{K}\, p \wedge \neg \mathrm{K} \neg p$ may sound unacceptable from the intuitive standpoint, since it seems to conflict with the idea, which also an intuitionist should accept, that there are undecided, hence unknown, statements. Here it is important, again, to pay attention to the intuitionistic meaning of the logical constants, in particular of negation. The assertibility of $\neg(\neg \mathrm{K}\, \alpha \wedge \neg \mathrm{K} \neg \alpha)$ means that a method is known to transform every proof of $\neg \mathrm{K}\, \alpha \wedge \neg \mathrm{K} \neg \alpha$ into a contradiction, hence that a logical obstacle is known to the possibility that there is a *proof* of $\neg \mathrm{K}\, \alpha \wedge \neg \mathrm{K} \neg \alpha$; it does not exclude the *fact* that neither α nor $\neg \alpha$ are known. We will see in a moment whether the existence of such a fact can be acknowledged within the intuitionistic conceptual framework. Before, I want to comment upon the existence of a logical obstacle to the possibility that there is a *proof* of $\neg \mathrm{K}\, \alpha \wedge \neg \mathrm{K} \neg \alpha$. This is neither unacceptable nor unexpected if we keep present that the operator K is, in intuitionistic logic, a truth-operator; for it is a principle valid in general, i.e. for *every* internal notion of truth, that the formula expressing the proposition "p is neither true nor false" is inconsistent. Take for instance the formula $-\mathrm{T}\, \alpha \,\&\, -\mathrm{T} -\alpha$, expressing the same proposition within classical logic, and reason exactly in the same way as in (14), simply replacing $\neg$ with $-$, and $\wedge$ with &. The crucial step is the

[24] An example is "Prime(n)", where n is some very large number.

inference of $\mathrm{T}\,\alpha$ from α; in other terms, the inconsistency of the formula expressing the proposition "α is neither true nor false" depends on the validity of the principle $\alpha \to \mathrm{T}\,\alpha$ (together with propositional laws that are common to classical and intuitionistic logic). We have seen that the reason why that principle is intuitionistically valid is the assumption that proofs are epistemically transparent; of course this very assumption may be questioned,[25] but the issue of its truth or falsity is utterly different from the question whether there are intuitive truths that, as a matter of fact, are unknown.

2.4 Unknown Statements

I have said that the assertibility of $\neg(\neg\,\mathrm{K}\,\alpha \wedge \neg\,\mathrm{K}\,\neg\alpha)$ does not exclude the existence of the fact that neither α nor $\neg\alpha$ are known. Can the intuitionist *assert* the existence of such a fact? I think not, and in this section I shall try to motivate this opinion.

Let me observe first that "$\mathrm{K}\,\alpha$", in all its possible readings, clearly is an *empirical* statement, not a mathematical one. I have argued elsewhere that the negation of many empirical statements, and in particular of K-statements, cannot be plausibly equated to intuitionistic negation $\neg$, and I have proposed that it be equated to Nelson's strong negation $\sim$.[26] So, if we add $\sim$ to the language $\mathscr{L}_{\mathrm{IPL}^{\mathrm{K}}}$ of Intuitionistic Propositional Logic plus the operator K, and we assume for simplicity that all the empirical sentences of $\mathscr{L}_{\mathrm{IPL}^{\sim,\mathrm{K}}}$ have proofs,[27] we must add, to the Definition 2 of the notion of proof of $\mathrm{K}\,\alpha$, a definition of the notion of proof of $\sim\mathrm{K}\,\alpha$. Here is my proposal:

Definition 5 Whenever one is presented with something that is not a proof of α, a proof of $\sim\mathrm{K}\,\alpha$ is the observation that what one is presented with is not a proof of α.

It should be noticed that Dummett's remark—that intuitionistic truth, when it is equated with the actual possession of a proof, does not commute with negation—is certainly correct when it is understood as referring to strong negation. For example, the observation that what one is presented with is not a proof that it is raining is not the same thing as the observation that what one is presented with is a proof that it is not raining. As a consequence, Dummett's objection to the validity of the (T) schema as a criterion for being a truth operator seems to cause trouble in this case. However, in this case the argument (9) is no longer valid: the second step is an application of contraposition, but contraposition is not valid for strong negation. As a consequence, the fact that strong negation does not commute with truth does not entail the invalidity of the (T) schema. We can therefore conclude that, even when we add to intuitionism strong negation, the knowledge operator K is a truth operator.

[25] A discussion of this assumption is beyond the limits of this paper.

[26] Reference [24].

[27] In general empirical sentences have (non-conclusive) justifications. A definition of the notion of justification for the sentences of $\mathscr{L}_{\mathrm{IPL}^{\sim,\mathrm{K}}}$ presupposes a solution of Gettier problems. I have suggested such a definition in [23].

It seems to me that, if the existence of undecided statements can be expressed at all in an intuitionistic language, it should be expressible in $\mathscr{L}_{\mathrm{IPL}^{\sim,\mathrm{K}}}$. Take for instance g, Goldbach's Conjecture: for the statement

(15) Goldbach's conjecture is undecided (unknown)

the following formula seems to be a plausible formalization in $\mathscr{L}_{\mathrm{IPL}^{\sim,\mathrm{K}}}$:

$$\sim \mathrm{K}\, g \wedge \sim \mathrm{K} \neg g. \tag{16}$$

Williamson argues against this formalization:

> if $\sim$ is to count intuitionistically as any sort of negation at all, $\sim A$ should at least be inconsistent with A in the ordinary intuitionistic sense.[28]

In other words, the schema

$$\sim\alpha \rightarrow \neg\alpha; \tag{17}$$

should be valid; then, from (16) one could derive $\neg \mathrm{K}\, g \wedge \neg \mathrm{K} \neg g$, which, by (2) and (13), is equivalent to $\neg g \wedge \neg\neg g$: a contradiction. However, the assumption that (17) is valid *for all* α of $\mathscr{L}_{\mathrm{IPL}^{\sim,\mathrm{K}}}$ seems to be a sort of *petitio principii*, since, on the one hand, it almost amounts to assuming what one wants to conclude, i.e. that (16) is inconsistent, and, on the other hand, the motivation for it seems insufficient. Notice that (17) *is* valid for all α belonging to $\mathscr{L}_{\mathrm{IPL}^{\sim}}$,[29] so, according to Williamson's criterion, $\sim$ does count intuitionistically as a sort of negation; the possible invalidity of (17) when α contains occurrences of K can therefore be imputed to the interplay between the meanings of K and $\sim$. On the other hand, (17) is clearly invalid when α contains occurrences of K. Consider the instance

$$\sim \mathrm{K}\, \alpha \rightarrow \neg \mathrm{K}\, \alpha: \tag{18}$$

it asserts the existence of a function f associating to every proof of the antecedent a proof of the consequent. A proof of the antecedent is the observation that what one is presented with is not a proof of α; this observation is true in two cases: when one is presented with a proof of $\sim\alpha$, and when one is presented with some x that is neither a proof of $\sim\alpha$ nor a proof of α. In the second case f should associate to x a function f' associating to every proof of K p a contradiction; but f' cannot exist: as x is not a proof of $\sim\alpha$, the existence of a y that is a proof of α cannot be ruled out, and if one observes that y is a proof of α, that observation is a proof of $\mathrm{K}\,\alpha$.

However, if we look at the interplay between intuitionistic logical constants, strong negation and K from the standpoint of Kripke semantics, the assertibility of (18) seems to be out of the question. A *Kripke model* for $\mathscr{L}_{\mathrm{IPL}^{\sim}}$ is a quadruple

[28] Reference [26], p. 139.

[29] Reference [9].

$\mathscr{M} = \langle W, \geq, D, V \rangle$, where W is a non-empty set (of nodes), $\geq$ is a reflexive partial order on W, and V is a partial function from atomic formulas and nodes to $\{0, 1\}$ satisfying the following conditions:

- If $V(p, w) = 0$ and wRw', then $\mathrm{Val}(p, w') = 0$;
  if $V(p, w) = 1$ and wRw', then $\mathrm{Val}(p, w') = 1$ (stability).
- For every $w \in W$, $V(\bot, w) = 0$.
  For every $w \in W$, $V(\sim\bot, w) = 1$.

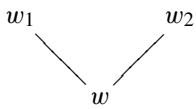
The notion $\vDash_w \alpha$ (α is true at w) is defined by induction on α as follows:

$$
\begin{aligned}
&\vDash_w p \text{ iff } V(p, w) = 1\\
&\vDash_w \sim p \text{ iff } V(p, w) = 0\\
&\vDash_w \sim\bot\\
&\vDash_w \alpha \wedge \beta \text{ iff } \vDash_w \alpha \text{ and } \vDash_w \beta\\
&\vDash_w \sim(\alpha \wedge \beta) \text{ iff } \vDash_w \sim\alpha \text{ or } \vDash_w \sim\beta\\
&\vDash_w \alpha \vee \beta \text{ iff } \vDash_w \alpha \text{ or } \vDash_w \beta\\
&\vDash_w \sim(\alpha \vee \beta) \text{ iff } \vDash_w \sim\alpha \text{ and } \vDash_w \sim\beta\\
&\vDash_w \alpha \to \beta \text{ iff, for every } w' \geq w, \text{ if } \vDash_{w'} \alpha \text{ then } \vDash_{w'} \beta\\
&\vDash_w \sim(\alpha \to \beta) \text{ iff } \vDash_w \alpha \text{ and } \vDash_w \sim\beta.
\end{aligned}
$$

Now, if we add the operator K to $\mathscr{L}_{\mathrm{IPL}^\sim}$, the only definition I can see that is faithful to Definition 5 is the following:

(19)
$$
\begin{aligned}
&\vDash_w \mathrm{K}\,\alpha \text{ iff } \vDash_w \alpha\\
&\vDash_w \sim\mathrm{K}\,\alpha \text{ iff, for some } w' \geq w,\ \vDash_{w'} \sim\alpha.
\end{aligned}
$$

Call any Kripke model for $\mathscr{L}_{\mathrm{IPL}^\sim}$ in which these clauses hold a model for $\mathscr{L}_{\mathrm{IPL}^{\sim,\mathrm{K}}}$. Consider now the following model $\mathscr{M}$ of $\mathscr{L}_{\mathrm{IPL}^{\sim,\mathrm{K}}}$, where $V(g, w)$ is undefined, $V(g, w_1) = 0$ and $V(g, w_2) = 1$:



$\mathscr{M}$ falsifies at w all the following formulas:

$$
\begin{aligned}
&\sim\mathrm{K}\,g \to \neg\,\mathrm{K}\,g\\
&\sim\mathrm{K}\,g \to \sim g\\
&\sim\mathrm{K}\,g \to \neg g\\
&\sim\mathrm{K}\,\neg g \to \neg\neg g.
\end{aligned}
$$

On the other hand, $\vDash_w \sim\mathrm{K}\,g \wedge \sim\mathrm{K}\,\neg g$; but the constraint of monotonicity is not met: not $\vDash_{w_1} \sim\mathrm{K}\,g$ and not $\vDash_{w_2} \sim\mathrm{K}\,\neg g$.

3 Neo-Verificationist Approaches

Does the paradox of knowability threaten the neo-verificationist, who normally equates truth with knowability rather than with knowledge?

Let us observe, first of all, that, even within the intuitionistic conceptual framework, it would be possible to suggest a definition of truth different from the one given above:

Definition 6 $\mathrm{TR}\,\alpha =_{\mathrm{def.}} \exists\sigma(\mathrm{proves}(\sigma, \alpha))$.

Of course, if this definition is proposed within the intuitionistic conceptual framework, the metalinguistic existential quantifier is to be understood intuitionistically: a proof of $\exists\sigma(\mathrm{proves}(\sigma, \alpha))$ is a procedure p whose execution yields, after a finite time, a pair $\langle\sigma, \pi\rangle$, where σ is a construction[30] and π is a proof of "σ proves α".

It is easy to see that Definition 6 is materially adequate. Define the following function f: if σ is a proof of α, $f(\sigma)$ is the following procedure p: (i) take σ; (ii) effect the observation π that σ proves α; (iii) construct the pair $\langle\sigma, \pi\rangle$. Since proofs are epistemically transparent, the observation π terminates after a finite time, and the pair $\langle\sigma, \pi\rangle$ is therefore a proof of $\exists\sigma(\mathrm{proves}(\sigma, \alpha))$, i.e. of $\mathrm{TR}\,\alpha$. Conversely, define the following function g: if p is a procedure whose execution yields, after a finite time, a pair $\langle\sigma, \pi\rangle$, where σ is a construction and π is a proof of "σ proves α", then $g(p)$ is σ; since proof observation is factive, σ is a proof of α.

Definitions 4 and 6 are not extensionally equivalent. Consider the sentence "Prime(n)", where n is some very large natural number, and suppose that the primality test has never been applied to n. Then one of the two statements "Prime(n)" and "$\neg$ Prime(n)" is true, according to Definition 6, since (i) we know the primality test, which is a procedure with the required properties, and (ii) we know that the primality test, if it were applied, would answer either that n is prime or that n is not prime. When truth is defined according to Definition 6, the truth of α is not a cognitive state, but empirical accessibility to a cognitive state which is a proof of α. On the other hand, neither "Prime(n) nor "$\neg$ Prime(n)" is true, according to Definition 4, since we have a proof of neither statement, owing to the fact that the primality test has not been applied to n. Hence, according to Definition 6 there are statements that are true although they are not known now, and possibly not even in the future; while according to Definition 4 there are no statements of this kind: there are only unknown statements waiting to be *made* true (i.e. known) by our activity of proving mathematical statements or coming to know empirical statements.

The essential point to notice in this connection is that the following formula is intuitionistically valid:

$$\exists\sigma(\mathrm{proves}(\sigma, \alpha)) \leftrightarrow \mathrm{K}\,\alpha, \tag{20}$$

[30]Throughout the present paper I adhere to the intuitionistic idea that proofs (of mathematical statements) belong to a domain of mental constructions. As a matter of fact, I find it much more appropriate to conceive proofs as belonging to the category of cognitive states; on this point see [24], §2.

since both subformulas are equivalent to the same formula α. How is this possible? The validity of (20) puts dramatically into evidence a peculiarity of intuitionistic logic which deserves being stressed. Suppose that the procedure p described above, were it applied, would give as a result that n is prime. According to Definition 6, what determines the truth of "Prime(n)" before the execution of p is a mere fact (if it is a fact): the fact that the execution of p will give as a result a proof of "Prime(n)". Now, the essential characteristic of intuitionistic logic, as Heyting conceives it, is its being a *logique du savoir*, opposed to classical logic as a *logique de l'être*[31]; this entails that the intuitionistic meaning of the logical constants, implication in particular, must be explained in terms of cognitive states instead of facts and relations between facts.[32] Hence, the mere fact that "Prime(n)" is true before the execution of p plays no role in determining the assertibility or the non-assertibility of any formula of intuitionistic logic; in particular, it does not conflict with the validity of (20).

As a consequence, if one wanted to define a notion of intuitionistic truth by means of Definition 6 instead of 4, one would face a dilemma: either to accept (20), whose validity follows from the fact that the biconditional is read intuitionistically, giving up the possibility of expressing the fact that there are true but unknown statements; or to insist that there are intuitionistically true but unknown statements, giving up the intuitionistic reading of the logical constants occurring in the semantic metalanguage.

The moral drawn from this dilemma by the realist is clear: there are statements that are intuitionistically true but unknown; hence, as shown by the paradox, there are also statements that are intuitionistically true but unknowable; therefore (K) must be rejected. Equally clear is the moral drawn by the intuitionist: both linguistic and metalinguistic logical constants must be read intuitionistically, hence (20) is valid, and the notion of truth defined by Definition 6 either is to be rejected, or has intuitive consequences that cannot be expressed in an intuitionistic language. There is, however, a third answer that can be, and has been, proposed—an answer I should call "hybrid": it consists in defining truth by Definition 6, in insisting that there are intuitionistically true but unknown statements, in giving up the intuitionistic reading of the metalinguistic logical constants, adopting for them a classical reading, and in accepting (K). This position is instantiated by whoever accepts (K) rejecting at the same time (RAR); for the reason why (RAR) is judged unacceptable can be only that it is understood as expressing the thought that every α is either false or known, i.e. is understood on the basis of the classical reading of the implication occurring in its formalization.

---

[31] «Heyting [10] has opposed intuitionistic logic as the logic of knowledge (*logique du savoir*) to classical logic of existence (*logique de l'être*).» ([11], p. 107).

[32] This is the content of what Heyting calls *principle of positivity*: «Every mathematical or logical theorem must express the result of a mathematical construction» ([11], p. 108. See also [10], p. 231). In the case of implication, in particular, Heyting holds that within classical logic «il n'y a pas de place pour une implication proprement dite, car chaque proposition est vraie ou fausse, et on ne conçoit pas comment sa vérité pourrait dépendre de celle d'autres propositions.» [10, p. 226] On the contrary, «il est tout naturel que la démonstration d'une proposition dépende de la démonstration d'une autre proposition.» (p. 233).

Among the supporters of the hybrid position there are many neo-verificationists, in my opinion. Be it as it may, it seems to me that this position incurs a paradox strictly analogous to the paradox of knowability. Assume

$$q \wedge \neg \mathrm{K}\, q; \tag{21}$$

then, by Definition 6, there is a proof of $q \wedge \neg \mathrm{K}\, q$; let's call such proof σ; then

$$\mathrm{proves}(\sigma, (q \wedge \neg \mathrm{K}\, q)); \tag{22}$$

by the definition of proof of a conjunction,[33]

$$\sigma = \langle \sigma_1, \sigma_2 \rangle, \quad \text{where } (\mathrm{proves}(\sigma_1, q) \;\&\; \mathrm{proves}(\sigma_2, \neg \mathrm{K}\, q)). \tag{23}$$

Since σ_1 proves q, and proofs are epistemically transparent, it is possible to perform the observation σ_3 that σ_1 proves q, and this observation is a proof of $\mathrm{K}\, p$; then

$$\mathrm{proves}(\sigma_3, \mathrm{K}\, q); \tag{24}$$

if we now construct the pair $\sigma' = \langle \sigma_3, \sigma_2 \rangle$, we have that

$$\mathrm{proves}(\sigma', (\mathrm{K}\, q \wedge \neg \mathrm{K}\, q)), \tag{25}$$

hence

$$\Sigma \sigma\, (\mathrm{proves}(\sigma, \bot)); \tag{26}$$

on the other hand, the meaning of $\bot$ is characterized by saying that there is no proof of $\bot$, hence the formula

$$-\Sigma \sigma\, (\mathrm{proves}(\sigma, \bot)) \tag{27}$$

is assertible: a contradiction. Therefore,

$$-(q \wedge \neg \mathrm{K}\, q), \tag{28}$$

from which

$$q \supset \mathrm{K}\, q. \tag{29}$$

[33]Cesare Cozzo argues, in [1], p. 76, that the existence of σ_1 and σ_2 is not a contradiction because the existence of σ_2 does not imply that there is a not proof of q. This may be conceded, but it does not solve the paradox, as the next steps show.

A possible way out consists in giving up the intuitionistic idea that proofs are epistemically transparent; in this way the step from (23) to (24) is blocked. But the price to pay is very high: as proof, or more generally verification, is the key-notion of a neo-verificationist theory of meaning, the non-transparency of proofs/verifications would create the same difficulties the neo-verificationists impute to the realist theory of meaning because of the non-transparency of truth-conditions (essentially, the non-satisfiability of the manifestability requirement imposed onto knowledge of meaning).

Another way out has been proposed by Dummett. As a matter of fact, Dummett has tackled the paradox in two papers,[34] suggesting two different answers; since the former has been explicitly withdrawn by him,[35] I will consider only the latter. Dummett's solution consists in accepting

$$\alpha \to \neg\neg \mathrm{K}\,\alpha, \tag{30}$$

rejecting at the same time (2). This is legitimated, firstly, by the fact that only (30), not (2), follows intuitionistically from (1); secondly, by the fact that, if one reads negation intuitionistically,

> '$\neg\neg \mathrm{K}\,\alpha$' means 'There is an obstacle in principle to our being able to deny that α will ever be known', in other words 'The possibility that α will come to be known always remains open'[36]

—which is precisely what the verificationist believes to hold good for every true α. Dummett does not explain why (2) should be rejected; he only remarks that what (2) says is «contrary to our strong intuition» (p. 51). As I remarked at the beginning, what (2) says is not contrary to our intuition if (2) is read intuitionistically; on the contrary, it certainly *is* contrary to our intuition if what it says is that either the fact that α does not obtain, or the fact that α is (or will ever be) known obtains; but this is precisely the classical reading of the implication occurring in (2). Hence, Dummett is reading classically the implication in (30), intuitionistically the double negation. Such a hybrid reading is not justified; as a consequence, Dummett's solution seems quite *ad hoc*.

4 How Is a Rational Discussion Possible?

One essential ingredient of the solution I have proposed is the remark that, when the logical constants are understood intuitionistically, the formalization (2′) of (RAR) becomes perfectly acceptable. On the other hand, when the logical constants are

[34] References [6, 8].

[35] «I do not stand by the resolution of this paradox I proposed in "Victor's Error", a piece I wrote in a mood of irritation with the paradox of knowability.» [7, p. 348].

[36] Reference [8], p. 52.

understood classically, (2) is utterly unacceptable. This situation is far from surprising; on the contrary, it illustrates a general truth reminded above: the classical meaning of the logical constants is deeply different from their intuitionistic meaning. Consider for instance the schema "α or not α": classically understood (i.e., formalized as $\alpha + -\alpha$) it expresses the intuitively true principle that every proposition is either true or false, (*intuitively* true because our common-sense or pre-theoretic intuitions about the world are predominantly realistic) whereas intuitionistically understood (i.e. formalized as $\alpha \vee \neg\alpha$) it expresses the intuitively false principle that every proposition is decidable in the sense that there is either a proof or a refutation of it.

However, this situation generates a serious problem: the problem whether a rational discussion between a supporter of classical logic and a supporter of intuitionistic logic is possible at all. How is it possible that there is real disagreement or real agreement between them, given that both disagreement and agreement about a principle presuppose that the same meaning is assigned to it by both parties, while, as we have just seen, the meaning of one and the same formula drastically changes across classical and intuitionistic readings?

It seems to me that there are at least two alternative strategies to tackle the problem. The first consists in placing the discussion between the two parties *before* the formalization of the intuitive notions (as the logical constants, the notion of truth, and so on) into a formal language. The discussion, in this case, concerns questions like the following:
(i) Which intuitive notions should be formalized? For instance: inclusive or exclusive disjunction? Which notion of implication? Which notion of truth?
(ii) Which intuitive notion should be chosen as the key-notion of the theory of meaning, i.e. as the notion in terms of which the meaning of the expressions of the formal language (in particular of the logical constants) is to be characterized? For instance: (bivalent) truth (as the realist claims), or knowability/existence of a proof (as the neo-verificationist claims), or knowledge/actual proof (as the intuitionist claims)?

In this case the problem can be solved, *provided that* each party accepts the *intelligibility* of the key-notion adopted by the other party; for only in this case a rational discussion is possible: the same intuitive notions are accessible to both parties, and the disagreement concerns the legitimacy, the adequacy, the fruitfulness, etc. of adopting one notion or another as the key-notion. From this standpoint, Brouwer's idea that such classical notions as bivalent truth or actual infinity are unintelligible is to be abandoned, in favor of a slightly different claim: that those classical notions, precisely because they are intelligible, turn out to be incapable to play the foundational role the classicist gives them. Of course, such a claim should be motivated by a rational argument; which means that a rational discussion would be possible.

The second strategy consists in placing the discussion between the two parties *after* the formalization of the intuitive notions. In this case the problem of course arises, owing to the fact that the choice of different key-notions for the theory of meaning induces differences in the meaning of the logical constants. However, there may be tactics to solve it.

I hold the first alternative is better, but I have not an a priori argument; I will argue for my thesis by considering what seems a very plausible tactics and explaining why, in my opinion, it is not viable.

The tactics is based on the idea of translating one logic into the other, analogously to the case of the translation of a language into another. As a matter of fact, there are several so-called 'translations', both of classical logic/mathematics into intuitionistic logic/mathematics—the so-called negative translations (by Kolmogorov, Gödel, Gentzen, Kuroda and others); and of intuitionistic logic/mathematics into extensions of classical logic/mathematics (Shapiro, Horsten, Artemov). I shall not enter here into a detailed discussion of this tactics. I want only to stress an obvious fact: that the so-called 'translations' are not translations at all. A translation, in general, must be correct, and it is correct if it is meaning-preserving, i.e. if, for every expression E of $\mathscr{L}$ (the language to be translated), its translation $\mathrm{Tr}(E)$ into $\mathscr{L}'$ has the same meaning as E (whatever meaning is). But there is no reason to believe that the 'translations' mentioned above are meaning-preserving. Consider for instance the BHK clause for implication; Shapiro himself admits that the notion of "transformations of proofs" cannot be captured in the language of Epistemic Arithmetic, and Smoryński has observed that the 'translation' of intuitionistic logic into epistemic logic «does not capture the full flavor of talk about methods» (p. 1497).[37] To make another example, Kuroda's negative translation is based on a simple idea: that intuitionistic double negation is a sort of 'equivalent' of classical truth; this is surely true if one aims at a faithful 'immersion' of classical logic into intuitionistic logic (i.e. at a representation preserving theoremhood), but not if one aims at a genuine translation, for the classical truth of α, expressed by its occurrence within any formula, is something very different from the existence of an obstacle in principle to our being able to deny that α, expressed by $\neg\neg\alpha$. Moreover, there seems to be a conceptual reason for the impossibility of a genuine translation of one logic into another: on the one hand, a translation is correct only if it is meaning-preserving; on the other hand, classical logic explains the meaning of the logical constants in terms of a notion (bivalent truth) the intuitionist considers unintelligible or illegitimate, and also the converse is true (the classicist finds mysterious the intuitionistic notion of general method or effective function): so it seems unlikely that one of them finds in his own language an expression with the same meaning of an expression of the other's language.

5 Conclusion

The Paradox of Knowability is a paradox if the logical constants occurring in its formalization are understood according to the realist explanation of their meaning; but in a discussion between realists and anti-realists one cannot assume that anti-realists understand them in this way, for the paradox is intended to be an argument by which the former try to convince the latter to abandon their views on the meaning

[37] Reference [13], p. 9.

of the logical constants, and such an argument cannot be convincing if, in order to be formulated, it requires anti-realists to give up preventively their views. Vice versa, the paradox completely vanishes when the logical constants occurring in its formalization are understood according to the BHK explanation, since it is now necessary to distinguish two notions of truth: internal intuitionistic truth, which coincides with knowledge, and intuitive truth, essentially consisting in correspondence to external reality; in the former sense it is obvious that every truth is known, in the latter it is equally obvious—also for the anti-realist—that not every truth is known, and also that not every truth is knowable. From this point of view, the view of the paradox as an argument against anti-realism is the result of a wrong way of conceiving the rules of a rational discussion between classicist/realist and intuitionist/anti-realist.

In conclusion, the Paradox of Knowability leaves the debate between realists and anti-realists at the same point it was before its discovery. The crucial point of the debate is which notion between truth and evidence should be adopted as the key notion of the theory of meaning, or—if we accept the (in my opinion misleading) idea that meaning is to be explained in any case in terms of truth-conditions—which notion of truth, between bivalent and non-bivalent truth, the theory of meaning should be built on; in this case, the criterion for distinguishing realism from anti-realism cannot be the acceptance or refusal of the intuitive principle (K), but the acceptance or refusal of the principle of bivalence, according to Dummett's original suggestion.

Acknowledgments I am indebted to Julien Murzi and Luca Tranchini for helpful comments on earlier versions of this paper. The work reported here was supported by the MIUR fund No. 20107738C5_002.

References

1. Cozzo, C.: What can we learn from the paradox of knowability? Topoi **13**(2), 71–78 (1994)
2. Dummett, M.: Realism. In: [4], pp. 145–165. Paper read at the Oxford Philosophical Society (1963)
3. Dummett, M.: Elements of Intuitionism. Clarendon Press, Oxford (1977)
4. Dummett, M.: Truth and Other Enigmas. Duckworth, London (1978)
5. Dummett, M.: The Logical Basis of Metaphysics. Duckworth, London (1991)
6. Dummett, M.: Victor's error. Analysis **61**, 1–2 (2001)
7. Dummett, M.: Reply to Wolfgang Künne. In: Auxier, R.E., Hahn, L.E. (eds.) The Philosophy of Michael Dummett. The Library of Living Philosophers, pp. 345–350. Open Court, Chicago (2007)
8. Dummett, M.: Fitch's paradox of knowability. In: [18], pp. 51–52 (2009)
9. Gurevich, Y.: Intuitionistic logic with strong negation. Studia Logica **36**(1/2), 49–59 (1977)
10. Heyting, A.: La conception intuitionniste de la logique. Les études philosophiques **2**, 226–233 (1956)
11. Heyting, A.: Intuitionism in mathematics. In: Klibansky, R., (ed.) Philosophy in the Mid-century. A Survey, pp. 101–115. La Nuova Italia, Firenze (1958)

12. Heyting, A.: On truth in mathematics. Verslag van de plechtige viering van het honderdvijftigjarig bestaan der Koninklijke Nederlandse Akademie van Wetenschappen, pp. 277–279. North Holland, Amsterdam (1958)
13. Horsten, L.: In defense of epistemic arithmetic. Synthese **116**, 1–25 (1998)
14. Kripke, S.: Outline of a theory of truth. J. Philos. **72**, 690–716 (1975)
15. Milne, P.: Tarski, truth and model theory. In: Proceedings of the Aristotelian Society, vol. XCIX, pp. 141–167 (1999)
16. Murzi, J.: Knowability and bivalence: intuitionistic solutions to the paradox of knowability. Philos. Stud. **149**(2), 269–281 (2010)
17. Salerno, J.: Knowability Noir: 1945–1963. In: [18], pp. 29–48 (2009)
18. Salerno, J. (ed.): New Essays on the Knowability Paradox. Oxford University Press, Oxford (2009)
19. Smoryński, C.A.: Review of Intensional Mathematics by Shapiro. J. Symb. Log. **56**, 1496–1499 (1991)
20. Tarski, A.: The Concept of Truth in Formalized Languages. Logic, Semantics, Metamathematics, pp. 152–278. Clarendon Press, Oxford (1936)
21. Tarski, A.: The semantic conception of truth and the foundations of semantics. Philos. Phenomenol. Res. **4**(3), 341–376 (1944)
22. Usberti, G.: Anti-realist truth and truth-recognition. Topoi **31**(1), 37–45 (2012)
23. Usberti, G.: Gettier problems, C-justifications, and C-truth-grounds. In: Moriconi, E., Tesconi, L. (eds.) Second Pisa Colloquium in Logic, Language and Epistemology, pp. 325–361. ETS, Pisa (2014)
24. Usberti, G.: A notion of C-justification for empirical statements. In: Wansing, H. (ed.) Dag Prawitz on Proofs and Meaning, pp. 415–450. Springer, Cham (2015)
25. Williamson, T.: Knowability and constructivism. Philos. Q. **38**(153), 422–432 (1988)
26. Williamson, T.: Never say never. Topoi **13**(2), 135–145 (1994)
27. Williamson, T.: Knowledge and Its Limits. Oxford University Press, Oxford (2000)

Explicit Composition and Its Application in Proofs of Normalization

Jan von Plato

Abstract The class of derivations in a system of logic has an inductive definition. One would thus expect that crucial properties of derivations, such as normalization in natural deduction or cut elimination in sequent calculus or consistency in arithmetic be proved by induction on the last rule applied. So far it has not been possible to implement this simple requirement uniformly. It is suggested that such proofs can be carried through by a *'Hilfssatz'* methodology that is hidden in Gentzen's original unpublished proof of the consistency of arithmetic: to prove that a suitably chosen property of derivations is maintained under the composition of two derivations. As examples, new proofs by induction on the last rule in a derivation are given for normalization and strong normalization in natural deduction.

Keywords Natural deduction · Strong normalization · Explicit composition · Bar induction

1 Introduction

The rules of inference of a logical system define an inductive class of formal derivations. The most natural way to prove properties for the class is by induction on the construction of derivations, i.e., by induction on the last rule applied. It is often a crucial component in such proofs to show that the property in question is maintained under the composition of two derivations, even if this aspect is regularly ignored and the composability of derivations taken for granted. Results that show composition to maintain properties of derivations were called *Hilfssätze* in work of Gentzen that remained unpublished in its time. His original proof of the consistency of arithmetic of 1935 contained a *Hilfssatz* by which the 'reducibility of sequents' is maintained under composition. After he changed this proof into one that used transfinite induction, all traces of the *Hilfssatz* disappeared (see von Plato 2015 [8] for details).

J. von Plato (✉)
Department of Philosophy, University of Helsinki, Helsinki, Finland
e-mail: jan.vonplato@helsinki.fi

T. Piecha and P. Schroeder-Heister (eds.), *Advances in Proof-Theoretic Semantics*,
Trends in Logic 43, DOI 10.1007/978-3-319-22686-6_8

A formal implementation of the *Hilfssatz* methodology requires that composition be made into an explicit rule that is added to the logical rules of a calculus. The following results are shown as illustrations of the use of such an explicit composition rule: (1) A proof of normalization by a *Hilfssatz* for intuitionistic natural deduction. (2) A proof of strong normalization by bar induction.

2 Notation for Natural Derivations

The rules of natural deduction are production rules by which the class of formal derivations is defined inductively. Whenever there is such a definition, the most natural way to prove properties of the corresponding class is by induction on the last rule applied. This is so also in proof theory; a proof of normalization for intuitionistic natural deduction is given as a first example.

For a uniform treatment, we use natural deduction with general elimination rules and the related notion of normal derivability in which the condition is that the major premisses of elimination rules have to be assumptions. The modified rules are, with the multiplicity $n, m \geqslant 0$ of closed formulas indicated by exponents as in A^n, B^m (Table 1).

The normalizability result to be presented can be worked out also for the standard rules that can be seen as special cases of the general ones (Table 2).

It will be convenient in this situation to leave out the degenerate derivations of the minor premisses, to have exactly the Gentzenian rules.

In the standard tree notation for natural derivations, as above, the composition of two derivations can be indicated schematically, as in:

$$\begin{array}{c}\Gamma\\ \vdots\\ D\end{array} \quad \text{and} \quad \begin{array}{c}D\ \Delta\\ \vdots\\ C\end{array} \quad \text{compose into} \quad \begin{array}{c}\Gamma\\ \vdots\\ D\ \Delta\\ \vdots\\ C\end{array}$$

Table 1 General E-rules for &, $\supset$, $\forall$

$\dfrac{A \,\&\, B \quad \begin{array}{c}\overset{1}{A^n},\ \overset{1}{B^m}\\ \vdots\\ C\end{array}}{C}\ \&E,1$	$\dfrac{A \supset B \quad A \quad \begin{array}{c}\overset{1}{B^n}\\ \vdots\\ C\end{array}}{C}\ \supset E,1$	$\dfrac{\forall x A \quad \begin{array}{c}\overset{1}{A(t/x)^n}\\ \vdots\\ C\end{array}}{C}\ \forall E,1$

Table 2 Gentzen's E-rules as special cases of general E-rules.

$\dfrac{A \,\&\, B \quad \overset{1}{A}}{A}\ \&E,1$	$\dfrac{A \,\&\, B \quad \overset{1}{B}}{B}\ \&E,1$	$\dfrac{A \supset B \quad A \quad \overset{1}{B}}{B}\ \supset E,1$	$\dfrac{\forall x A \quad \overset{1}{A(t/x)}}{A(t/x)}\ \forall E,1$

Composition has the condition that the eigenvariables and discharge labels of the two derivations be distinct, if not, they can be changed.

No trace is left of the composition in the rightmost derivation. As the calculus is defined by its logical rules, composition in natural deduction is usually left implicit. To represent the composition of two derivations formally and to reason about its properties in a convenient form, we write the logical rules and the additional rule of composition in *sequent calculus style*, with the open assumptions of each formula D in a derivation written out as a multiset Γ in a sequent $\Gamma \rightarrow D$.

More formally, we define a root-first translation into sequent calculus style. If the last rule is $\& I$, we have:

$$\dfrac{\begin{array}{c}\Gamma\\ \vdots\\ A\end{array} \quad \begin{array}{c}\Delta\\ \vdots\\ B\end{array}}{A \,\&\, B}\ \&I \quad \rightsquigarrow \quad \dfrac{\begin{array}{c}\vdots\\ \Gamma \rightarrow A\end{array} \quad \begin{array}{c}\vdots\\ \Delta \rightarrow B\end{array}}{\Gamma, \Delta \rightarrow A \,\&\, B}\ \&I$$

$\vee I$ is similar, and $\supset I$ is:

$$\dfrac{\begin{array}{c}1\\ A^n, \Gamma\\ \vdots\\ B\end{array}}{A \supset B}\ \supset I,1 \quad \rightsquigarrow \quad \dfrac{\begin{array}{c}\vdots\\ A^n, \Gamma \rightarrow B\end{array}}{\Gamma \rightarrow A \supset B}\ \supset I$$

The translation continues from the premisses until assumptions are reached. The logical rules of the calculus **NLI** are obtained by translating the rest of the logical rules into sequent notation. The nomenclature **NLI** was used in some early manuscripts of Gentzen to denote a "natural-logistic intuitionistic calculus" (Table 3).

Table 3 Calculus **NLI**

$\dfrac{\Gamma \rightarrow A \,\&\, B \quad A^n, B^m, \Delta \rightarrow C}{\Gamma, \Delta \rightarrow C}\ \&E$	$\dfrac{\Gamma \rightarrow A \quad \Delta \rightarrow B}{\Gamma, \Delta \rightarrow A \,\&\, B}\ \&I$
$\dfrac{\Gamma \rightarrow A \vee B \quad A^n, \Delta \rightarrow C \quad B^m, \Theta \rightarrow C}{\Gamma, \Delta, \Theta \rightarrow C}\ \vee E$	$\dfrac{\Gamma \rightarrow A}{\Gamma \rightarrow A \vee B}\ \vee I_1 \quad \dfrac{\Gamma \rightarrow B}{\Gamma \rightarrow A \vee B}\ \vee I_2$
$\dfrac{\Gamma \rightarrow A \supset B \quad \Delta \rightarrow A \quad B^m, \Theta \rightarrow C}{\Gamma, \Delta, \Theta \rightarrow C}\ \supset E$	$\dfrac{A^n, \Gamma \rightarrow B}{\Gamma \rightarrow A \supset B}\ \supset I$
$\dfrac{\Gamma \rightarrow \forall x A(x) \quad A(t)^n, \Delta \rightarrow C}{\Gamma, \Delta \rightarrow C}\ \forall E$	$\dfrac{\Gamma \rightarrow A(y)}{\Gamma \rightarrow \forall x A(x)}\ \forall I$
$\dfrac{\Gamma \rightarrow \exists x A(x) \quad A(y)^n, \Delta \rightarrow C}{\Gamma, \Delta \rightarrow C}\ \exists E$	$\dfrac{\Gamma \rightarrow A(t)}{\Gamma \rightarrow \exists x A(x)}\ \exists I$

Natural deduction in sequent style

The calculus is completed by adding initial sequents of the form $A \rightarrow A$, with A an arbitrary formula, and the zero-premiss rule $\perp E$ by which $\perp \rightarrow C$ can begin a derivation branch.

We say that the closing of an assumption formula in E-rules and in rule $\supset I$ is *vacuous* if $n = 0$ or $m = 0$. Similarly, the closing of an assumption is *multiple* if $n > 1$ or $m > 1$. With $n = 1$ or $m = 1$, the closing of an assumption is *simple*. Vacuous and multiple closing of assumptions is seen in:

$$\dfrac{\begin{array}{c}\Gamma\\ \vdots\\ B\end{array}}{A \supset B}\supset I \qquad \dfrac{\begin{array}{c}\overset{1}{A},\ \overset{1}{A},\ \Gamma\\ \vdots\\ B\end{array}}{A \supset B}\supset I,1$$

The former case corresponds to the situation in sequent calculus in which a formula active in a logical rule stems from a step of weakening, the latter to a situation in which it stems from a step of contraction, as shown in von Plato (2001) [5].

The composition of two derivations is an essential step in the normalization of derivations. It can now be written quite generally in the form:

$$\dfrac{\Gamma \rightarrow D \quad D, \Delta \rightarrow C}{\Gamma, \Delta \rightarrow C}\,Comp$$

Iterated compositions appear as so many successive instances of rule *Comp*.

In a *permutative conversion*, the height of derivation of a major premiss derived by $\vee E$ or $\exists E$, i.e., number of successive steps of inference, is diminished. The effect of the general rules is that such conversions work for all derived major premisses of elimination rules:

Definition 1 A derivation in natural deduction with general elimination rules is *normal* if all major premisses of E-rules are assumptions.

As a first step towards normalization, we need to show that derivations in natural deduction can be composed:

Lemma 1 (Closure of derivations with respect to composition) *If given derivations of the sequents $\Gamma \rightarrow D$ and $D, \Delta \rightarrow C$ in **NLI** are composed by rule Comp to conclude the sequent $\Gamma, \Delta \rightarrow C$, the instance of Comp can be eliminated.*

Proof We show by induction on the height of derivation of the right premiss of *Comp* that it can be eliminated.
1. Base case. The second premiss of *Comp* is an initial sequent, as in:

$$\dfrac{\Gamma \rightarrow D \quad D \rightarrow D}{\Gamma \rightarrow D}\,Comp$$

The conclusion of *Comp* is identical to its first premiss, so that *Comp* can be deleted.

If the second premiss is of the form $\bot \to D$, the first premiss is $\Gamma \to \bot$. It has not been derived by a right rule, so that *Comp* can be permuted up in the first premiss. In the end, a topsequent $\Gamma' \to \bot$ is found as the left premiss of *Comp*, by which $\bot$ is in Γ', so that the conclusion of *Comp* is an initial sequent.
2. Inductive case with the second premiss of *Comp* derived by an I-rule. There are two subcases, a one-premiss rule and a two-premiss rule. In the former case, *Comp* is permuted up to the premiss, with a lesser height of derivation as a result. In the latter case, we use the notation (D) to indicate a possible occurrence of D in a premiss:

$$\dfrac{\Gamma \to D \qquad \dfrac{(D), \Delta' \to C' \quad (D), \Delta'' \to C''}{D, \Delta', \Delta'' \to C}\,Rule}{\Gamma, \Delta', \Delta'' \to C}\,Comp$$

Rule *Comp* is permuted to any premiss that has an occurrence of D, say the first one, with the result:

$$\dfrac{\dfrac{\Gamma \to D \quad D, \Delta' \to C'}{\Gamma, \Delta' \to C'}\,Comp \qquad \Delta'' \to C''}{\Gamma, \Delta', \Delta'' \to C}\,Rule$$

3. Inductive case with the second premiss of *Comp* derived by an E-rule, as in:

$$\dfrac{\Gamma \to D \qquad \dfrac{(D), \Delta \to A \,\&\, B \quad (D), A^n, B^m, \Theta \to C}{D, \Delta, \Theta \to C}\,\&E}{\Gamma, \Delta, \Theta \to C}\,Comp$$

As in case 2, *Comp* is permuted up, to whichever premiss has an occurrence of the composition formula D, with a lesser height of derivation as a result. The other cases of E-rules are entirely similar. QED.

In the case of a multiple discharge, a detour conversion will lead to several compositions, with a multiplication of the contexts as in the example

$$\dfrac{\dfrac{\Gamma \to A \quad \Delta \to B}{\Gamma, \Delta \to A \,\&\, B}\,\&I \qquad A, A, B, \Theta \to C}{\Gamma, \Delta, \Theta \to C}\,\&E$$

The conversion is into

$$\dfrac{\Delta \to B \qquad \dfrac{\Gamma \to A \qquad \dfrac{\Gamma \to A \quad A, A, B, \Theta \to C}{A, B, \Gamma, \Theta \to C}\,Comp}{B, \Gamma, \Gamma, \Theta \to C}\,Comp}{\Gamma, \Gamma, \Delta, \Theta \to C}\,Comp$$

Such multiplication does not affect the normalization process. Note well that normalization depends on the admissibility of composition which latter has to be proved *before* normalization.

3 Normalization by *Hilfssatz*

In normalization, derived major premisses of E-rules are converted step by step into assumptions. There are two situations, depending on whether the major premiss was derived by an E-rule or an I-rule:

Definition 2 (*Normalizability*) A derivation in **NLI** is *normalizable* if there is a sequence of conversions that transform it into normal form.

The idea of our proof of the normalization theorem is to show by induction on the last rule applied in a derivation that logical rules maintain normalizability.

The cut elimination theorem is often called *Gentzen's Hauptsatz*, main theorem. He used the word *Hilfssatz*, auxiliary theorem or lemma, for an analogous result by which composition of derivable sequents maintains the reducibility of sequents, a property defined in his original proof of the consistency of arithmetic (Gentzen 1935 [2, p. 106]). Henceforth any result in proof theory in which it is shown that a property of sequents or derivations is maintained under composition shall be called a *Hilfssatz*. Normalizability will be the first such property to be proved.

Theorem 1 (Normalizability for intuitionistic natural deduction) *Derivations in* **NLI** *convert to normal form.*

Proof Consider the last rule applied. The base case is an assumption that is a normal derivation. In the inductive case, if an I-rule is applied to premisses the derivations of which are normalizable, the result is a normalizable derivation. The same holds if a normal instance of an E-rule is applied. The remaining case it that a non-normal instance of an E rule is applied. The major premiss of the rule is then derived either by another E-rule or an I-rule, so we have two main cases with subcases according to the specific rule in each. Derivations are so transformed that normalizability can be concluded either because the last rule instance resolves into possible non-normalities with shorter conversion formulas, or because the height of derivation of its premisses is diminished.

1. *E-rules*: Let the rule be $\&E$ followed by another instance of $\&E$, as in:

$$\dfrac{\dfrac{\Gamma \rightarrow A \,\&\, B \quad A^n, B^m, \Delta \rightarrow C \,\&\, D}{\Gamma, \Delta \rightarrow C \,\&\, D}\,\&E \quad C^k, D^l, \Theta \rightarrow E}{\Gamma, \Delta, \Theta \rightarrow E}\,\&E$$

By the inductive hypothesis, the derivations of the premisses of the last rule are normalizable. The second instance of $\&E$ is permuted above the first:

$$\dfrac{\Gamma \to A \,\&\, B \qquad \dfrac{A^n, B^m, \Delta \to C \,\&\, D \quad C^k, D^l, \Theta \to E}{A^n, B^m, \Delta, \Theta \to E}\,\&E}{\Gamma, \Delta, \Theta \to E}\,\&E$$

The height of derivation of the major premiss of the last rule instance in the upper derivation has diminished by 1, so the subderivation down to that rule instance is normalizable. The height of the major premiss of the other rule instance has remained intact and therefore normalizability follows.

All other cases of permutative convertibility go through in the same way.

2. *I-rules*: The second situation of convertibility is that the major premiss has been derived by an I-rule, and there are five cases:

2.1. Detour convertibility on &:

$$\dfrac{\dfrac{\Gamma \to A \quad \Delta \to B}{\Gamma, \Delta \to A \,\&\, B}\,\&I \qquad A^n, B^m, \Theta \to C}{\Gamma, \Delta, \Theta \to C}\,\&E$$

Let us assume for the time being that $n = m = 1$. The detour conversion is given by:

$$\dfrac{\Delta \to B \qquad \dfrac{\Gamma \to A \quad A, B, \Theta \to C}{B, \Gamma, \Theta \to C}\,Comp}{\Gamma, \Delta, \Theta \to C}\,Comp$$

The result is not a derivation in **NLI** . We proved in Lemma 1 that *Comp* is eliminable. The next step is to show that *Comp* maintains normalizability. This will be done in the *Hilfssatz* to be proved separately. By the *Hilfssatz*, the conclusion of the upper *Comp* is normalizable, and again by the *Hilfssatz*, also the conclusion of the lower *Comp*. If $n > 1$ or $m > 1$, *Comp* is applied repeatedly, the admissibility of an uppermost *Comp* giving the admissibility of the following ones. If $n = 0$, the instance of *Comp* with the left premiss $\Gamma \to A$ falls out of the derivation, and similarly with $m = 0$. If $n = m = 0$, the right premiss of rule & E before conversion is $\Theta \to C$, and it is taken in place of the original conclusion $\Gamma, \Delta, \Theta \to C$. This situation is called a 'simplification convertibility' in Prawitz (1965) [3]. In all cases, the result of conversion is uniquely defined.

2.2. Detour convertibility on $\vee$. There are two cases, as in:

$$\dfrac{\dfrac{\Gamma \to A}{\Gamma \to A \vee B}\,\vee I \qquad A^m, \Delta \to C \quad B^n, \Theta \to C}{\Gamma, \Delta, \Theta \to C}\,\vee E$$

$$\dfrac{\dfrac{\Gamma \to B}{\Gamma \to A \vee B}\,{\vee I} \quad A^m, \Delta \to C \quad B^n, \Theta \to C}{\Gamma, \Delta, \Theta \to C}\,{\vee E}$$

As in 2.1, assume for the time being that $n = m = 1$. The detour conversion is given by:

$$\dfrac{\Gamma \to A \quad A, \Delta \to C}{\Gamma, \Delta \to C}\,{Comp} \qquad \dfrac{\Gamma \to B \quad B, \Theta \to C}{\Gamma, \Theta \to C}\,{Comp}$$

The multiplicities are treated as in 2.1, except for the case of $m = 0$ or $n = 0$. Then the given derivation has a simplification convertibility, say when $m = n = 0$:

$$\dfrac{\dfrac{\Gamma \to A}{\Gamma \to A \vee B}\,{\vee I} \quad \Delta \to C \quad \Theta \to C}{\Gamma, \Delta, \Theta \to C}\,{\vee E}$$

There is a conversion, but it is not uniquely defined: Either one of the original minor premisses of $\vee E$ can be taken. Similarly, if say $n > 0$ and $m = 0$, either a composition with composition formula A can be made, or a simplification conversion.

2.3. Detour convertibility on $\supset I$:

$$\dfrac{\dfrac{A^n, \Gamma \to B}{\Gamma \to A \supset B}\,{\supset I} \quad \Delta \to A \quad B^m, \Theta \to C}{\Gamma, \Delta, \Theta \to C}\,{\supset E}$$

In the conversion, multiple discharge of assumptions is again resolved into iterated compositions, so we may assume $n = m = 1$ and have the conversion:

$$\dfrac{\dfrac{\Delta \to A \quad A, \Gamma \to B}{\Gamma, \Delta \to B}\,{Comp} \quad B, \Theta \to C}{\Gamma, \Delta, \Theta \to C}\,{Comp}$$

If $m = 0$, there is a simplification convertibility with the uniquely defined result $\Theta \to C$.

2.4. Detour convertibility on $\forall$:

$$\dfrac{\dfrac{\Gamma \to A(y)}{\Gamma \to \forall x A(x)}\,{\forall I} \quad A^n(t), \Delta \to C}{\Gamma, \Delta \to C}\,{\forall E}$$

As before, assume for the time being that $n = 1$. The eigenvariable y in the derivation of $\Gamma \rightarrow A(y)$ is replaced by the term t and the detour conversion given by:

$$\dfrac{\Gamma \rightarrow A(t) \quad A(t), \Delta \rightarrow C}{\Gamma, \Delta \rightarrow C}\,\mathit{Comp}$$

The multiplicities are treated as before.

2.5. Detour convertibility on $\exists$:

$$\dfrac{\dfrac{\Gamma \rightarrow A(t)}{\Gamma \rightarrow \exists x A(x)}\,\exists I \quad A^n(y), \Delta \rightarrow C}{\Gamma, \Delta \rightarrow C}\,\exists E$$

As before, assume for the time being that $n = 1$. The eigenvariable y in the derivation of $A(y), \Delta \rightarrow C$ is replaced by the term t, and the detour conversion is:

$$\dfrac{\Gamma \rightarrow A(t) \quad A(t), \Delta \rightarrow C}{\Gamma, \Delta \rightarrow C}\,\mathit{Comp}$$

Multiplicities are treated as before. QED.

It remains to give a proof of the *Hilfssatz*:

Hilfssatz 1 (*Closure of normalizability under composition*) *If the premisses of rule Comp are normalizable, also the conclusion is.*

Proof The proof is by induction on the length of the composition formula D with a subinduction on the sum of the heights of derivation of the two premisses.
1. $D \equiv P$. With an atomic formula P, we have

$$\dfrac{\Gamma \rightarrow P \quad P, \Delta \rightarrow C}{\Gamma, \Delta \rightarrow C}\,\mathit{Comp}$$

P is never principal in the right premiss, so that *Comp* can be permuted up with a lesser sum of heights of derivation as a result. There are two cases, a one-premiss rule and a two-premiss rule. For the latter, we use again the notation (P) to indicate a possible occurrence of P in a premiss:

$$\dfrac{\Gamma \rightarrow P \quad \dfrac{(P), \Delta' \rightarrow C' \quad (P), \Delta'' \rightarrow C''}{P, \Delta', \Delta'' \rightarrow C}\,\mathit{Rule}}{\Gamma, \Delta', \Delta'' \rightarrow C}\,\mathit{Comp}$$

Rule *Comp* is permuted to the premiss that has an occurrence of P, say the first one, with the result:

$$\dfrac{\dfrac{\Gamma \to P \quad P, \Delta' \to C'}{P, \Delta' \to C'}\,Comp \qquad \Delta'' \to C''}{\Gamma, \Delta', \Delta'' \to C}\,Rule$$

In the end, the second premiss of *Comp* is an initial sequent, as in:

$$\dfrac{\Gamma \to P \quad P \to P}{\Gamma \to P}\,Comp$$

The conclusion of *Comp* is identical to its first premiss, so that *Comp* can be deleted.
2. $D \equiv \bot$. Because $\bot$ is never principal in the left premiss, *Comp* is permuted up as in the proof of admissibility of composition.
3. $D \equiv A \,\&\, B$. If $A \,\&\, B$ is not principal in the right premiss, *Comp* can be permuted as in 1.

If $A \,\&\, B$ is principal, there has to be a normal rule instance in the right premiss, as in:

$$\dfrac{\Gamma \to A \,\&\, B \qquad \dfrac{A \,\&\, B \to A \,\&\, B \quad A^n, B^m, \Delta \to C}{A \,\&\, B, \Delta \to C}\,\&E}{\Gamma, \Delta \to C}\,Comp$$

Comp is permuted up to the first premiss:

$$\dfrac{\dfrac{\Gamma \to A \,\&\, B \quad A \,\&\, B \to A \,\&\, B}{\Gamma \to A \,\&\, B}\,Comp \qquad A^n, B^m, \Delta \to C}{\Gamma, \Delta \to C}\,\&E$$

Comp is now deleted and a generally non-normal instance of rule $\&E$ created. If the major premiss is concluded by an E-rule, a permutative conversion is done and no instance of *Comp* created. If the last rule is $\&I$, a detour convertibility with the conversion formula $A \,\&\, B$ is created. A detour conversion will lead to new instances of *Comp*, but on strictly shorter formulas.

The other cases of composition formulas are treated in a similar way. QED.

Lemma 1, closure of derivations with respect to composition, merely shows that a derivation in natural deduction can be got from two composable derivations. The *Hilfssatz* adds the property of preservation of normalizability. It is even important to give the details for the composition of derivations as in the proof of Lemma 1, for the algorithm of normalization depends crucially on the steps needed for the admissibility of composition. Even so, one searches in vain for more than a mere indication of this proof in the logical literature.

4 Strong Normalization by Bar Induction

Derivations are denoted by $d_0, d_1, d_2, \ldots$, and let $N(d)$ express that d is a normal derivation, i.e., that all major premisses of E-rules are initial sequents. This property can be decided by an inspection of the derivation. The choice sequences in normalization are defined as follows:

Definition 3 (*Conversion choice sequence for a derivation*) Given a derivation d, a *conversion choice sequence* for d is a succession of conversions on d with the restriction that whenever d has a permutative convertibility, it has to be chosen.

The restriction is in fact not necessary, but it will make the proof go through smoothly. It is not met if disjunction and existence are left out of the language and the standard elimination rules used, so there is sense in calling the result of this Section a strong normalization theorem.

We shall indicate by $PF(d)$ that a derivation d is free of permutative conversions.

The notation $\overline{\alpha}_n(d) \equiv d_n$ stands for the derivation that is obtained from a given derivation d after n steps of conversion $\overline{\alpha}_n$. The notation $\alpha_1(\overline{\alpha}_n(d)) \equiv \alpha_1(d_n)$ stands for the result of a one-step continuation of the sequence of conversions $\overline{\alpha}_n$.

Definition 4 (*Normalizing and strongly normalizing derivations*)

i. A derivation d is *normalizing* whenever $\exists\alpha\exists x N(\overline{\alpha}_x(d))$.
ii. A derivation d is *strongly normalizing* whenever $\forall\alpha\exists x N(\overline{\alpha}_x(d))$.

We write $WN(d)$ for the former and $SN(d)$ for the latter.

We shall use the standard formulation of bar induction in the proof of strong normalization, with the two predicates $PF(d)$ and $SN(d)$. It has to be established that: (1) The base case predicate $PF(d)$ is decidable. (2) Every conversion choice sequence of a given derivation d has an initial segment such that a permutation-free derivation is obtained. (3) Permutation-free derivations are strongly normalizing. (4) If every one-step continuation of conversions of a derivation d is strongly normalizing, also d is strongly normalizing.

Theorem 2 (Strong normalization for intuitionistic natural deduction) *Derivations in* **NLI** *are strongly normalizing.*

Proof We show in turn that the four conditions of bar induction are satisfied by the predicates $PF(d)$ and $SN(d)$. Let d_0 be the given derivation that we assume to be non-normal.

1. *Decidability*: $PF(d)$ is decidable as noted above.
2. *Termination of permutative conversions*: Let a derivation d have permutative convertibilities. As seen in the proof of normalization, each such conversion diminishes the height of derivation of the major premiss in question by 1 and leaves the other heights unaltered. Therefore permutative conversions terminate in a bounded number n of steps in a derivation d_n such that $PF(d_n)$.

3. *If PF(d), then SN(d)*: The proof is by induction on the last rule in d and we can assume d not to be normal and the derivations of the premisses to be strongly normalizing. By $PF(d)$, all non-normalities are detour convertibilities. Any conversion chosen resolves into compositions, and a *Hilfssatz* needs to be proved by which composition of derivations maintains strong normalizability. This is done below.
4. *If* $\forall\alpha_1 SN(\alpha_1(d_n))$, *then* $SN(d_n)$: Each one-step continuation of the conversion of d_n is by assumption strongly normalizing, therefore the derivation d_n is by definition strongly normalizing.
   By 1–4, $SN(d_0)$. QED.

It remains to add a proof of the *Hilfssatz* used in condition 3:

Hilfssatz 2 (*Closure of strong normalizability under composition*) *Given strongly normalizing derivations of* $\Gamma \to D$ *and* $D, \Delta \to C$, *their composition into a derivation of* $\Gamma, \Delta \to C$ *is strongly normalizing.*

Proof As before, the proof is by induction on the length of the composition formula D, with a subinduction on the sum of heights of derivation of the premisses of rule *Comp*, and goes through virtually identically to the proof of *Hilfssatz* 1. QED.

5 Concluding Remarks and Further Applications

Looking at the single detour conversion schemes in the proof of Theorem 1, we notice that simplification convertibility with disjunction in case 2.2 leaves two possible results of conversion. For the rest of detour conversions, the local transformations produce unique converted derivations, and that property is sufficient for the overall result: Bar induction is a principle by which such *local control* of a suitably chosen property is turned into *global structure*, one could put it.

There is at each stage of strong normalization a finite number of non-normalities from which to choose the conversion to be made. Therefore strong normalization is a consequence of the variety of bar induction known as the fan theorem. The consistency of arithmetic was originally proved by bar induction by Gentzen and soon replaced by a proof through transfinite induction (see von Plato 2015 [8], and Siders and von Plato (2015) [4] for an explicit formulation of Gentzen's bar induction). As with Gentzen's proof, also the present proof could be carried through by the use of transfinite ordinals. What the least ordinal needed is, is at present not known, but because the fan theorem suffices for the result, Gentzen's ε_0 gives a strict upper bound.

The proofs of normalization and strong normalization through *Hilfssätze* should work without problems for classical natural deduction with the rule of indirect proof and the same definition of normality as above, as in von Plato and Siders (2012) [9].

The proofs can obviously be worked through also for standard natural deduction, along the lines of my paper (von Plato 2011 [6]).

Two more applications of explicit composition can be noted here:

1. *The interpretation of arbitrary cuts in natural deduction*: A comparison of natural deduction in sequent calculus style with sequent calculus proper shows that a non-normal instance of an E-rule corresponds exactly to the case of a cut in which the right premiss of cut has been derived by a corresponding left rule. In the translation from sequent derivations with cuts to natural deduction, such cuts turn into non-normalities. The rest of the cuts are translated as explicit delayed compositions. What corresponds to cut elimination is seen from the admissibility of composition in natural deduction: An uppermost instance of *Comp* is permuted up until it either reaches an assumption and vanishes or hits a normal instance of an E-rule and gets turned into a non-normality. After the delayed compositions have been eliminated, there remain the proper non-normalitites and these can be eliminated in any order whatsoever. When in the normal derivation the major premisses are left unwritten, a sequent derivation is obtained. The overall procedure gives strong cut elimination in precisely the same sense in which there is strong normalization in natural deduction. Details are found in Sect. 13.4 of von Plato (2013) [7].

2. *Normalization and strong normalization of* λ*-terms*: Any proof of normalization and strong normalization can be turned into a corresponding proof for typed λ-terms. The term structure is particularly transparent with general elimination rules, for the selector terms have now, with implication elimination as an example, the following structure (von Plato 2001 [5, p. 566]):

$$\frac{c : A \supset B \quad a : A \quad \begin{array}{c}[x : B]\\ \vdots \\ d : C\end{array}}{gap(c, a, (x)d) : C}$$

A selector term is *normal* if its first argument is a variable, in particular, for the above "generalized application" as it is called in von Plato 2001 [5], the nested "tower" of applications, met with the standard application function, does not occur for normal terms. Permutative conversions reduce a suitably defined notion of depth of selector terms, and detour conversions reduce to substitutions. A *Hilfssatz* is used to prove that strong normalizability of λ-terms is maintained under such substitution.

References

1. Gentzen, G.: Untersuchungen über das logische Schließen. Mathematische Zeitschrift **39**, 176–210, 405–431 (1934–35)
2. Gentzen, G.: Der erste Widerspruchsfreiheitsbeweis für die klassische Zahlentheorie. First published in Archiv für mathematische Logik **16**(1974), 97–118 (1935)
3. Prawitz, D.: Natural Deduction: A Proof-Theoretical Study. Almqvist & Wiksell, Stockholm (1965)
4. Siders, A., von Plato, J.: Bar induction in the proof of termination of Gentzen's reduction procedure. In: Kahle, R., Rathjen, M. (eds.) Gentzen's Centenary: The Quest of Consistency, pp. 127–130. Springer, New York (2015)
5. von Plato, J.: Natural deduction with general elimination rules. Arch. Math. Log. **40**, 541–567 (2001)
6. von Plato, J.: A sequent calculus isomorphic to Gentzen's natural deduction. Rev. Symb. Log. **4**, 43–53 (2011)
7. von Plato, J.: Elements of Logical Reasoning. Cambridge University Press, Cambridge (2013)
8. von Plato, J.: From *Hauptsatz* to *Hilfssatz*. In: Kahle, R., Rathjen, M. (eds.) Gentzen's Centenary: The Quest of Consistency, pp. 89–126. Springer, New York (2015)
9. von Plato, J., Siders, A.: Normal derivability in classical natural deduction. Rev. Symb. Log. **5**, 205–211 (2012)

Towards a Proof-Theoretic Semantics of Equalities

Reinhard Kahle

Abstract We have a fresh look on Frege's *mode of presentation*, taking into account proofs of equalities as a key concept. Revisiting the classical example of *Morning star* and *Evening star* the account leads to a proposal for a proof-theoretic semantics of equalities.

Keywords Frege · Mode of presentation · Sinn und Bedeutung · Proof-theoretic semantics · Equality

1 Frege's Question

Gottlob Frege opened his seminal paper *Sinn und Bedeutung* [3] by asking what the epistemic difference is between the equations $a = a$ and $a = b$.[1] His proposal to distinguish between *sense* and *denotation* (or *reference*) of a term turned out to be one of the most fruitful conceptual advances in the history of philosophical logic.

Modern *Possible Worlds Semantics* draws on this distinction: the sense of a term refers to the full variety of possible worlds (in the way that we have to consider the denotation of a term in every possible world), while the (Fregean) denotation has to take into account only the actual world.

As appealing as this view might be, there are (at least) two problems with it. First, it comes with a concept of rigid designators. Second, it is not applicable to mathematics, because mathematical equations hold equally in every possible world.

While many approaches try to attack the problem from a semantic perspective, here we would like to provide a syntactic account, which takes up Frege's original question

[1] In fact, he doesn't put this directly as a question, but rather states that "$a = a$ holds *a priori* and, according to Kant, is to be labeled analytic, while statements of the form $a = b$ often contain very valuable extensions of our knowledge and cannot always be established *a priori*." [5, p. 157].

R. Kahle (✉)
Departamento de Matemática, CMA and CENTRIA, FCT, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal
e-mail: kahle@mat.uc.pt

T. Piecha and P. Schroeder-Heister (eds.), *Advances in Proof-Theoretic Semantics*, Trends in Logic 43, DOI 10.1007/978-3-319-22686-6_9

of the (epistemic[2]) difference of $a = a$ and $a = b$. With restriction to singular terms, we will propose a fresh understanding of Frege's *mode of presentation*. It is motivated by the question how we actually *prove* a particular equation, and it can be considered as a *proof-theoretic semantics of equalities*.

2 Equality Versus Identity

To set the stage for the further discussion we would like to assume that we always use a first-order language for which we may have some non-logical axioms, and a fixed structure with universe $\mathfrak{A}$ in which this language is interpreted[3] by an interpretation function $(\cdot)^{\mathfrak{M}}$. Let us use Latin characters for terms of the language, and German (Gothic) ones for elements of the structure. Equality is understood as the relation $t = s$ on the *syntactical* level between terms of the first-order language[4]; identity stands for the (trivial) relation $\mathfrak{a} \equiv \mathfrak{a}$ on the *semantic* level, which holds only between an object in the structure and itself.[5] The fact that *identity* is not entirely trivial comes from its use for terms in combination with the interpretation in the form $(t)^{\mathfrak{M}} \equiv (s)^{\mathfrak{M}}$.[6]

In this setting we can recast Frege's first observation that "if we were to regard equality as a relation between that which the names 'a' and 'b' designate, it would seem that $a = b$ could not differ from $a = a$ (i.e. provided $a = b$ is true). A relation would thereby be expressed of a thing to itself, and indeed one in which each thing stands to itself but to no other thing." [5, p. 157]. In our terminology we may say that we are not concerned with the semantical identity relation $\mathfrak{a} \equiv \mathfrak{a}$, but with the syntactical equality relation $t = s$.

It is standard to axiomatize equality in first-order logic as a universal congruence relation, i.e., an equivalence relation compatible with all operations (functions and relations). As such, it mimics on the syntactic level just the properties which identity exhibits on the semantic level. But the domain of identity is simply $\mathfrak{A} \times \mathfrak{A}$, and the elements of $\mathfrak{A}$ are unique in the sense that $\mathfrak{a} \equiv \mathfrak{a}$, but $\mathfrak{a} \not\equiv \mathfrak{b}$ for two elements $\mathfrak{a}, \mathfrak{b} \in \mathfrak{A}$. On the syntactic side, however, the equality relation is defined for terms, and, clearly, two terms, though being interpreted by the same object $\mathfrak{a}$, may well be different.

---

[2] In this paper, we restrict ourselves to an epistemic perspective, and we will not go into metaphysical issues. This seems also to be Frege's position, as he speaks explicitly about "our knowledge" (see the citation in the previous footnote).

[3] Here, the notion of structure includes the possibility that its universe is taken from the "real world".

[4] We make here a slight abuse of the word "relation"; strictly speaking, a relation is a semantic concept and "syntactic relation" cannot be anything more than a formula formed by use of a relation symbol.

[5] For a recent philosophical discussion of the concept of identity see also [14].

[6] This distinction of equality and identity may help to unravel Frege's initial footnote in *Sense and Denotation* explaining that he uses equality "in the sense of identity and understand '$a = b$' in the sense of 'a is the same as b' or 'a and b coincide'." [5, p. 157, slightly changed translation]; of course, Frege didn't have the modern distinction of syntax and semantics at hand.

Frege introduced the *sense* of a sign as its *mode of presentation*. Generally, this is taken as some kind of illustration rather than a definition. It is our aim to provide a more formal explication of *mode of presentation* by drawing on the difference of equality and identity.

3 The Mode of Presentation

Frege did not define the notion of mode of presentation, but he did give two quite illustrating examples of it.

The first one is taken from geometry: a particular point may be presented as intersection of two lines a and b or as intersection of the lines b and c. Though the intersections take place at the very same point, we would say that the two modes of presentation differ: one refers to the lines a and b, the other to the lines b and c.

Assuming a suitable axiom system for geometry, this system will provide terms which serve as definitions for the intersections expressed by, say, $Intsec(a, b)$ and $Intsec(b, c)$. Assuming that both terms refer to the same point $\mathfrak{p}$ of the plane, it requires some reasoning in the given axiomatic framework to derive the equality $Intsec(a, b) = Intsec(b, c)$. This equality is epistemically different from a simple reflexive equality, like $Intsec(a, b) = Intsec(a, b)$.

We propose to use $Intsec(a, b)$ to obtain a mode of presentation of $\mathfrak{p}$, and $Intsec(b, c)$ to obtain another mode of presentation of the same point.[7] We may say that a term t of our formal language *expresses* (to use Frege's wording) a mode of presentation if it may be used as a mathematical expression to define a newly introduced constant A. We do not say that the term *is* the mode of presentation, as—with Frege—the latter is surely not a syntactic object (this would be the *mode of designation*, [5, p. 157]). The way the mode of presentation should be located between the purely syntactical level and the semantical level will be discussed in more detail below. But let us note, that our mode of presentation is clearly different from any form of reference in model-theoretic terms.

Let us now turn to the more prominent example given by Frege. By "morning star", Venus is presented as the star[8] visible in the morning, by "evening star" as visible in the evening. Thus, the sense of "morning star" differs from that of "evening star", although both refer to the same object. We may use "the star visible in the morning" and "the star visible in the evening" as the expressions which give us the mode of

[7] This view is even better illustrated in the example Frege gives in his *Begriffsschrift* in the paragraph on *identity of content*, [4, p. 20f]. This paragraph could be used as further support of our account here, yet, Frege, by the time of the *Begriffsschrift*, didn't bring forward the notions of *sense* or *mode of presentation*.

[8] In the discussion of this example, "star" is, of course, to be understood as a folk term.

presentation, using the same argument as above: these expressions may serve as terms t defining a constant A ("morning star", "evening star", or even "Venus").[9]

Thus, we may extend our working definition of *mode of presentation* given above for mathematical terms to terms in general, saying that a term t may express a mode of presentation if it can be used as *definiens* in a clause like "Let A be t." Later we shall see how proofs enter.

4 Morning Star Versus Evening Star Revisited

Frege's example of difference of senses in "morning star" and "evening star" became a classic. It is intuitively clear that there are two different senses, although there is only one reference.

Possible worlds semantics does not cope well with this example. Taking Kripke's [9] famous distinction of rigid and non-rigid (use of) terms into account, one can consider "morning star" and "evening star" as definite descriptions[10] which should be non-rigid. But "Venus", as a proper name, is supposed to be rigid. Now, however, in the worlds in which "morning star" and "evening star" are supposed to be different, we would have "two copies" of Venus, let's call them Venus_M and Venus_E. Leaving aside the question which of them should be *the* Venus, the problem is that for these two Veneres the astronomical laws have to fail—otherwise they would coincide again.[11] Is it really the case that—to understand the difference of the sense of morning star and evening star—we would have to consider worlds with different astronomical laws? In our view, the difference in the *sense* of morning star and evening star should not depend on the astronomical laws at all—it depends, to go back to Frege, only in the mode of their presentation.

In our account, we would take (appropriate) terms t_M and t_E representing "morning star" and "evening star" in a sufficiently formalized astronomical theory as definite descriptions which both could serve as defining a planet. It is now a new task to *prove* the equality $t_M = t_E$ by use of astronomical laws (together with the empirical astronomical observations which are formalized as statements involving t_M and t_E). We may say that the fact that the denotations of t_M and t_E are equal follows from

[9] Of course, Venus should be *defined* by only one of these expressions—unless it is already known that they coincide (though, it would look quite odd to give two different *definitions* of one and the same object).

[10] Kripke treats the terms "Phosphorus" and "Hespherus" as proper names; we take here "morning star" and "evening star" as elliptic definite descriptions extendable to "the brightest non-lunar object in the morning/evening sky". For more on rigid designators, see [10].

[11] Consider an alternative world where the astronomical observations of Venus_M coincide with the observations of Venus in the real world. If this alternative world have the same astronomical laws as the real world, Venus_M *has to* appear at the same position as Venus in our world, i.e., at the place of Venus_E, i.e., Venus_M and Venus_E have to be identical.

the identity of $(t_M)^{\mathfrak{M}} \equiv (t_E)^{\mathfrak{M}} \equiv \mathfrak{Venus}$ in the real world, while the equality of the modes of description $t_M = t_E$ follows from the proof in our astronomical theory. The need of *performing* this proof explains the epistemic difference between identities and equalities.[12]

5 Equality

We here consider only equalities between terms, which may refer to mathematical objects or to objects of our real world.

As said, in first-order logic, equality is axiomatized as a universal congruence relation, thus directly linked to *extensionality* (the congruence axioms include the compatibility with all functions and relations).

Working in an epistemic context, however, one may note that not all (true) equalities might be known by an agent[13] $\mathscr{A}$. Thus, the equalities known by $\mathscr{A}$ may not be complete with respect to the identities which hold in the intended model of $\mathscr{A}$'s knowledge. This incompleteness has to be understood with respect to the combination of interpretation and identity as described in Sect. 2: for two terms t and s, $(t)^{\mathfrak{M}} \equiv (s)^{\mathfrak{M}}$ may hold, but $\mathscr{A}$ doesn't know $t = s$.

The incompleteness can arise from two different sources. On the one hand, an agent may have an "underaxiomatized" representation of the world. On the other hand, agents are not supposed to be logically omniscient, and will miss (fail to know) those equations which they haven't yet proved.

The first case may apply in the morning star/evening star example, when the agent does not know the astronomical laws to derive the fact that both terms refer to the same object.[14]

The second case may apply to the geometric example, if the agent didn't perform the mathematical proof of the equality of the two intersections.

In both cases, the equalities the agent knows are *incomplete* with respect to the identities which hold in the appropriate model. Now, the equality relation $=$ of $\mathscr{A}$ (considered as the set of equalities known by $\mathscr{A}$) may serve to express some *intensionality with respect to the outer extensionality*, given by $\equiv$ (or all true equalities).

[12]We leave here aside the fact that essentially nobody actually performs this proof, but learns the equality $t_M = t_E$ in school and, thus, adds it somehow as an axiom to the belief set. But as it should be with everything we learn in school, it should be possible, in principle, to replace our "learned axioms" by actual proofs, if we would study the respective topic in sufficient detail.

[13]The term *agent* is heavily burdened by its use in Artificial Intelligence. However, because of a lack of alternatives, we use "agent" here in the way as it became recently fashionable in philosophy to designate "something having knowledge".

[14]To satisfy our remark of footnote 12 we may stipulate that this agent also didn't learn this equality in school or elsewhere.

If we analyze $\mathscr{A}$'s knowledge we should allow the substitution of two terms only if $\mathscr{A}$'s knowledge comprises the corresponding equality—the underlying identity in the model is irrelevant. With only these identities in mind, we may observe the intensional phenomena in $\mathscr{A}$'s knowledge.

6 Equality of Senses

One of the fundamental challenges for every theory of senses is the notion of equality of senses.

In our setting the notion of sense is naturally relativized to (the knowledge of) an agent $\mathscr{A}$. A very naive attempt would be to introduce a notion of equality of senses relativized to an agent $\mathscr{A}$, identifying the sense expressed by two terms if and only if $\mathscr{A}$ can prove the equality $t = s$. This would allow to separate the denotation from the senses of two terms denoting the same object in cases where $\mathscr{A}$ does not have the proof of the corresponding equality at hand. But it would compromise Frege's original idea, as the senses of "morning star" and "evening star" should clearly stay different even if somebody knows that both denote Venus.

Still, we may obtain an interesting notion of equality of senses if we allow for the closure of the mode of presentation under *some* equalities. This can be illustrated best by use of the geometric example: we said that $Intsec(a, b)$ and $Intsec(b, c)$ should be considered as different modes of presentation of the point $\mathfrak{p}$. It seems to be, however, that $Intsec(a, b)$ and $Intsec(b, a)$ do not give us different modes of presentation of the same point. In technical terms, this means that the mode of presentation is not changed when we invoke the symmetry of the relation *Intsec*.

It is not our aim to specify concrete criteria concerning which (type of) equations should be taken into account for the equality of senses. In contrast, we think that equality of senses should not only be relativized to an agent (or an agent's knowledge) but that it could also be graduated and that it depends on the chosen axiomatic context.

The rôle of the axiomatic context can be exemplified by the natural numbers: if they are introduced as a commutative semigroup, commutativity is, of course, "build in" and $t + s$ should not have a sense different from that of $s + t$. If, however, the natural numbers are introduced by use of the Peano Axioms, the commutativity of addition requires a rather non-trivial proof by induction, and, in this context, one might say that the sense of $t + s$ differs from the one of $s + t$, as the required recursion over (only) one of the summands to calculate the value may lead to substantially different computations.

This last example shows that, for our notion of mode of presentation, the underlying axiomatic setting forms an integral part of the sense of a term.[15]

[15]To elaborate this approach one could take into account, for instance, *background knowledge* as constitutive for senses. We may also invoke *definitional knowledge* obtained by *definitional reflection*, [12, Sect. 2.3.2].

7 Proof-Theoretic Semantics

So far, we gave some kind of answer to what we called Frege's questions stressing the epistemic character of a possibly incomplete set of proven equalities of an agent, in contrast to identity in a model. We will now turn to the idea of proof-theoretic semantics.

According to ([8], p. 503),

> [p]roof-theoretic semantics [assigns] proofs or deductions an autonomous semantic role from the very onset, rather than explaining this role in terms of truth transmission. In proof-theoretic semantics, proofs are not merely treated as syntactic objects […], but as entities in terms of which meaning and logical consequence can be explained.

This approach is already quite successfully pursued for the usual logical operations (see [7, 12] and this volume). It is our aim to extend it to some further concepts, like equalities here or necessity in [6].

In the case of *equality* a proof-theoretic semantics requires that, from the very onset, one would have to dispense with any (model-theoretic) notion of identity. From a technical point of view, one could say that the proof-theoretic semantics of the *equality relation* is given by the axioms involving this relation. But what would be the proof-theoretic semantics of a particular *equation*? The terms in such an equation have now, where any model-theoretic interpretation is gone, of course, an autonomous status.

From a proof-theoretic perspective, "morning star" and "evening star" should, of course, be different. Their *mode of presentation* is given by the way the axioms introduce them as terms. This includes implicitly the full axiomatic framework which now makes part of the mode of presentation.

Whatever the concrete axioms might be, they should state that the "morning star" is visible (on some days) in the morning, and the "evening star" in the evening, respectively. As discussed above, the equality between them needs a proof. For the *proof-theoretic semantics* of the terms it should not even be relevant whether such a proof is performed or not—its sheer need gives rise to consider the proof-theoretic semantics as different for the two terms, determined only by the axioms governing them. Only in the case of "immediate" ("trivial" or maybe "elementary") equalities—like in the case of the symmetry of *Intsec*—a term might be manipulated without changing its sense.

As related approaches we would like to mention here Tichý's *Transparent Intensional Logic* (TIL), [1, 13] and Moschovakis's *Sense and Denotation as Algorithm and Value*, [11].

TIL does not dispense with possible worlds, but assigns them a secondary rôle in the analysis of senses. These are introduced as abstract procedures, called *constructions*, which are applied to *an object*, in dependence of a possible world, to decide whether this object (for instance, Venus) fulfills the intension (e.g., being the morning star). With respect to our approach one can ask whether, and if so, in which way the abstract procedures can be related to the proofs we take as a basis.

Such a relation would be given, at least partly, by the Curry–Howard correspondence for Moschovakis's approach. He introduces senses as algorithms which compute (denotational) values. Based on the well-known correspondence of algorithms and proofs, we could adapt Moschovakis's slogan by describing our (broader) approach to intensionality as *Sense and Denotation as Proof and Truth*. Conversely, Moschovakis's account could also be dubbed a *recursion-theoretic semantics*.

Acknowledgments Research supported by the Portuguese Science Foundation, FCT, through the projects *Hilbert's Legacy in the Philosophy of Mathematics*, PTDC/FIL-FCI/109991/2009; *The Notion of Mathematical Proof*, PTDC/MHC-FIL/5363/2012; *Hilbert's 24th Problem*, PTDC/MHC-FIL/2583/2014, and UID/MAT/00297/2013.

References

1. Duží, M., Jespersen, B., Materna, P.: Procedural Semantics for Hyperintensional Logic. Logic, Epistemology, and the Unity of Science, vol. 17. Springer, Berlin (2010)
2. Frege, G.: Begriffsschrift, Nebert, Halle a.d.s. (English translation [4])
3. Frege, G.: Über Sinn und Bedeutung. Zeitschrift für Philosophie und philosophische Kritik, NF 100 pp. 25-50 (1892) (English translation [5])
4. Frege, G.: Concept script. In: van Heijenoort, J. (ed.) From Frege to Gödel, Harvard University Press, Cambridge (1967) (English translation of [2])
5. Frege, G.: On sense and meaning. In: McGuiness, B. (ed.) Collected Papers on Mathematics, Logic, and Philosophy, p. 157-177. Basil Blackwell (1974) (English translation of [3] by M. Black)
6. Kahle, R.: A proof-theoretic view of necessity. Synthese **148**(3), 659–673 (2006)
7. Kahle, R., Schroeder-Heister, P. (eds.) Proof-Theoretic Semantics. Springer (2006). Special issue of Synthese 148(3), 503-743
8. Kahle, R., Schroeder-Heister, P.: Proof-theoretic semantics-introduction. Synthese 148(3), 503-506 (2006)
9. Kripke, S.: Naming and Necessity. Harvard University Press, Cambridge (1980)
10. LaPorte, J.: Rigid Designators. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (Summer 2011 edn.) (2011), http://plato.stanford.edu/archives/sum2011/entries/rigid-designators/
11. Moschovakis, Y.: Sense and denotation as algorithm and value. In: Väänänen, J., Oikkonen, J. (eds.) Logic Colloquium '90, pp. 210-249. Lecture Notes in Logic, vol. 2, Association for Symbolic Logic (1994)
12. Schroeder-Heister, P.: Proof-theoretic semantics. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (Winter 2012 edn.) (2012), http://plato.standard.edu/archives/win2012/enties/proof-theoretic-semantic/
13. Tichý, P.: The Foundations of Frege's Logic. De Gruyter, Berlin (1988)
14. Wehmeier, K.F.: How to live without identity–and why. Australas. J. Philos. **90**(4), 761–777 (2012)

On the Proof-Theoretic Foundations of Set Theory

Lars Hallnäs

Abstract In this paper we discuss a proof-theoretic foundation of set theory that focusses on set definitions in an open type free framework. The idea to make Cantor's informal definition of the notion of a set more precise by saying that any given property defines a set seems to be in conflict with ordinary modes of reasoning. There is to some extent a confusion here between *extensional* perspectives (sets as collections of objects) and *intensional* perspectives (set theoretic definitions) that the central paradoxes build on. The solutions offered by Zermelo-Fraenkel set theories, von Neumann-Bernays set-class theories and type theories follow the strategy of retirement behind more or less safe boundaries. What if we revisit the original idea without making strong assumptions on closure properties of the theoretical notion of a set? That is, take the basic definitions for what they are without confusing the borders between intensional and extensional perspectives.

Keywords Set theory · Foundations · Proof theory · Definitional reflection · Partial inductive definitions · Functional closure

1 Introduction

Foundations of set theory relates to answers of the following two main questions:

(A) What is a set?
(B) What does it mean to reason with sets?

With respect to (A) Cantor's informal definition of the notion of a set seems perfectly intuitive.

> By an "aggregate" (*Menge*) we understand any collection into a whole (*Zusammenfassung zu einem Ganzen*) M of definite and separate objects *m* of our intuition or our thought. [2, p. 85]

L. Hallnäs (✉)
University of Borås, Borås, Sweden
e-mail: Lars.Hallnas@hb.se

T. Piecha and P. Schroeder-Heister (eds.), *Advances in Proof-Theoretic Semantics*,
Trends in Logic 43, DOI 10.1007/978-3-319-22686-6_10

It is natural to think of *collection into a whole* as an *act of abstraction*. The question is how to understand this. In view of the paradoxes by Russell and others, the idea to make this more precise by saying that any given property defines a set seemed to be in conflict with intended *natural* modes of reasoning. What was wrong with this idea?

It might be an issue of confusing extensional and intensional perspectives. The idea of a set as a gathering of given objects into a whole paints a picture of sets as collections $(a, b, \ldots)$. We have given objects and we collect them into a whole by so to speak bracketing them. This extensional view of sets has a clear expression in the cumulative hierarchy. Abstracting with respect to a given property introduces a more intensional perspective, i.e., the way in which we actually define a set with the intention to capture a collection of objects.

> Russell's antinomy came as a veritable shock to those few thinkers who occupied themselves with foundational problems at the turn of the century. [4, p. 2]

There is something strange about this reaction. Why do we expect that such a, very general, more intensional characterisation will capture just sets as collections of objects in an intuitive extensional sense, i.e., as bracketing a given collection of objects? There is no reason to think that these two notions and perspectives should coincide, i.e., that the intensional characterisation would produce just *nice* sets, namely collections of given objects. It is in this respect of interest to note that the definition, i.e., the defining property $x \notin x$ of the Russell set R is a very elementary one. Its proof-theoretic behaviour can, for example, be observed already in intuitionistic propositional logic [3].

So if we accept the idea of abstraction with respect to any given defining property, i.e., full comprehension, as a foundation for set theory, we have an answer to question (A), that is, what a set is. But how should we then understand the paradoxes? The Russell paradox for instance seems to show that something is wrong with respect to question (B). The paradoxical argument builds on several basic assumptions, where one of the most important ones is the assumption that '*R is a set*' is a well-defined notion with respect to intended intuitive logical reasoning, which is a very strong assumption with respect to the given definition. So this is one way to view Russell's paradox; too strong assumptions on basic theoretical notions.

The solutions offered by Zermelo-Fraenkel set theories, von Neumann-Bernays set-class theories and type theories follow the strategy of retirement behind more or less safe boundaries (see [4]). There are several ideas about proof-theoretically founded restrictions on the comprehension scheme [5], [9]. Compare further the set theory of Fitch (see [4], [9]), the notion of a Frege structure [1] and notions of structural rules in relation to paradoxes [14].

Now what if we revisit the original idea without making strong assumptions on closure properties of the theoretical notion of a set? That is, take the basic definitions for what they are without confounding intensional and extensional perspectives.

2 Defining Sets

If we think of set definitions as abstractions λX, saying that a property, or functional expression, X defines a set, we may derive the following definitions of membership and equality for sets:

- $A \in \lambda X$ iff $X(A)$,
- $A = B$ iff $(x \in A \iff x \in B)$ for all sets x
  (i.e., $(A = \lambda X \ \& \ B = \lambda Y \implies \lambda X = \lambda Y) \iff (X(x) \iff Y(x))$ for all sets x).

In the same manner the axiomatic approach, ZF and other similar set theories, introduce axioms stating the existence of sets for certain specific *safe* defining properties, such as for example the subset property

$$x \in P(A) \text{ iff } x \text{ is a subset of } A$$

but also other types of axioms such as axioms introducing measurable cardinals and other large cardinals.

Although the axioms of power set and replacement, together with axioms of infinity (large cardinals starting with $\aleph_0$), provide for strong means to build sets following the cumulative hierarchy intuition of the universe of sets, they still represent a theory marked by withdrawal from foundational disasters to more favourable positions. It is not only matters of a first order formalization of *safe* axioms, but also from a more general intensional perspective a lack of elementary foundational principles. There is a very elementary and suggestive extensional picture through the cumulative hierarchy, but this is lacking with respect to definitional issues.

Why is $(x = x)$, for example, not an admissible set defining condition?

1. It contradicts the idea of sets as collections of given objects, i.e., $\lambda(x = x)$ is a member of $\lambda(x = x)$.
2. We cannot comprehend the given objects we are supposed to collect into a whole by abstraction.

In both cases we say that $(x = x)$ does not define a *set* in the sense of a *total* object that behaves nicely with respect to the intended reading of logical constants and the notion of membership. But this does not really answer the question. It just says that whatever $\lambda(x = x)$ may define it is not a set in the *extensional* sense as a collection of given objects.

The problem here is an example of what we in many cases meet as we try to define a notion where it is difficult to map out the exact borders by elementary means, the notion of a total computable function being a canonical example. From a foundational and theoretical point of view it would be nice if it were possible to make sense in some way of the initial, and very elementary, ideas of Frege and others [4].

Let us look at a very naïve and simplistic attempt to *define* sets based on the idea of sets as introduced by abstraction of defining properties. In defining sets this way it is natural to make a distinction between set expressions, i.e., sets, terms etc., and propositional expressions, i.e., propositions, formulas etc. But if we accept more open definitions this does not seem necessary, and for reasons of simplicity we will just make a distinction between sets (A) and set theoretical reasoning (B) in what follows. This would also be in line with reading *Ockham's razor* as saying that basic classifications and distinctions are matters of proofs and not foundational definitions. The definition of sets is:

- T and F are sets,
- $A \to B$, $A \in B$, $A = B$ are sets if A and B are sets,
- $S(f)$, $\exists(f)$ and $\forall(f)$ are sets if the world of sets is closed under the given function f.

This would answer question (A). To answer question (B) we add the following *derived* definition:

- T is true,
- $A \to B$ is true if (A is true $\implies$ B is true),
- $A \in S(f)$ is true if $f(A)$ is true,
- $A = B$ is true if ($x \in A$ is true $\iff$ $x \in B$ is true) for all sets x,
- $\exists(f)$ is true if $f(x)$ is true for some set x,
- $\forall(f)$ is true if $f(x)$ is true for all sets x.

Russell's paradox tells us of course directly that there are no such definitions satisfying the intended closure properties we have written down above. But from an intensional, i.e., definitional, point of view, we actually intend to define something by writing down these clauses. The question is just what that is, in what ways we can interpret these acts of defining?

3 Functional Closure, Local Logic and the Notion of Absoluteness

There are three major issues to observe in the *definitions* given above in Sect. 2:

1. We introduce functional constructions, $S(f)$, $\exists(f)$ and $\forall(f)$, by a defining condition asking for the notion we define to be closed under a given function.
2. We introduce a conditional construction $A \to B$ by a defining condition asking B to follow from A.
3. What we actually state in the 'definitions' are closure conditions for notions we hope to be able to define in one way or another.

3.1 The Functional Closure

The idea of *function closure* (in the realm of monotone inductive definitions) is that we have some things $a, b, \ldots$ given and also some functions $f, g, \ldots$. We then define a notion X by saying that

- $a, b, \ldots$ is an X,
- if x is an X, then $f(x), g(x), \ldots$ is an X.

Implicitly this means that X is defined by these clauses and nothing else. From an intensional and foundational point of view in 'generating' X the things that $f, g, \ldots$ act on in X are not given, besides the initial things $a, b, \ldots$, they are introduced as we build X. Once defined, X is then the smallest collection of things including $a, b, \ldots$ and being closed under $f, g, \ldots$.

Similarly the idea of a functional closure is that we have some things $a, b, \ldots$ and functions $f, g, \ldots$ given and also functionals $F, G, \ldots$. Analogously, from an intensional and foundational point of view, the functions 'in' X that $F, G, \ldots$ act on, i.e., functions that X is closed under, are not given, but introduced as we build X. In both cases we take for granted certain things as primitive notions. In the first case some given objects and functions and in the second case some given objects (not necessary in all cases), some functions and functionals. In both cases what we rely on is, so to speak, inscribed in fundamental circles of reasoning. The objects we generate in building up the function closure are of course given in an abstract manner of speaking. The same thing holds for the functions we generate in building up the functional closure:

- $a, b, \ldots$ is an X,
- if x is an X, then $f(x), g(x), \ldots$ is an X,
- if X is closed under f, then $F(f), G(f), \ldots$ is an X.

In *non-foundational* and mathematically precise definitions we assume there is given a universe of objects, a function space and some functions and functionals defined on this universe/function space.

3.2 Local Logic

When defining $A \rightarrow B$ *is true* in terms of *if* A *is true, then* B *is true* it is really an issue what we mean by *if* A *is true, then* B *is true* as a defining condition. A reasonable interpretation of this is that what we mean to say is that B follows from A on the basis of information provided by the given definition, i.e., that we can prove B to follow from A in the local logic that the given definition implicitly defines. With respect to set theory this means that the sets we introduce, or to be more precise the set definitions we introduce, open up for reasoning relative to a *local set theoretic context*. What this could mean will be explained below.

3.3 Absoluteness

It is one thing to use *if A is true, then B is true* as a defining condition in a definition and quite another thing to state *if A is true, then B is true* as a closure condition for a given definition. In view of an analogy between models of set theoretical axioms and definitions of set theoretical concepts we might introduce the notion of *absoluteness* (cf. [8]) also in this definitional context. Whereas in the first case we compare how a set theoretical notion (formula) behaves in a model in relation to its behavior in another model, which intuitively means outside the model if the second model is the true cumulative hierarchy V, in the latter case we compare how a definitional notion/condition behaves inside the definition, in the local logic of the definition, with how it behaves outside the definition in the world of *intended* interpretation of defining conditions.

A set theory S is a pair of definitions $S\Phi$ and $TS\Phi$, following the ideas discussed above in Sect. 2, for a given collection of functions Φ. A defining condition A is *(left) absolute* (with respect to S), if for all defining conditions B

$$B \text{ follows from } A \text{ in } TS\Phi \text{ iff } (A \text{ is true in } TS\Phi \implies B \text{ is true in } TS\Phi).$$

What this means is that deriving something from A in $TS\Phi$ is the same as implication. One closure condition that is generally self-evident is the following one

$$a \text{ is true by definition } D \text{ iff there is a defining condition } A \text{ in } D \text{ of a true by } D.$$

This is the basic axiom of definitional theory.

Take the Russell set $S(\lambda(x \in x \to F))$ (let us call it r) and let R be a set theory that includes this set. The set r is not (left) absolute in R. $r \to F$ is true in R, that is F follows from r in R. But whereas r is true in R, F is obviously not since it is not even defined in R. The argument follows from the basic definitional axiom together with an assumption that the local logic of the definition has a reasonable behavior with respect to the intended interpretation of involved logical constants. This argument demonstrates that negation is not an absolute notion, which from a proof-theoretic point of view would be a reasonable way to interpret the Russell paradox, i.e., falsity is an absolute notion, while negation is not.

This notion of absoluteness can further be specialized as follows:
A defining condition A is

1. *(right) absolute* (with respect to S) if

$$A \text{ follows from } B \text{ in } TS\Phi \iff (B \text{ is true in } TS\Phi \implies A \text{ is true in } TS\Phi),$$

2. *upward absolute* if

$$B \text{ follows from } A \text{ in } TS\Phi \implies (A \text{ is true in } TS\Phi \implies B \text{ is true in } TS\Phi),$$

3. *downward absolute* if

$$A \text{ is true in } TS\Phi \implies (B \text{ is true in } TS\Phi \implies B \text{ follows from } A \text{ in } TS\Phi),$$

4. etc.

To say that a defining condition, or a set, is (left/right) absolute means that the condition, or set, with respect to local reasoning has the same meaning inside the local logic as outside it.

4 A Proof-Theoretic Interpretation

Even if we note that there are no definitions having the closure properties stated in Sect. 2 above, there is still the possibility to read these *definitions* from a more strict intensional point of view. We then look at the closure conditions as clauses in two *partial inductive definitions* ([6, 7, 13]). The idea is basically to look at *if* ..., *then* ... and *is closed under* in terms of the notion of logical consequence that defines the local logic of the definitions in question, i.e., that *if* A, *then* B is read as B *follows from* A *by the given definition*.

As a mathematical object a *(partial inductive) definition* D consists of a collection of equations

$$a = A$$

for $a \in U$ for some given universe of discourse and where A is a defining condition built up from elements in U, $\top$ and $\bot$ using constructions $\bigwedge_I$ and $\Rightarrow$. Let $D(a)$ be the collection of conditions defining a in D if there are any and $\{\bot\}$ otherwise. The *local logic* of D, $\vdash_D$, is then given by the following elementary (monotone) inductive definition

$$\Gamma, a \vdash_D a$$

$$\Gamma \vdash_D \top \qquad\qquad \Gamma, \bot \vdash_D C$$

$$\frac{\Gamma \vdash_D A_i \qquad (i \in I)}{\Gamma \vdash_D \bigwedge_I A_i} \qquad \frac{\Gamma, A_i \vdash_D C}{\Gamma, \bigwedge_I A_i \vdash_D C}\,(i \in I)$$

$$\frac{\Gamma, A \vdash_D B}{\Gamma \vdash_D A \Rightarrow B} \qquad \frac{\Gamma \vdash_D A \qquad \Gamma, B \vdash C}{\Gamma, A \Rightarrow B \vdash_D C}$$

$$\frac{\Gamma \vdash_D A}{\Gamma \vdash_D a}\,(A \in D(a)) \qquad \frac{\Gamma, A \vdash_D C \qquad (A \in D(a))}{\Gamma, a \vdash_D C}$$

The function closure with respect to $X \subset U$ and functions $f_1 \dots f_n$ with arities $k_1 \dots k_n$ over U, is then formally defined by the following definition

$$\begin{aligned} a &= \top \quad (a \in X) \\ f_i(x_1 \dots x_{k_i}) &= (x_1 \dots x_{k_i}) \quad (i \leq n) \end{aligned}$$

$Def(D(X, f_1 \dots f_n))$ is then the smallest set containing X and being closed under the functions $f_1 \dots f_n$.

Similarly the *functional closure* with respect to $X \subset U$, functions $f_1 \dots f_n$ with arities $k_1 \dots k_n$ over U, a functional $F : [U \to U] \to U$ and a set $\Phi \subset [U \to U]$, is given by a definition $D(X, f_1 \dots f_n, F, \Phi)$:

$$\begin{aligned} a &= \top \quad (a \in X) \\ f_i(x_1 \dots x_{k_i}) &= (x_1 \dots x_{k_i}) \quad (i \leq n) \\ F(f) &= \textstyle\bigwedge_U (x \Rightarrow f(x)) \quad (f \in \Phi) \end{aligned}$$

Now we might rewrite the definitions $S\Phi$ and $TS\Phi$ in the following way:

$$S\Phi \begin{cases} T = \top \\ F = \top \\ A \to B = \bigwedge(A, B) \\ A \in B = \bigwedge(A, B) \\ A = B = \bigwedge(A, B) \\ S(f) = \bigwedge\limits_{S\Phi}(x \Rightarrow f(x)) \\ \exists(f) = \bigwedge\limits_{S\Phi}(x \Rightarrow f(x)) \\ \forall(f) = \bigwedge\limits_{S\Phi}(x \Rightarrow f(x)) \end{cases}$$

$$TS\Phi \begin{cases} T = \top \\ A \to B = A \Rightarrow B \\ A \in S(f) = f(A) \\ A = B = \bigwedge\limits_{S\Phi}((x \in A \Rightarrow x \in B), (x \in B \Rightarrow x \in A)) \\ \exists(f) = f(x) \quad (S\Phi) \\ \forall(f) = \bigwedge\limits_{S\Phi}(f(x)) \end{cases}$$

Reading them as foundational definitions we have to accept certain notions as primitive notions; the conditions T and F, the function $\rightarrow$, the functionals S, $\exists$ and $\forall$, the notion of a function and indexing families over the sets we define. In principle what amounts to understanding the functional closure as a primitive foundational notion. The resulting *formal* systems, defining the local logics of the definitions, are consequently formal systems in an informal sense. They define what a proof is as a foundational notion, providing a proof-theoretic foundation of set theory, that is, using proof-theoretical notions in an abstract and open manner (cf. the notion of a *general proof theory* in [10–12]).

5 Sets

From an extensional perspective viewing sets as collections of given sets, the notion of an elementary set connects to hierarchies of what we somehow can visualize, i.e., low levels of the cumulative hierarchy. From an intensional point of view, where the act of abstraction with respect to a given defining property/function is in focus, a natural notion of an elementary set must build on characteristics of the definition. The Levy hierarchy [8] of course shows strong connections between both perspectives for ZF, but the situation here is a bit different as we look at set definitions in much more open set theories. It is for instance clear that a set such as $S(\lambda(x = x))$ is a very elementary set with respect to its defining function.

Let us say that a set

- $S(f)$ is a Φ-*set* if $S\Phi$ is closed under f, i.e., that $f(x)$ follows from x in $S\Phi$ for all sets x in $S\Phi$,
- $S(f)$ is *elementary* if it is a Φ-set for all Φ.

Both $S(\lambda(x = x))$ and $S(\lambda(x \notin x))$ are elementary sets. A simple example of a non-elementary set is $S(\lambda(x = S(\lambda(y = a))))$.

6 Foundational Issues

It is clear that consistency is not an explicit issue in the present context. Falsity (i.e., F) is by definition something that is not defined and can thus never be proved in a set theory $S\Phi$. But consistency of course relates to issues of cut elimination for sequent calculi, which relates to upward absoluteness. So assume we have a set theory S where all basic defining conditions are absolute, or at least upward absolute. From the point of view of set theoretic reasoning the sets definable in these theories are somehow 'nice' sets.

Stating that there are functions Φ with certain properties is what here corresponds to axioms of set theory, and proving or believing that the theory $S\Phi$ is absolute in some sense corresponds to defining a model for the axioms. But while the true

cumulative hierarchy V is the universe in which these models live, a general set theory $S\Omega$ with no restrictions on functions is the context in which these definitional theories $S\Phi$ live. The big difference is that V is an extensional context, i.e., the true world of pure sets, whereas $S\Omega$ is an intensional context based on a very general notion of set definitions not presupposing a rationale of welldefinedness. What set theories $S\Phi$ reflect is not inner models, but the locality of proof logics.

The idea of reduction is somehow inherent in the notion of foundations, i.e., that we build on elementary foundations. Although we evidently just walk around in ontological circles, this idea of reduction is not meaningless. A very clear and conceptually elementary model provides a reduction in the sense that we see clearly why given axioms make sense. The argument that the idea of a reduction is an illusion since the construction of the model involves all the power of the axioms themselves does not make for a strong case. It is the suggestive simplicity and clearness of the picture the model paints that is important, i.e., that we really can *see* the construction. Simplicity with respect to definitional principles builds another type of foundations; the local logic of given definitions. The foundational construction here is the functional closure interpreted as a partial inductive definition. What is important is then that we can 'see' the proofs that build the sets and the set theoretical arguments in a very elementary sense. A typical example making the difference clear is the power set $P(A) = S(PA)$ where PAx is $\forall z(z \in x \rightarrow z \in A)$. To envision $P(A)$ as a collection of given objects involves very abstract acts of visualising for large sets A. Can we *see* the set, can we trust the axiom? It is of course clear that $S(PA)$ as a set theoretical definition opens up for logical complexity in reasoning, but in this case it is a matter of visualising proofs with respect to a given definition. Can we *see* the proofs, can we trust the definition?

The definition itself is in some sense elementary, the proofs defining reasoning in theories $S\Phi$ are also elementary in some sense. Thus there is a reduction in foundations in some sense. But in actual set theoretical practice we need to trust certain closure conditions on the definitions allowing for nice forms of reasoning for what we believe to be nice theories $S\Phi$. The major challenge here is to develop set theory within the framework of theories $S\Phi$ and explore the meaning of classical set theoretical issues in this context.

Since $S\Omega$ is closed, in the sense that each definable function f is reflected in a set $S(f)$, we have the following

Theorem *V (modulo large cardinals beyond $\aleph_0$) has a definable reflection in $S\Omega$.*

References

1. Aczel, P.: Frege structures and the notions of proposition, truth and set. In: J. Barwise, H.J. Keisler, K. Kunen (eds.) The Kleene Symposium. North-Holland (1980)
2. Cantor, G.: Contributions to the Founding of the Theory of Transfinite Numbers. Dover Publications (1955)
3. Ekman, J.: Normal proofs in set theory. Ph.D. thesis, Department of Computing Science, Chalmers University of Technology (1994)
4. Fraenkel, A.A., Bar-Hillel, Y., Levy, A.: Foundations of Set Theory. North-Holland, Amsterdam (1973)
5. Hallnäs, L.: On normalization of proofs in set theory. Ph.D. Thesis, Dissertationes Mathematicae CCLXI, Warszawa (1988)
6. Hallnäs, L.: Partial inductive definitions. Theor. Comput. Sci. **87**, 115–142 (1991)
7. Hallnäs, L., Schroeder-Heister, P.: A proof-theoretic characterization of logic programming II: programs as definitions. J. Log. Comput. **1**(5), 635–660 (1991)
8. Levy, A.: A hierarchy of formulas in set theory. In: Memoirs of the American Mathematical Society, vol. 57. American Mathematical Society, Providence (1965)
9. Prawitz, D.: Natural Deduction. A Proof-Theoretical Study. Almqvist & Wiksell, Stockholm (1965)
10. Prawitz, D.: Ideas and results in proof theory. In: Fenstad, J.E. (ed.) Proceedings of the Second Scandinavian Logic Symposium. North-Holland (1971)
11. Prawitz, D.: Towards a general proof theory. In: Suppes, P. (ed.) Logic. Methodology and the Philosophy of Science IV. North-Holland, Amsterdam (1973)
12. Prawitz, D.: On the idea of a general proof theory. Synthese **27**, 63–77 (1974)
13. Schroeder-Heister, P.: Rules of definitional reflection. In: Proceedings of the 8th Annual IEEE Symposium on Logic in Computer Science. Los Alamitos, Montreal (1993)
14. Schroeder-Heister, P.: Paradoxes and structural rules. In: Novaes, C.D., Hjortland, O.T. (eds.) Insolubles and Consequences: Essays in honour of Stephen Read. College Publications, London (2012)

A Strongly Differing Opinion on Proof-Theoretic Semantics?

Wilfrid Hodges

Abstract Responding to an invitation from Peter Schroeder-Heister, the paper reacts to some criticisms of 'model theory' voiced among proof theorists interested in proof-theoretic semantics. It argues that the criticisms are poorly targeted: they conflate model theory with model-theoretic semantics and with the model-theoretic definition of logical consequence, which are three largely unrelated areas of study. On defining the meanings of logical constants, and of natural language expressions in general, the paper lays out some methodological requirements that any satisfactory definitions would need to meet, for example about generalisability from one context of use to other contexts. On defining logical consequence, the paper argues that some points made recently by Schroeder-Heister and Kosta Došen are largely sound and probably uncontroversial if clearly stated, but their impact is blurred by some question-begging formulations.

Keywords Proof-theoretic semantics · Model-theoretic semantics · Definition of logical consequence · Tarski

It was very kind of Peter Schroeder-Heister to invite me to contribute to this meaty conference. He said:

> . . . you would fit very well into this meeting, even though (or perhaps because) you have opinions that strongly differ from [those] of the majority of people at the conference. Perhaps you can give a talk in defence of model theory, as far as the foundations of logic are concerned. (1)

That's a fantastic invitation, and I went to the meeting resolved to disagree with as many people as possible.

In the event it was not so easy. Partly there was serious research being done in proof theory, and I am not a proof theorist. Partly there were a good number of entirely sensible and friendly people. But also I often found it hard to see what the issues were. I think this was not entirely my fault. Straw men were being set up and

W. Hodges (✉)
Okehampton, Devon, England, UK
e-mail: wilfrid.hodges@btinternet.com

T. Piecha and P. Schroeder-Heister (eds.), *Advances in Proof-Theoretic Semantics*,
Trends in Logic 43, DOI 10.1007/978-3-319-22686-6_11

knocked down. I could see this most clearly when the straw men were described as model theorists, because I do know something about model theory, and some of the views being attributed to model theorists were not ones I recognised. This impression was strengthened when I read a recent paper of Peter's in *Synthese* [14].

So I had plenty to disagree with, but not in a very satisfactory way. It's more edifying to discuss substantive issues than to clear away misunderstandings. But the clearance work has to be done first. I will try to keep it both brief and profitable.

I thank Peter Schroeder-Heister and Kosta Došen for some valuable discussions.

1 Straw Model Theory

A good place to start will be an elegant paper of Dag Prawitz [11] from 1974. There is a lot that I agree with in the paper, but I was pulled up sharp when he said:

> In model theory, one concentrates on questions like what sentences are logically valid and what sentences follow logically from other sentences [11, p. 66]. (2)

I can say with absolute confidence that I never met a model theorist who 'concentrates on questions like what sentences are logically valid and what sentences follow logically from other sentences'. On his next page Prawitz discusses Alfred Tarski's proposal for defining logical consequence, from his paper of 1936 [17]. So it seems likely that Prawitz reached the view stated in (2) by assuming that Tarski's 1936 paper is of interest to model theorists. This is not in fact the case. Nothing in the paper is of any interest to model theorists, except perhaps those with an interest in the prehistory of their subject.

Peter Schroeder-Heister adds another ingredient to the mix in his recent paper [14], namely model-theoretic semantics. This is a discipline concerned with describing meanings, so Peter rightly connects it with questions about how one should describe the meanings of logical constants. But its origins are quite different from those of the model-theoretic truth definition, and it belongs to a different research community. Model theorists don't do model-theoretic semantics either. I do know one person who contributes to model-theoretic semantics using techniques of model theory, namely Dag Westerståhl; but there are not many of him. In short, the three areas of research—model theory, the definition of logical consequence and model-theoretic semantics—are quite different and they have hardly anything in common beyond a connection with models in the sense associated with Alfred Tarski.

So now let me unpick the historical relations between these areas. (All the comments on Tarski below draw out material from my [10].)

There are some other research areas that connect with Tarski's notion of models but not with each other. One is mental model theory as pursued by the cognitive

scientist Ruth Byrne [2], and another is the model-theoretic syntax advocated by the linguists Geoffrey Pullum and Barbara Scholz [12].

1.1 Tarski's Definition of Logical Consequence

During the years 1929–1933 Tarski put together a definition of the concept 'ϕ is a true sentence of the language L' [16], which has become known as 'Tarski's definition of truth'. Tarski stated some very strict conditions that his definition had to meet. All symbols of the language L (apart from punctuation—we ignore this below) must be fully meaningful. The definition is written in the formalised metalanguage of L, but justified in the informal meta-metalanguage. It must use only higher-order logic, concepts expressible in the language L itself, and some syntactic notions. It must be extensionally correct: the objects satisfying it must be exactly the objects that we count intuitively as true sentences of L. The extensional correctness must be informally provable in the meta-metatheory of L. This is not the place to go into further details. The paper became well known through a German translation in 1935. It makes no reference to models, and model theorists don't cite it.

In 1935 Tarski was persuaded to attend the International Congress of Philosophers in Paris. Worrying about what he could say to impress the philosophers, he formed the idea of presenting the truth definition as a vehicle for giving formal definitions of various notions from logical metatheory, among them the notion of logical consequence. The result was a pair of papers, [17] presenting the definition of logical consequence, and [18] discussing the general idea of defining semantic notions [5, pp. 95ff].

The paper [17] on logical consequence answered a methodological question, not a question of conceptual analysis. You can't do conceptual analysis until you have a concept to analyse. But when Tarski wrote, there was no agreed concept of logical consequence to be analysed. (One should look first at what was available in the literature of his time. For example Hilbert and Ackermann [8, p. 1] have a proof-theoretic notion of *Logische Folgerung*, while Carnap [3, p. 10] speaks of one proposition being a *Grund* for another, without any clear definition. Tarski may also have factored in earlier ideas, like Bolzano's *Ableitbarkeit* and various medieval notions of *consequentia*.) Tarski makes exactly this point in his opening paragraph, noting that 'every precise definition of this concept will show arbitrary features to a greater or [lesser] degree' [19, p. 409]. In fact the term 'logical consequence' itself seems to have become common in philosophical logic only as a result of Tarski's paper.

To see what the methodological question was, we need to put the paper in context. Gödel had recently shown that there is no maximal proof calculus for pure logic of second or higher order. Ramsey [13] had discussed languages with infinite conjunctions, and both Bernays [1, pp. 86ff.] and Tarski himself [19, p. 288] had considered proof rules with infinitely many premises. So some very general questions about proof calculi were in the air, and some robust and well-motivated definitions were

needed for handling them. Tarski seems to have clarified the central question in his own mind along the following lines:

> What are the weakest constraints that we can put on a rule for deriving propositions from sets of propositions in a formal language, which make it reasonable to count any rule satisfying these constraints as an inference rule?

He proposed to label these constraints as saying that the conclusion of the rule is a 'logical consequence' of its premises.

Now for Tarski in 1935 there were two kinds of formal language. In the first kind, which we can call 'pure' languages, all symbols are logical. In the second kind, which we can call 'applied', there are also nonlogical symbols, but these symbols are all required to be fully meaningful. For pure languages, Tarski adopted just the constraint that whenever the premises are true the conclusion must be true too. This constraint looks trivial, but in Paris in 1935 it served the purpose of advertising his recent formal definition of 'true'.

For applied languages Tarski had to decide what to do about *analytical relations* between the meanings of the nonlogical constants. For example Hilbert in his Göttingen lectures around 1920 (which formed the basis of his book with Ackermann) had observed that 'Tony Blair is a parent' entails 'Tony Blair has a child' (my adaptation of Hilbert's more traditional example). Would it be appropriate to allow an inference rule that takes one from the first sentence to the second? Tarski decided no. An inference rule should be invariant under systematic changes of the meanings of the nonlogical symbols; but if we swap the meanings of 'has a child' and 'has bright red hair', then the proposed inference rule would take a true premise to a false conclusion.

It's noticeable that Tarski's own text says almost nothing about relations between the meanings of the nonlogical constants (there is a brief parenthetical remark in the middle of P. 415 in [19]), but has at least a page on the importance of the difference between (i) changing a symbol to one with a different meaning and (ii) replacing the symbol by a variable that arbitrary objects can be assigned to. That tells me that Tarski in 1935 was really more interested in fine-tuning the notion of satisfaction than in accommodating the philosophers in Paris.

The paper does use the word 'model', though not in the modern sense. The name 'model-theoretic definition of logical consequence' is not Tarski's, and I think it came into use only after the later developments that we turn to next.

1.2 Model Theory

During the 1930s and 1940s Tarski maintained a strict distinction between mathematics and metamathematics. Because of this, he was still in 1938 reluctant to accept that a set of formal axioms could serve to define the class of structures which satisfy them—as for example the class of rings consists of the structures that satisfy the

axioms of ring theory. But mathematical developments put him under pressure to change his mind. By 1950 he was ready to embrace what we now know as model theory, and he devoted the early 1950s to setting up the basics of the theory.

In the course of this work, Tarski rejigged his old truth definition, so that instead of defining 'ϕ is a true sentence of the language L' it defined 'ϕ is a sentence true in the structure M for the language L', where now L is a formal language whose nonlogical symbols have no meaning and the structure M is used to assign meanings to these symbols. This new truth definition is known as the 'model-theoretic truth definition'. You can find it in standard textbooks of model theory. But in practice model theorists mostly use just the separate recursive clauses of the definition, for example that $\bar{a}$ satisfies $\forall y \phi(\bar{x}, y)$ in M if and only if for every element b of M, $\bar{a}b$ satisfies $\phi(\bar{x}, y)$ in M. These clauses are all older than Tarski's work. The definition as a whole does guarantee that the relation 'ϕ is true in the structure M' is set-theoretically definable, though today most logicians would reckon that this is intuitively obvious. Occasionally it's useful to know that the definition can be written as a set-theoretic formula of a particular form.

The model-theoretic truth definition uses an adaptation of the idea of satisfaction that Tarski introduced in his 1933 truth definition and exploited in the 1936 paper. If you apply that model-theoretic adaptation to the 1936 definition of logical consequence, you get

ϕ is a logical consequence of T if and only if every model of T is a model of ϕ (3)

where now ϕ is a sentence and T a set of sentences, in a language whose nonlogical symbols are meaningless. It happens that the righthand clause of (3) is a relation that appears very often in model theory, so it would be useful to have a name for it. On the basis of the facts above, Tarski in 1953 [20, p. 8] proposed reading the relation as 'ϕ is a logical consequence of T'. Model theorists have tended to follow Tarski's lead and pronounce the relation as 'T entails ϕ' or 'ϕ is a consequence of T'. The use of the name has nothing to do with any interest in the concept of logical consequence itself.

Tarski's 1953 essay [20] seems to have had some unintended consequences among philosophers. A number of people conflated the 1936 definition with the 1953 one, and called both of them the 'model-theoretic definition of logical consequence'. I think the conflation is unfortunate, because the question we discussed in 1.1.1 above, about analytical relations between meanings, is one of the most important questions addressed in the 1936 definition, but it is meaningless for the languages of first-order model theory. Later, during the 1980s, the 'model-theoretic definition of logical consequence' attracted the attention of some philosophers who reassessed it as a contribution to conceptual analysis.

Peter in his invitation to me (1) referred to a 'defence of model theory, as far as the foundations of logic are concerned'. I think I'll give this a miss. To me, model theory is a way of addressing certain kinds of question in mathematics, chiefly but not exclusively in geometry, algebra and number theory. The main link to foundations

of logic is that some techniques of model theory made their way into axiomatic set theory around 1960 and continue to have an influence in large cardinal theory.

1.3 Model-Theoretic Semantics

So far, nothing that I've mentioned is directly to do with semantics, i.e. the study of meanings. Tarski called his truth definition the 'semantic definition of truth', most probably because of a formal similarity with what Kotarbiński had called 'semantic definitions'. In his truth paper [19, p. 193f.] he listed some notions that he called 'semantic': denotation, definability, truth. The notion 'meaning' was not in his list, and this is certainly not an accident.

During the 1960s a number of papers appeared that were about extending model theory from non-modal formal languages to modal ones. Some people described this as giving 'model-theoretic semantics' for modal logics. I suppose that originally 'giving a semantics' meant giving a model theory that would allow one to talk in a concrete and precise way about truth and satisfaction of modal formulas. But a subtle shift started to take place. In a standard model for modal logic, each relation symbol has an 'intension', which is a function taking each possible world to a set that is the extension of the relation symbol in that world. You can think of extensions as references, and intensions as meanings—though a lot of people have criticised these analogies. So you can think of a model for the modal logic as assigning to each meaningful expression of the language an intension that represents the 'meaning' of that expression. Around 1970 Richard Montague adapted all these notions to the study of fragments of natural languages, building on earlier work of Rudolf Carnap. From that date onwards it became common to refer to Montague-style model theories of natural language as 'model-theoretic semantics'. (Though Barbara Partee, a pioneer in this area, describes her field as 'formal semantics'.) From the mid 1970s onwards, the people who did model-theoretic semantics were mostly linguists or philosophers of language. The earlier model-theoretic semantics had been done mostly by philosophical logicians, and almost never by model theorists.

Model-theoretic semantics is useless for lexicography—you learn nothing about the meaning of the Greek noun *skindapsós* by being told that its intension maps every possible world to the set of all the things in that world that fit the description *skindapsós*. But it comes into its own for describing how the meaning of a compound phrase depends on the meanings of its constituents. Earlier we illustrated how the clauses of Tarski's truth definition tell us what things satisfy a compound formula, in terms of what things satisfy its immediate subformulas. Tarski had one clause for each logical operator: the logical operators $\rightarrow$, $\neg$, $\forall$ etc., are all of them expressions whose meaning is explained by saying how the meaning of a compound formed by means of them depends on the meanings of the constituent expressions. In modal logic and its variants we add to those logical constants other expressions like 'necessarily', 'believes', 'until'. Formal semanticists push the boat out and apply similar machinery to 'himself', 'hardly ever' and 'so much as' (for example).

Model-theoretic semantics and the model-theoretic definition of logical consequence were always completely separate. You might reckon that there is a link, because both of them are involved with giving meanings. But there are major differences. First, in studying logical consequence we are only concerned with the meaning of one expression; model-theoretic semantics aims to get a purchase on language as a whole. Second, Tarski always assumed that the expression 'logical consequence of' was not in the formal language L; it was an expression of the metatheory. Of course one can put it into the object language, but Tarski himself avoided doing this, because he had proved that languages containing enough of their own metatheory generate contradictions. So a person who wants to add 'logical consequence of' to the object language has the extra task of proving that the resulting language is still consistent. And third, the aim with logical consequence was to give a definition of it, under suitable constraints. Model-theoretic semantics doesn't give definitions, it gives truth-conditions.

So it was curious to read the introduction to Peter Schroeder-Heister's [14] and find him claiming that 'classical model-theoretic semantics' makes various assumptions about how logical consequence should be defined. I assumed at first that he was using 'model-theoretic semantics' as a name for the model-theoretic definition of logical consequence. But then almost at once he talks about model-theoretic treatment of the logical operators, and that really is in the realm of model-theoretic semantics. Well, it's not good history but it's an intriguing question all the same. Could there be a theory that helpfully combines definition of metatheoretic notions with the techniques of model-theoretic semantics? What problems would it run into? What constraints should it aim to observe? What kinds of new result could we expect from combining the two things? I think it's clear that Peter himself doesn't want to go down this road, but somebody else might. (Maybe somebody already has, in which case I give them my apologies and best wishes.)

2 Defining Meanings in General

We can separate out two strands in the aims of proof-theoretic semantics. One is to use proof theory to specify the meanings of logical constants. This can be generalised to specifying the meanings of other expressions too. (Peter tells me he would welcome faster progress in this direction, for example using more advanced proof-theoretic tools like those used to handle inductive definitions.) The other is to give a good description of logical consequence from the point of view of proof theory. I assume Peter's invitation was to comment on both of these aims. In this section I tackle meanings in general, and in the next section I turn to logical consequence.

2.1 Defining Meanings: Specialise Then Generalise

In the introduction to his Stanford Encyclopedia entry on 'Proof-Theoretic Semantics' [15] Peter says:

> ... the meaning of a term should be explained by reference to the way it is used in our language.

That's a very reasonable starting-point. I wasn't clear whether Peter takes 'our language' to be English (or German), or a formal language used in logic, but I'll assume the former. Paraphrasing Peter's statement a little, the meaning of an expression E in a language L is what you need to know in order to use E in L. But we should exclude purely grammatical information about E, so a safer statement is

> The meaning of an expression E in a language L is the further information that you need in order to use E in L, if you already know the grammatical facts about E.

There is more to be said on this, but not here.

Straight away we hit a problem. Life is open-ended, and so is language. The same expression can be used in indefinitely many different situations, and *a priori* there is no reason to think we can write down the rules for using the expression in a manageable description that covers all cases. This certainly applies to the logical constants 'and', 'every' and so on, which occur throughout the language and not just in contexts of logical argument.

So in practice we do what linguists have to do constantly in their studies. We narrow down to a set of contexts that we can handle, and we give rules for using the expression in those contexts. Then we rely on general facts about life and language to determine how the expression would be used in other contexts. I will call the narrow set of contexts the *primary applications*, and I will call the arguments used for generalising from the primary applications to the whole language the *transfer arguments*.

The 'Frege-Geach problem' illustrates these notions. In 1965 Peter Geach wrote a paper [7] in which—among other things—he attacked the view that you can explain the meaning of the sentence

> He hit her. (4)

by saying that it ascribes a certain kind of action to 'him'. Geach argues that this explanation won't carry over to contexts where (4) is used but not asserted, for example when it follows the word 'If'. In contexts where (4) is not asserted, it doesn't ascribe anything. But, says Geach, the explanation needs to be carried over to these contexts, because we can apply modus ponens and argue

> He hit her. If he hit her then q. Therefore q.

Moreover the two occurrences of the sentence, 'by itself and in the "if" clause, must have the same sense if the *modus ponens* is not to be vitiated by equivocation' [7, p. 462f].

I used to think that Geach's argument was a very clever way of refuting all sorts of plausible theories. I still think it's clever, but now it seems to me to prove almost nothing. When we explain how an expression is used in certain contexts, transfer arguments will always be needed to infer how it is used in other contexts. In fact looking again at Geach's paper, I see that this agrees with his conclusion:

> . . . it is up to [the person giving this kind of explanation] to give an account of the role of "*p*" that will allow of its standing as a premise. This task is pretty consistently shirked. [7, p. 463]

The key point that Geach contributes is that the validity of the modus ponens argument is a constraint on possible transfer arguments.

We must ask: Who has the responsibility for handling the transfer arguments?

To illustrate with 'and': a person who is explaining 'the way it is used in our language' will need to explain its use not just between propositions in deductions, but also such uses as

formally correct and materially adequate; black and white. (5)

There are subtleties here: a formally correct and materially adequate definition is a formally correct definition that is also materially adequate, but a black and white cat is not a black cat that is also white. How did we know this?

You might argue that this property of 'black and white' is something for the linguists to worry about, and not a thing that proof theorists could be expected to have views on. But on the other hand linguists can't make bricks without straw: if the proof theorists expect the linguists to explain how the proof-theoretic meaning of 'and' transfers to uses like those in (5), then they must be prepared for the linguists to complain that the proof-theoretic meaning just isn't enough to generalise from. Somebody has to take responsibility for the join-up.

The point is very general. For example an explanation of the meaning of 'He hit her' in terms of truth conditions raises the question how we can infer what it means to say

Last Friday Zayd hit Amr very hard, to teach him a lesson.

Obviously if you specified the meaning of 'hit' as the set of ordered pairs (a, b) such that a hit b, then you are going to have serious problems answering this question. (I stole this example from the great 11th century semanticist Abd al-Qāhir al-Jurjānī. Today people working on the semantics of tree-adjoining grammars wrestle with the same problem.)

2.2 Representing the Meaning

When we describe the meaning of an expression, we always do it in some format: maybe a picture, or a diagram, or a formal definition in words, or a physical demonstration, or an abstract set, or . . . In other words, the information about the expression always has to be packaged up as an object—I will call the object the *semantic value* of the expression—in some *form of representation*. This places on us the burden on making sure that both we and the people we are speaking to can read the representation, i.e. that we can *understand what information the semantic value is supposed to convey*.

There is a great temptation for logicians just to throw symbols on the page and hope that they are self-explanatory. For example we might write, as a partial explanation of 'and':

$$\frac{(\phi \text{ and } \psi)}{\phi} \tag{6}$$

But what does this diagram mean? Does it mean for example one of the following?

(a) If we are entitled to assert (ϕ and ψ) then this fact entitles us to assert ϕ.
(b) If we have already asserted (ϕ and ψ) then we are entitled to assert ϕ.
(c) If we are committed to defending (ϕ and ψ) then we are committed to defending ϕ.
(d) If $(\phi \wedge \psi)$ is true then so is ϕ.
(e) In any situation S, if $(\phi \wedge \psi)$ is true in S then ϕ is true in S.

Some of these statements are deducible from others by general principles. Let me straight away generalise the notion of transfer arguments to include the arguments that justify these deductions. These arguments generalise not from one context of use to another, but from one kind of statement about use to another kind of statement about use.

Note that if we use reading (e), then there is a very plausible argument to show that the natural deduction rules for $\wedge$ and the standard truth table for $\wedge$ give *exactly the same information* about $\wedge$, so that in this case the difference between a proof-theoretic semantics and a model-theoretic one becomes purely one of notation. But in any case a person who wants to compare model-theoretic semantics with proof-theoretic semantics for logical operators will need to answer the question above for (6), and similar ones for the other natural deduction diagrams and for truth tables. This applies to intuitionist logical operators just as much as to classical ones.

There seem to be more ways of reading a formal derivation than there are of reading a truth table. Derivations, particularly in Hilbert-style or natural deduction formalisms, look a bit like formalised natural language arguments. But usually they are missing the explanatory tags that we put all over the place in natural language arguments: 'Then', 'But', 'Suppose', 'I grant that', 'I think I can show that', 'I claim that' etc. etc.

To illustrate the possibilities, let me sketch how Ibn Sīnā thought we should read arguments in which an assumption is made and then discharged [9]. He observed

that when we introduce an assumption ϕ by saying 'If ϕ', we don't always repeat the 'If ϕ' whenever we state a proposition that depends on the assumption. (That's certainly so if ϕ is introduced with 'Let' or 'Suppose'. But Ibn Sīnā is right; one can find enough examples where it's true with 'If' too.) So, he argued, we must intend that 'If ϕ then' should be *understood* at the beginning of all relevant propositions down to the point where the assumption is discharged. So we should *understand*

$$\dfrac{\begin{array}{c}[\phi] \qquad \Psi \\ \vdots \\ \chi\end{array}}{(\phi \to \chi)} \quad \text{as meaning} \quad \begin{array}{c}(\phi \to \phi) \qquad \Psi \\ \vdots \\ (\phi \to \chi)\end{array} \tag{7}$$

In the 'understood but not stated' derivation on the right, the formula $(\phi \to \phi)$ at the top is an axiom, and the discharging step that derives $(\phi \to \chi)$ from χ falls away. A general metarule asserts that for every step $\Delta, \alpha \vdash \beta$ we have a step $\Delta, (\phi \to \alpha) \vdash (\phi \to \beta)$. (This analysis is extraordinarily close to Frege's explanation of making and discharging assumptions, though it was given over 800 years before Frege. But as Peter noted at the meeting, Ibn Sīnā and Frege had different motivations. In fact Ibn Sīnā wanted to understand the real intentions of the person giving the proof, whereas Frege aimed through *Begriffsschrift* to display the true 'logical weaving' of informal proofs that begin 'Let ...' [6, pp. 379ff].)

Ibn Sīnā's position is in effect a claim about what kind of contentful argument is expressed by the natural deduction rules. So it's directly relevant to how we can read the proof rule of $\to$-introduction as carrying information about the meaning of $\to$.

The discussion so far has used only natural deduction proof rules. It would be possible to give a semantics using $\vdash$ as a primitive notion, so that for example we define $\wedge$ by

$$(\phi \wedge \psi) \vdash \phi, \qquad (\phi \wedge \psi) \vdash \psi, \qquad \phi, \psi \vdash (\phi \wedge \psi). \tag{8}$$

(There are well-known variants of this definition.) The difficulty with taking $\vdash$ as primitive is that until we have a definition of $\vdash$, there is going to be no purchase for transfer arguments. In particular we won't be able even to raise the question whether (8) gives the same information as a truth table for $\wedge$, frankly because until $\vdash$ is explained, we don't know what information (8) is giving us.

One last point: some kinds of semantics refer to the semantic value of an expression as the 'denotation' of the expression. This is just a name, no more. It certainly doesn't entail that the semantics treats expressions as proper names of their semantic values. To single out some kinds of semantics as 'denotational' is like singling out the semantics that are written in Turkish; the classification is pointless.

3 Defining Logical Consequence

In both his truth definition and his definition of logical consequence, Tarski set new standards of carefulness about the requirements he was imposing on the definitions: what concepts could be used in the definitions, and what assumptions could be used in the justifications of the definitions. You can attack his definitions either by showing that they failed to meet the requirements, or by arguing that the requirements were inappropriate for his purposes. Or of course you can propose some different requirements that suit a different agenda. This third option wouldn't be an attack on Tarski; it would be an alternative venture.

Here is an example of an alternative venture. Suppose you want the definition of logical consequence to have the following property:

> For any propositions ϕ and ψ, if the definition of 'ψ is a logical consequence of ϕ' is that $\Gamma(\phi, \psi)$, then the statement $\Gamma(\phi, \psi)$ states criteria that can be used for convincing ourselves that ψ is (or is not) a logical consequence of ϕ.

To make this realistic, maybe we should add 'at least in simple or straightforward cases'. Also if you were a cognitive scientist, you might want to strengthen to 'the criteria that we would in fact use for convincing ourselves ...'; then the definition would express a theory about how we think.

It's not hard to show that Tarski's definition doesn't have this property. For Tarski the statement $\Gamma(\phi, \psi)$ takes the form

> For every interpretation or model M, if M makes ϕ true then M makes ψ true.

Because of the quantifier over all M, in practice the only way of showing that $\Gamma(\phi, \psi)$ holds will normally be to show the stronger statement

> For every interpretation or model M, 'M makes ψ true' is a logical consequence of 'M makes ϕ true'.

But this is just a more complicated variant of 'ψ is a logical consequence of ϕ', so it can't provide the criteria we asked for.

Prawitz presents this argument very clearly [11, p. 67f.]. But the basic point is older. It goes back at least to Ibn Sīnā, who used it to argue that you can't use the notion 'true in situation S' as a device for making the validity of an inference intuitively clear. (This appears in his *Qiyās* iii.2, unfortunately still available only in Arabic.) Several people including me have suggested that the argument poses at least a theoretical difficulty for those mental model theorists who maintain that we do in fact reason by making the kind of move that Ibn Sīnā criticised. So I don't think that proof-theoretic semanticists who present the argument should assume they are in any way swimming against the tide.

Looking around the literature in proof-theoretic semantics, I don't in fact see anything that I would regard as a criticism of Tarski's definition. Things that are

phrased as attacks on the definition are usually pleas for a different agenda. Nothing compels us to stick to the agendas of eighty years ago.

A striking pair of papers by Peter Schroeder-Heister [14] and Kosta Došen [4] raise a number of questions about the nature of definitions, and about what can be defined in terms of what. I very much welcome the questions—the general theory of definition has had a very patchy treatment by logicians in the last century—and I agree with most of the positive points that Peter and Kosta make. But some of their claims about the views of other people seem to me mighty strange.

At the heart of their arguments against 'model-theoretic semantics' is the question what can be defined in terms of what. This was a question of constant interest to the traditional Aristotelian logicians, and a large part of what they said about it strikes me as codswallop. Ouch—on general principle one shouldn't say that sort of thing about the logic of a distant culture. But what else can you say about people who insist that the only correct definition of 'human' is 'mortal rational animal', and give only circular arguments in support of this view?

There are still people who operate a broadly Aristotelian notion of the hierarchy of concepts. One notable example is the linguist Anna Wierzbicka [21, cf. p. 10]. She seems to operate by a kind of introspection of concepts. The main difficulty of introspection is that you can never be sure what is the source of the information that it serves up. I think in fact there are two main kinds of reason for regarding concept C as prior to concept D in the hierarchy of definitions. Both these reasons can in principle be lifted out of introspection and made objective, which is always an improvement.

The first kind of reason is that because of the way our minds work, we wouldn't be able to understand D unless we already understood C. For example could you understand what it is to be vengeful if you didn't already understand what it is to be angry? Could you understand what it is to be infectious if you didn't understand what it is to be ill? Or closer to home, could you come to have a concept of satisfaction if you didn't already have a concept of truth? In theory at least, questions of these kinds can be answered by seeing what you can teach to children, or whether there are natural languages in which there is a word for D but no word for C. There are surely important cognitive facts to be discovered here, but I for one would rather leave it to the experts.

The second kind of reason is not cognitive but semantic. An example is that you can define 'x is a mother' in terms of 'x is the mother of y' by quantifying out the y, but there is no logical operation that goes in the opposite direction. To handle examples like this, it's almost essential to put in the variables, because the whole point is that 'mother of' has an extra argument that is missing in 'mother'—it has an extra degree of freedom. In fact Tarski and his teacher Leśniewski seem to have been the first logicians who insisted on putting variables where they are needed, though Frege had already raised the point.

Kosta's paper does draw attention to one place where variables are needed. He points out (in his §4) that a notation for derivations which only allows us to put a variable for the conclusion is much less useful than a notation that allows us to a variable for a hypothesis as well. This is clearly correct, and I can say so with an easy

conscience because I have already (in (7) above) used a notation that does precisely have variables for the hypotheses. My notation is very standard, but in fact it's not the one that Kosta himself recommends. In effect Kosta, working in a categorial framework, calls for a notation that sets out the variables in the concept

f is a derivation of B from A. (9)

My notation doesn't show the f, but if needed one could write an f in the middle of the triangle. Also Kosta's notation can be written in a line; this is an advantage in text, but possibly a hindrance for writing out pictures of complex derivations. On the other hand my notation has the advantage that it allows one to write several hypotheses, whereas Kosta's arrow notation allows just one source for the arrow; for my application in (7) above, that would have been a fatal flaw. As all this illustrates, there are some quite subtle relationships between notation and concept, and they are very sensitive to the purpose that the notation will be put to, and the mathematical context in which it will be used.

But elsewhere Kosta forgets the variables. For example he asks [4, §5]:

> Can inferences be reduced to consequence relations? So that having an inference from A to B means just that B is a consequence of A. (10)

where should the variables go? I suggest that the concept of an inference needs three variables, essentially as in Kosta's notation (9) for derivations:

x is an inference from y to z. (11)

The notion of consequence carries just two variables:

x is a consequence of y. (12)

Kosta's question (10) asks whether (11) is definable from (12), and he expects the answer No.

Clearly Kosta is right: (11) is not definable from (12) (and *a fortiori* not 'reducible to' (12)) for the glaring semantic reason that (11) carries an extra argument. This is not just an accident of Kosta's formulation. It's an essential part of the notion of z being inferable from y that people can perform an act called making an inference from y to z, but it is certainly not part of the notion of consequence that people can make a consequence. And I agree with Kosta that this is a point worth making. I also agree with him that for purposes of the foundations of logic, a psychological analysis of 'making an inference' is not the right way to go.

But then why does Kosta add this comment?

> This reduction of inference to implication, which squares well with the second dogma of semantics, is indeed the point of view of practically all of the philosophy of logic and language in the twentieth century.

(He explains that 'implication' serves for 'consequence' here, so it is the same reduction as above.) Kosta seems here to be saying that the vast mass of twentieth century researchers in philosophy of logic and language all make a mistake not far short of adding 2 to 4 and getting 11. Sad to say, he is right that there are one or two professionals in this field who lack this elementary competence; I could document this but I won't. But 'practically all . . .': that seems to me an unreasonable accusation to make with no evidence offered.

Kosta also refers to 'the second dogma of semantics'. As Kosta formulates it in his §3 (adjusting a similar statement in Peter's [14]), this dogma states

> The correctness of the hypothetical notions reduces to the preservation of the correctness of the categorical ones.

If I understand this right, the notion of z being inferable from y is 'hypothetical' because one gets to z by using y as a 'hypothesis'. The act of doing this is essentially the same as the act of making an inference from y to z, so we are hovering around the same semantic distinction as before. But I don't think I recall ever hearing anybody argue that the notion of making an inference can be *defined* in terms of something being a Tarskian consequence of something else. Rather the opposite: Tarski gave his definition at least partly so that a usable notion of consequence was available to people who weren't interested in the notion of making an inference. It's a big world, there are lots of different things to be interested in. Preferring to work on B rather than A is not a kind of dogma.

Kosta adds that the second dogma 'may be understood as a corollary' of a dogma that categorical notions have 'primacy' over hypothetical notions. [4, §3] In the mainstream semantic and model-theoretic literature that I've seen, nobody talks about 'prior' notions or about one notion having 'primacy' over another. So the burden is on those who use these terms to explain what they mean by them, and what evidence they have for attributing views that involve these terms to semanticists. Otherwise it's they that are the dogmatists.

Peter has asked whether people who use Tarski's truth definition regard satisfaction as prior to truth. It's a reasonable question, but I think that the answer is a straight No, except in a technical sense that is probably not much relevant to this paper. Tarski's truth definition goes by recursion on the complexity of formulas. It's a common mathematical experience that when we define or prove something by recursion, it can be nontrivial to formulate the notion that we carry up through the recursion. Often it will need to carry extra features that can be discarded at the end of the recursion. The notion of satisfaction was a technical requirement of just this sort, needed for the recursive definition. But if the question is about having informal *concepts* of truth and satisfaction, then my own view has always been that satisfaction has to be understood in terms of truth and not the other way round. I should add that this is a question I came to through trying to give an intuitive introduction to model theory for non-model-theorists. It's not a question that model theorists ever have to deal with in their normal business.

References

1. Bernays, P.: Letter to Gödel, 18 January 1931. In: Feferman, S. (ed.) Kurt Gödel Collected Works Volume IV, Correspondence A-G, pp. 80–91. Clarendon Press, Oxford (2003)
2. Byrne, R.: Mental Models Website. http://www.psychology.tcd.ie/other/Ruth_Byrne/mental_models/theory.html. Cited 25 November 2013
3. Carnap, R.: Abriss der Logistik. Springer, Vienna (1929)
4. Došen, K.: Inferential semantics. In: H. Wansing (ed.) Dag Prawitz on Proofs and Meaning, pp. 147–162. Springer, Cham (2015)
5. Feferman, A.B., Feferman, S.: Alfred Tarski: Life and Logic. Cambridge University Press, Cambridge (2004)
6. Frege, G.: Über die Grundlagen der Geometrie. Jahresbericht der Deutschen Mathematiker-Vereinigung **15**, 293–309, 377–403, 423–430 (1906)
7. Geach, P.T.: Assertion. Philos. Rev. **74**, 449–465 (1965)
8. Hilbert, D., Ackermann, W.: Grundzüge der Theoretischen Logik. Springer, Berlin (1928)
9. Hodges, W.: Ibn Sina on reductio ad absurdum. Review of symbolic logic (to appear)
10. Hodges, W.: Tarski's theory of definition. In: Patterson, D. (ed.) New Essays on Tarski and Philosophy, pp. 94–132. Oxford University Press, Oxford (2008)
11. Prawitz, D.: On the idea of a general proof theory. Synthese **27**, 63–77 (1974)
12. Pullum, G.K., Scholz, B.C.: On the distinction between model-theoretic and generative-enumerative syntactic frameworks. In: De Groote, P., et al. (eds.) Logical Aspects of Computational Linguistics. Lecture Notes in Computer Science, vol. 2099, pp. 17–43. Springer, Berlin (2001)
13. Ramsey, F.P.: The foundations of mathematics. Proc. Lond. Math. Soc. **25**, 338–384 (1925)
14. Schroeder-Heister, P.: The categorical and the hypothetical: a critique of some fundamental assumptions of standard semantics. Synthese **187**, 925–942 (2012)
15. Schroeder-Heister, P.: Proof-theoretic semantics. In: Stanford Internet Encyclopedia of Philosophy (2012). http://plato.stanford.edu/entries/proof-theoretic-semantics/. Dated 5 December 2012
16. Tarski, A.: Pojęcie prawdy w językach nauk dedukcyjnych. Prace Towarzyctwa Naukowego Warszawskiego, Wydzial III Nauk Matematyczno-Fizycznych **34** (1933). Revised translation: The concept of truth in formalized languages. In: [19], pp. 152–278
17. Tarski, A.: O pojęciu wynikania logicznego. Przegląd Filozoficzny **39**, 58–68 (1936). Translated as: On the concept of logical consequence. In: [19], pp. 409–420
18. Tarski, A.: O ugruntowaniu naukowej semantyki. Przegląd Filozoficzny **39**, 50–57 (1936). Translated as: The establishment of scientific semantics. In [19], pp. 401–408
19. Tarski, A.: Logic, Semantics, Metamathematics: papers from 1923 to 1938. Corcoran, J. (ed.) Hackett Publishing Company, Indianapolis, Indiana (1983)
20. Tarski, A., Mostowski, A., Robinson, R.: Undecidable Theories. North-Holland, Amsterdam (1953)
21. Wierzbicka, A.: Semantics: Primes and Universals. Oxford University Press, Oxford (1996)

Comments on an Opinion

Kosta Došen

Abstract Wilfrid Hodges' opinion is that some ideas of Peter Schroeder-Heister and the author concerning logical consequence are largely sound and probably uncontroversial, but he criticizes some of their aspects. In this note Hodges' critique of the author is found misplaced.

Keywords Inference · Deduction · Consequence · Proof-theoretic semantics · Categorial proof theory

I am glad Wilfrid Hodges took in [9] an interest in my philosophical paper [3], but I am sorry his reading of it is marred by misunderstanding, and leads to imprudent reproaches. I don't find Peter Schroeder-Heister's ideas are evaluated correctly in [9], but I will make comments only on what is said there, in the last few pages, on my paper.

In the middle of Sect. 1.3 of [9], where the critique of my paper starts, Wilfrid says that "at the heart of their [Peter's and my] arguments against 'model-theoretic semantics' is the question what can be defined in terms of what". I was unaware that I was producing arguments against model-theoretic semantics, and as much unaware that I was dealing with Wilfrid's question. Concerning the arguments against model-theoretic semantics, the dogmas (assumptions that everybody makes, and nobody calls into question) discussed by Peter and me are accepted not only in that kind of semantics, but also in proof-theoretic semantics, and my paper discusses their acceptance in the later kind of semantics. Concerning defining, at two places (in Sects. 5 and 6 of [3]) I mentioned the inductive definition of derivations and codes for them, and the definition of inference, i.e. deduction (as I said in Sect. 2 of [3], I used the word "inference" to accord with Prawitz's usage), as an equivalence class of derivations. Elsewhere, I spoke of *definition* only when I mentioned the opinions of others. "The question what can be defined in terms of what", burdened (as the linguist Anna Wierzbicka) by the legacy of Aristotle, is hardly "at the heart of my

K. Došen (✉)
Faculty of Philosophy, University of Belgrade and Mathematical Institute,
Serbian Academy of Sciences and Arts, Knez Mihailova 36, p.f. 367,
11001 Belgrade, Serbia
e-mail: kosta@mi.sanu.ac.rs

T. Piecha and P. Schroeder-Heister (eds.), *Advances in Proof-Theoretic Semantics*,
Trends in Logic 43, DOI 10.1007/978-3-319-22686-6_12

arguments". My ideas, as should be quite clear from my paper, are very far from Aristotle. They come from categorial proof theory, a mathematical field at the border of category theory and proof theory.

Wilfrid seems to think that speaking of *priority* and *primacy* means one must be speaking about "what can be defined in terms of what". One may reasonably claim that for explaining how an organism functions physiological notions, like for example *homoeostasis*, have primacy over anatomical notions, like for example *parenchyma* or *stroma*. This does not mean that the later notions are *definable* in terms of the former ones, and not vice versa. In defining anatomical notions concerning organs one may, but need not, rely on physiological notions, but one would equally rely on anatomical notions—in particular on the notion of organ—when defining physiological notions. In general, in the order of *explanation*, it seems indisputable that one may claim precedence for the notions of a science that seeks laws accounting for phenomena over the notions of a taxonomical science (see [1], Sect. 10). Nearer to the field of logic, theoretical linguistics and its notions would have for explaining how language functions precedence, primacy, over descriptive linguistics and its notions.

In a different register, in the order of *exposition* and not the order of explanation, some notions can have for deep and natural reasons precedence over others without this meaning that the later notions are simply definable in terms of the former. In logic, one usually has that in the order of exposition the connectives of propositional logic have primacy, priority, over the quantifiers of predicate logic, without the latter being definable in terms of the former. In the foundations of mathematics, one usually has that in the order of exposition logical notions, the connectives and the quantifiers, together with the axioms concerning them, have priority over the set-theoretical membership relation, together with the axioms concerning it, without the latter being definable in terms of the former, as some authorities still expected a hundred years ago. I will return to matters of primacy towards the end of this note.

I argue in [3] and elsewhere that the notion of inference should not be understood as the notion of consequence *relation*. I don't understand what Wilfrid means by saying before (1.10) that I forgot the codes of inferences. It is quite the opposite. I argue that an inference should not be taken as an ordered pair made of the premise and the conclusion, an ordered pair which is a member of a consequence relation. With inferences we do not have a *relation*, but a *graph* in the sense of category theory, which is given by a function assigning to every arrow an ordered pair of objects (some graph-theorists call that a directed graph, and others, following [8], could call it a directed pseudograph).

Wilfrid finds after (1.9) that a notation for derivations (does he mean by that the same as I mean by *inference* or *deduction*?) in which he draws triangles "has the advantage that it allows one to write several hypotheses". The usual notation in the style of Gentzen with sequents plural on the left can claim the same merit. In the context where we are interested in *identity* of inferences there is no mathematical loss, and there is a gain in clarity, if we restrict ourselves to the categorial format with a single object as the source of an arrow. Since the premises are finite in number, we can replace a plurality of them by their conjunction, and the absence of them by the propositional constant true. If on the other hand we are interested in the question of

reducibility of inferences to normal form, then it might be worthwhile to move to a multicategorial, or operadic format, with a plurality of objects as sources. Moreover, this should be done in an analogue of bicategories, i.e. weak 2-categories (see [4]). One can also envisage working in polycategories (see [5]).

I argued at length against psychologism concerning inference towards the end of Sect. 4 of [3], and also in Sects. 6 and 7. In Wilfrid's remarks after (1.12) in [9], though at the end of the paragraph he admits that "a psychological analysis of 'making an inference' is not the right way to go", there are still psychologistic tones in his mentioning that "people can perform an act called making an inference". So I am afraid that the point Wilfrid ascribes to me with approval is not exactly mine. If one understands inference psychologistically, as much as Wilfrid seems to do, that point may be acknowledged, but I don't think it has much worth from a technical, proof-theoretical, point of view.

To speak of impersonal inferences, not performed as an act, not made by anybody, need not be natural. This may be something in the technical language of proof-theorists. The task of proof theory however is not to stick to ordinary language, but to speak about mathematical structures involved in deduction. One finds the very interesting and important partial algebras in question in categorial proof theory (see the elementary talk [2]). Model theory, as it was conceived up to now, is blind for their logical role.

Wilfrid's indignant remarks where I am accused to be saying that "the vast mass of twentieth century researchers in philosophy of logic and language all make a mistake not far short of adding 2 to 4 and getting 11" seem to stem from his assuming that a Kosta made of straw is accusing the philosophers interested in logic and language of confusing a psychologistic inference with a non-psychologistic consequence. Without putting psychologism into the picture, it was already shown by Gentzen that from a purely technical point of view it is worth studying inference syntactically, though Gentzen's sequents could be read as consequence, i.e. a generalized implication (this is how, for example, Church read them). Without psychologism, the difference between inference and consequence becomes mathematically even clearer in categorial proof theory, where one studies identity of inferences. Gentzen did not study that (though one may perhaps take that his results are pointing in that direction).

I believe that at least 95 % of logicians, and 99 % of philosophers of logic and language, do not care about the codes of inferences and identity of inferences formalized by systems of equations between these codes. They are quite happy with having inferences that amount to consequence *relations*. Inferences that have respectively the same premises and the same conclusions are for them always the same. Being one of the rare logicians working in categorial proof theory on identity of inferences (i.e. identity of deductions), in the footsteps of Lambek and Mac Lane, I dare advance the figures of these percentages.

The primacy of propositions over deductions, i.e. of asserting over deducing, is in the order of explaining how language functions, and is of the same kind as the primacy of asserting over naming, which Dummett speaks about in Chap. 1 of [6], and which is mentioned in Sect. 2 of [3]. Dummett's words are: "Frege's account, if it is to be reduced to a slogan, could be expressed in this way: that in the order of

explanation the sense of a sentence is primary, but in the order of *recognition* the sense of a word is primary." ([6], p. 4) In the penultimate paragraph of [9] Wilfrid says: "In the mainstream semantic and model-theoretic literature that I've seen, nobody talks about 'prior' notions or about one notion having 'primacy' over another."

One should first realize that in accordance with what was said about primacy and defining at the beginning of this note, and contrary to what it seems Wilfrid would have in the wake of Aristotle, this is not simply a matter of defining the notion of deduction in terms of the notion of proposition, or vice versa, or defining the notion of proposition in terms of the notion of name, or vice versa. The literature that would supply what Wilfrid says he has not seen or heard could start with that reference to Dummett and continue with the references to Frege [7] and Wittgenstein ([10] and [11]), which are also in [3]. It is surprising that after starting so auspiciously, on the shoulders of giants such as these last two, we don't manage to end up in the mainstream semantic literature.

I agree however with Wilfrid that model-theorists usually do not care about philosophical questions concerning meaning. I don't think this is because they have superior knowledge, but because together with interest they lack knowledge about these philosophical matters—as well as knowledge about many interesting and important mathematical matters of logic not in their realm.

Acknowledgments Work on this note was supported by the Ministry of Education, Science and Technological Development of Serbia. I am grateful to Wilfrid Hodges for discussing the matters raised in this note, and to the organizers of the Second Conference on Proof-Theoretic Semantics, Peter Schroeder-Heister and Thomas Piecha, for accepting to include it in the proceedings edited by them.

References

1. Došen, K.: Logical consequence: a turn in style. In: Dalla Chiara, M.L. et al. (eds.) Logic and Scientific Methods, Volume I of the 10th International Congress of Logic, Methodology and Philosophy of Science, Florence 1995, pp. 289-311. Kluwer, Dordrecht (1997). http://www.mi.sanu.ac.rs/kosta/publications.htm
2. Došen, K.: Algebras of deductions in category theory. In: Jokanović et al. (eds), Third Mathematical Conference of the Republic of Srpska, Proceedings, Trebinje 2013, Zbornik radova, vol. I, pp. 11-18. Univerzitet u Istočnom Sarajevu, Fakultet za proizvodnju i menadžment, Trebinje (2014). http://www.mi.sanu.ac.rs/kosta/DosenAlgebrasofDeductions.pdf; http://www.mk.rs.ba/wp-content/uploads/2015/02/TOM1-Copy.pdf, pp. 1–8 http://www.mi.sanu.ac.rs/kosta/publications.htm
3. Došen, K.: Inferential semantics. In: Wansing, H., (ed.) Dag Prawitz on Proofs and Meaning, pp. 147-162. Springer, Cham (2015). Preprint of 2012: http://www.mi.sanu.ac.rs/kosta/publications.htm
4. Došen, K., Petrić, Z.: Weak cat-operads (2010). Preprint v. 8: http://arXiv.org

5. Došen, K., Petrić, Z.: Graphs of plural cuts. Theor. Comput. Sci. 484, 41-55 (2013). http://arXiv.org
6. Dummett, M.A.E.: Frege: Philosophy of Language. Duckworth, London (1973)
7. Frege, G.: Die Grundlagen der Arithmetik: Eine logisch mathematische Untersuchung über den Begriff der Zahl. Verlag von Wilhelm Koebner, Breslau (1884) (English translation by J.L. Austin: The Foundations of Arithmetic: A Logico-Mathematical Enquiry into the Concept of Number, 2nd revised edn, Blackwell, Oxford, 1974)
8. Harary, F.: Graph Theory. Addison-Wesley, Reading, Mass. (1969)
9. Hodges, W.: A strongly differing opinion on proof-theoretic semantics? In: Piecha, T., Schroeder-Heister, P., (eds.) Advances in Proof-Theoretic Semantics. Springer, Berlin (2015). This volume
10. Wittgenstein, L.: Logisch-philosophische Abhandlung. Annalen der Naturphilosophie 14, 185-262 (1921) (English translation by C.K. Ogden: Tractatus logico-philosophicus, Routledge, London, 1922, new translation by D.F. Pears and B.F. McGuinness, Routledge, London, 1961)
11. Wittgenstein, L.: Philosophische Untersuchungen. Blackwell, Oxford (1953) (English translation by G.E.M. Anscombe: Philosophical Investigations, fourth edition with revisions by P.M.S. Hacker and J. Schulte, Wiley-Blackwell, Oxford, 2009)

On Dummett's "Proof-Theoretic Justifications of Logical Laws"

Warren Goldfarb

Abstract This paper deals with Michael Dummett's attempts at a proof-theoretic justification of the laws of (intuitionistic) logic, pointing to several critical problems inherent in this approach. It discusses in particular the role played by "boundary rules" in Dummett's semantics. For a revised approach based on schematic validity it is shown that the rules of intuitionistic logic can indeed be justified, but it is argued that a schematic conception of validity is problematic for Dummett's philosophy of logic.

Keywords Proof-theoretic justification · Logical laws · Dummett

Can logical laws be justified? Of course, the question can be answered, trivially, in the affirmative: a logical law can be justified by deriving it from other logical laws. But the question is meant to ask something deeper, something like: can *the* logical laws be justified, all of them. Or, at least, can the logical laws be justified on the basis of some small fragment of them, a fragment deductively weaker than the whole?

To this question it seems plausible that the answer is negative. Early analytic philosophers might have argued that since the logical laws provide the canons of justification, it does not even make sense to seek to justify them. (This view is, I take it, near to the surface, if not completely explicit, in Frege. It is the cornerstone of Carnap's thought, when he takes the specification of a linguistic framework—including all the logical laws—as a precondition for any rational inquiry or debate at all.) This philosophical view is supported by, or mirrored in, an obvious technical point: any justification would involve a deductive argument; this argument would use logical laws, so that the justification would presuppose what it is supposed to justify. Thus it would be circular, and not a justification at all. This is well illustrated

This paper was written in 1998. In 1999, at the first Tübingen conference on Proof-Theoretic Semantics, it was presented as a manuscript, with a talk by Michael Dummett devoted to it. It has been widely circulating since, and has been intensively discussed by many authors. It is here published in its original form.

W. Goldfarb (✉)
Department of Philosophy, Harvard University, Cambridge, MA, USA
e-mail: goldfarb@fas.harvard.edu

T. Piecha and P. Schroeder-Heister (eds.), *Advances in Proof-Theoretic Semantics*,
Trends in Logic 43, DOI 10.1007/978-3-319-22686-6_13

by soundness proofs for deductive systems: ordinarily, in showing soundness of a particular axiom or rule, one uses logical reasoning that is the direct analogue in the metalanguage of that very axiom or rule.

Nonetheless, as Michael Dummett has long urged (see, e.g., [1]), a negative answer might be too quick.

It might be proposed, for example, that it is the meaning of our words that have, as upshots, the acceptability of the logical laws; might not an account of those meanings therefore be able to play the role of supporting, or even fully justifying logical laws? To put a finer point on it, the suggestion is that logical laws are true by dint of the meanings of the words in them—specifically the meanings of the logical particles; and hence one might be able to find justifications of those laws simply by unfolding what the meanings of the logical particles are. The hope is that this might be done *without* invoking the full panoply of logical laws that use those particles, so as to obtain noncircular justifications.

In an odd sense, the idea goes back to Wittgenstein's discovery of truth-functional analysis: for the validity of the truth-functional laws follows at once from the stipulation of the truth-functions that the connectives represent. (I say "in an odd sense", since for Wittgenstein the logical laws have no content, and it is surely odd to speak of justifying something without content: what is there to justify?) But it should be noted that a strong assumption underlies Wittgenstein's procedure, namely his notion of propositions as bipolar—possibly true, possibly false, and determinately either one or the other. That is a highly suspect assumption, at least to those like Dummett who wish to question classical two-valued logic. So perhaps the question should be rephrased as: can we find noncircular justifications of logical laws by unfolding the meanings of the logical particles, without making strong meaning-theoretic assumptions?

Gerhard Gentzen's work in proof theory in the 1930s proved to be suggestive in this regard. Gentzen had developed logical systems in which the role of each connective was isolated, so that each basic inference rule was "about" one and only one connective. Indeed, he showed that two sorts of rules for each connective suffice. One sort allows for the introduction of the connective, and one for its elimination. In the context of a system for natural deduction (rather than in Gentzen's sequent calculus), the rule of $\wedge$-introduction is that which licenses the inference of $A \wedge B$ from premises A and B; the rules of $\wedge$-elimination license the inference of A from $A \wedge B$ and of B from $A \wedge B$. The rule of $\rightarrow$-introduction is the rule of discharge of premises: if B has been deduced from premises including A, then we may infer $A \rightarrow B$ while striking A from the list of premises. The rule of $\rightarrow$-elimination is just modus ponens, licensing the inference of B from A and $A \rightarrow B$. Gentzen suggested that introduction rules have much the same status as definitions: they fix the meaning of the connectives they introduce, at least in part. That is, an introduction rule for a connective gives the conditions under which a statement with that connective as its main connective can be inferred. Those conditions can be thought of as simply stipulated, and once stipulated, as constitutive of the meaning of the connective.

With respect to the project of justifying logical laws on the basis of the meaning of the logical particles, if we accept this view of introduction rules then clearly those

rules stand in no further need of justification. As Dummett puts it, they are "self-justifying". The question then is whether such self-justifying rules can be used to endow further logical rules with justification, in particular, rules beyond those that amount to iterated use of introduction rules.

In Chaps. 11–13 of *The Logical Basis of Metaphysics* [3], Michael Dummett formulates a method for providing what he argues are just such justifications. The introduction rules for the connectives are taken as furnishing the *canonical means* of establishing sentences whose main connectives are one of those the rules introduce. Dummett's method then seeks to show, of an inference, that any canonical argument for the premises of the inference can be transformed into a canonical argument for the conclusion. Dummett's claim is that if this can be shown, the inference is justified.

The clearest illustrative case is an inference by an elimination rule, say, an inference from $F \wedge G$ to G. A canonical argument for the premise $F \wedge G$ would end in an application of the rule of $\wedge$-introduction, that is, would end in an inference of $F \wedge G$ from F and G. But then the argument already contains a canonical argument for G. Thus, the inference is justified, since we can transform the given argument for $F \wedge G$ into a canonical argument for G simply by extracting the subargument for G. The basic idea here stems from Gentzen's [6] technique of normalization of proofs, which he devised to prove his cut-elimination theorem. Dummett's use of the technique as a justificatory procedure is inspired by a similar proposal of Dag Prawitz from the early 1970s (especially in [7]), although there are differences in formulation and in scope.

This method of justifying logical laws is important to Dummett for several reasons. First, it provides a sense in which logical inferences *can* be justified, in a way that is clearly noncircular, and so stills the doubt I mentioned at the start as to whether any such program could make sense. Moreover, although the method presupposes the self-justifying nature of introduction rules, and so relies on a view of the meaning-endowing nature of those rules, the method need not invoke a full-fledged and comprehensive theory of meaning, as Dummett's better-known arguments criticizing classical logic and supporting intuitionistic logic do. Since we seem to be no closer to obtaining a comprehensive Dummettian theory of meaning for natural language than we were when Dummett formulated his meaning-theoretic program 25 years ago, this avoidance of invoking such a theory makes the method more credible and presumably less open to controversy.

As it turns out, or so Dummett asserts, the method provides justification for intuitionistic logic but not for classical logic, at least not for the classical laws about negation. Thus it gives important support to his position that intuitionistic logic is preferable to classical. Indeed, it exhibits a virtue of intuitionistic logic—justifiability on the basis of laws that merely express the meaning of the connectives—that classical logic fails to have: "[Intuitionistic logic's] logical constants can be understood, and its logical laws acknowledged, without appeal to any semantic theory and with only a very general meaning-theoretical background." [3, p. 300] The failure of this method for the laws of classical negation thus allows an invidious distinction to be made.

In this paper I investigate Dummett's method, as it applies to sentential logic.[1] I shall show that, even in this restricted domain, Dummett's method won't do: it provides "justifications" for obviously invalid inferences. I shall consider how to repair the damage, and analyze the question of whether the repair restores confidence in the philosophical framework underlying Dummett's claim that his method does indeed justify. The results are, I think, suggestive of some overlooked, and possibly deep, difficulties in Dummett's overarching project of marrying intuitionism and a verificationistic theory of meaning.

1 Analysis of the Method

In order to make the method precise, we must define the notion of a canonical argument, for, to repeat, the idea is that an inference is justified if any canonical argument for its premises can be transformed into a canonical argument for its conclusion. The definition should make plausible the following: if a logically complex proposition is provable at all, then it could in principle be proved by a canonical argument. For only if that condition is met will Dummett's procedure have any plausible claim to justificatory force. In line with the underlying idea, it might be tempting to define a canonical argument as one composed only of introduction rules. This does not work, however, because of the nature of the introduction rule for the conditional, to wit: $F \to G$ may be inferred from a subsidiary argument from premise F to conclusion G, discharging the premise F. Since F itself may be logically complex, the argument from F to G cannot be restricted to those that use introduction rules only, or else many elementary logical truths will not be obtainable by canonical arguments, for example, $A \wedge B \to B$. That is, a canonical argument will end in $\to$-introduction:

$$\frac{\begin{array}{c}[A \wedge B] \\ B\end{array}}{A \wedge B \to B}$$

But if we are constrained to using only introduction rules, we will not be able to fill in the middle part. Hence the subsidiary arguments, the ones starting from premises that will eventually be discharged, cannot be constrained to contain only introduction rules. All that can be required of such subsidiary arguments is that they themselves be already recognized as justified. The result is a definition, by simultaneous induction on the complexity of the statements in the arguments, of the notion of "valid canonical argument" along with the notion of "valid argument":

[1] Dummett actually proposes the method for full first-order logic. Moreover, he aims at definitions that could apply to arbitrary new connectives, as well as our familiar ones.

> A *valid canonical argument* is a deduction whose premises are all atomic sentences and that uses only introduction rules except when auxiliary premises are introduced; at any point when such are introduced, the subargument from the point of the introduction of the first new premise to the last step before the discharge of the last new premise must be a valid argument.
>
> A *valid argument* is an inference I such that any valid canonical argument (i.e., any valid canonical argument with any atomic premises) for the premises of I can be transformed into a valid canonical argument, with the same atomic premises, for the conclusion of I.

We have simplified matters slightly by omitting what Dummett calls "boundary rules", which allow the inference of one atomic sentence from others.[2] For example, these may be empirical laws, connecting the primitive notions of the vocabulary. Dummett allows the employment of such rules in valid canonical arguments. For the moment we take there to be no such rules, since the mathematics is clearer without them. In the next section, we shall allow boundary rules and investigate their impact.

The validity of an argument depends only on its premises and conclusion, and not on any intervening steps. Hence the second definition is framed as applying to inferences, rather than deductions. The simultaneous induction works because discharge of premises increases logical complexity. Thus, whether a deduction with conclusion F is a valid canonical argument depends on the validity of arguments whose premises and conclusion are of strictly lesser logical complexity than F.

These definitions are far from transparent. Applying them involves tracking through the tree structures of deductions in natural deduction systems. Most importantly, the definitions do not readily yield any general information about the range of inferences that are valid or not.

However, the definitions can be greatly clarified if we focus not on the proof-theoretic layout but rather on the relation that holds between a set α of atomic sentences and a formula F when there is a valid canonical argument with conclusion F and premises among the atomic sentences in α. Let us use "$\alpha \Vdash F$" for this relation. Using this notation we may frame the definition of "valid" thus: an inference from premises $F_1, \ldots, F_n$ to conclusion G is valid iff, for all sets α, if $\alpha \Vdash F_i$ for each i, then $\alpha \Vdash G$. (It may seem that this reformulation ignores a constructivity requirement, implicit in the phrase "we can transform" of the original definition. However, since we are dealing with sentential logic only, all notions are decidable and all quantifiers in the metalanguage are constructively evaluable.)

We can now investigate the relation $\alpha \Vdash F$, by looking at how its behaviour for logically complex F depends on its behaviour on the constituents of F. A valid canonical argument for $F \wedge G$ is just a valid canonical argument for F and a valid canonical argument for G, put together by means of a final inference to $F \wedge G$, using the rule of $\wedge$-introduction. A valid canonical argument for $F \vee G$ is either a valid canonical argument for F followed by one application of $\vee$-introduction or else a

[2] These amount to the definitions given by Dummett in [3, p. 261], simplified by the absence of boundary rules and (more importantly) of the need to deal with free variables.

valid canonical argument for G followed by one application of $\vee$-introduction. These observations immediately yield:

$$\alpha \Vdash F \wedge G \text{ iff } \alpha \Vdash F \text{ and } \alpha \Vdash G \tag{1}$$

$$\alpha \Vdash F \vee G \text{ iff } \alpha \Vdash F \text{ or } \alpha \Vdash G \tag{2}$$

A valid canonical argument for $F \to G$ with atomic premises in α is a valid inference I to G from premises F and members of α, followed by an application of $\to$-introduction, discharging the premise F and yielding $F \to G$. The inference I will be valid provided that every valid canonical argument for F whose premises may include members of α and possibly some other atomic sentences can be transformed into a valid canonical argument for G whose premises are either in α or are among those others. This yields the condition:

$$\alpha \Vdash F \to G \text{ iff } \forall\beta(\text{if } \alpha \subseteq \beta \text{ and } \beta \Vdash F, \text{ then } \beta \Vdash G). \tag{3}$$

(1)–(3) show that the relation $\Vdash$ is, in fact, a familiar one from the semantics of intuitionistic logic, since they are nothing other than rules for the treatment of the connectives in the usual Kripke model semantics, when we take the sets α of atomic sentences as the nodes (worlds) of the model, and the relation $\alpha \subseteq \beta$ as the relation of extension. Thus the proof-theoretic trappings of Dummett's presentation conceal a notion whose structure is the same as the standard model-theoretic or semantic one.

One connective remains to be considered, namely, negation. As Dummett notes, the only way to treat negation that is consonant with his general procedure is to take $\neg F$ as an abbreviation for $(F \to \bot)$, where $\bot$ is a sentential constant governed by the following introduction rule: from premises that are all the atomic sentences, it may be inferred. Dummett allows there to be infinitely many atomic sentences; in fact, this treatment of negation fares poorly if there are not. For if $A_1, \ldots, A_n$ exhaust the atomic sentences, then the introduction rule just mentioned yields the validity of inferring $\neg(A_1 \wedge \ldots \wedge A_n)$ with no premises. Thus on logical grounds alone we would be able to infer that not every atomic statement is true, and this is surely an unacceptable result.

If there are infinitely many atomic sentences, then this treatment of negation can most easily be incorporated into our forcing relation by requiring that the domain of sets of atomic sentences α that we consider is always finite. Then the stipulation above becomes:

$$\alpha \Vdash \bot \text{ for no } \alpha. \tag{4}$$

The resulting rule for negation is then: $\alpha \Vdash \neg F$ iff $\forall\beta(\text{if } \alpha \subseteq \beta \text{ then not } \beta \Vdash F)$. This is just the standard rule for the treatment of negation in the semantics of intuitionistic logic.

The characterization of the forcing relation will be complete once we give the clause governing atomic sentences themselves. Since we are at the moment allowing

no boundary rules, we have:

$$\text{for any atomic sentence } A, \alpha \Vdash A \text{ iff } A \in \alpha. \tag{5}$$

As we've just seen, Dummett's notion of valid canonical model yields a relation $\Vdash$ that obeys just the usual semantic rules for models of intuitionism, as given by (1)–(4). However, there is a key difference between $\Vdash$ as used in Dummett's method and the ordinary model-theory of intuitionistic logic. In the latter, the validity of an inference would mean that at each node (world) in *every* Kripke model, if the premises are true then the conclusion is true. Dummett's method, in contrast, amounts to considering only one particular structure, the Kripke model in which every finite set of atomic sentences is a distinct node, and for every finite set of atomic sentences there is exactly one node at which all and only those sentences are true, namely, the node that is the set of those sentences. This restriction to one particular structure yields anomalous results.

Counterexample 1 *If F does not contain $\bot$, then the inference from no premise to $\neg\neg F$ is valid.*

Proof It is easily shown by induction on the construction of F that if F does not contain $\bot$ then for every α there exists β with $\alpha \subseteq \beta$ and $\beta \Vdash F$. But then for no γ do we have $\gamma \Vdash \neg F$. Hence, for every α, $\alpha \Vdash \neg\neg F$. □

By the way, since we have shown that, for F that do not contain $\bot$, $\gamma \Vdash \neg F$ for no F, we also have the conclusion that, for such F and any G, the inference from no premises to $\neg F \to G$ is valid.

Counterexample 2 *Let F be a sentence not containing $\bot$ and G a sentence having no atomic sentences in common with F. Then the inference from premise $F \to G$ to conclusion G is valid.*

Proof Suppose $\alpha \Vdash F \to G$; we must show $\alpha \Vdash G$. By (3), for any β with $\alpha \subseteq \beta$, if $\beta \Vdash F$ then $\beta \Vdash G$. Moreover, as noted in the previous proof, there exists a β such that $\alpha \subseteq \beta$ and $\beta \Vdash F$.

The following is easily shown by induction on the construction of sentences: for any sentence H and any sets γ and δ, if $A \in \gamma$ iff $A \in \delta$ for all atomic sentences A that occur in H, then $\gamma \Vdash H$ iff $\delta \Vdash H$.

Thus, if β' is the subset of β containing just those atomic sentences either in α or occurring in F, we have $A \in \beta$ iff $A \in \beta'$ for all A that occur in F. Since $\beta \Vdash F$, it follows that $\beta' \Vdash F$. Hence $\beta' \Vdash G$. Since F and G have no atomic sentence in common, no atomic sentence occurring in G is in $\beta' - \alpha$. Thus $A \in \beta'$ iff $A \in \alpha$ for all A that occur in G. Hence $\alpha \Vdash G$. □

Thus there are many inferences that turn out valid under Dummett's definition, and yet are logically valid in no plausible sense. The counterexamples show that such inferences exist even in the fragment of the language that does not contain $\bot$, and so does not contain negation. As a particularly vivid case, we have the validity of the inference from $A \to B$ to B whenever A and B are distinct atomic sentences! We must conclude that Dummett's method has no justificatory force whatsoever.

2 Boundary Rules

To see how the trouble arises in terms of canonical arguments, rather than the relation $\Vdash$, it is helpful to consider the case of the inference from $A \to B$ to B, where A and B are distinct atomic sentences. If there were to be a valid canonical argument for $A \to B$, it would have to enable us to transform any valid canonical argument for A into one for B. Since B is atomic, the only valid canonical argument for B is the one-step argument of taking B as a premise. Hence a valid canonical argument for $A \to B$ must have B as an (undischarged) premise; and so it will be transformable into a valid canonical argument for B. The problem, in short, is that there is no way of getting from A to B, except by taking B as premise.

Here, it might be thought, is where Dummett's boundary rules can play a role, since boundary rules license inferences from atomic formulas to atomic formulas. However, three considerations—one technical and two philosophical—show that the problems in the method cannot be avoided by boundary rules as Dummett envisages them.

First, if the counterexamples are to be avoided, there are going to have to be an inordinate number of boundary rules. To forestall the validity of the inference from $A \to B$ to B, there must be a rule allowing the inference of B from A (and possibly other premises not including B) for *any* pair (A, B) of distinct atomic sentences. To forestall the validity of the inference from no premise to $\neg\neg A$, there must be a rule allowing the inference of $\perp$ from A (again, possibly with other premises). Rules that avoid some anomalies may engender others. For example, if $\perp$ can be inferred by boundary rules from premises A and B, and from premises A and C, but not from A and any other premises, then although the inference from no premise to $\neg\neg A$ is no longer valid, the inference from $\neg A$ to $B \vee C$ is. It appears, then, that it is unreasonable to expect that boundary rules will avoid the difficulty.

(By the way, it is not clear that a rule allowing the inference of $\perp$ from atomic premises should count as a boundary rule at all. Dummett characterizes boundary rules as "rules governing ... non-logical expressions." Allowing $\perp$ as a conclusion violates this description. After all, a rule allowing the inference of $\perp$ from premises A and B is just a rule allowing the inference of $\neg B$ from A, and of $\neg A$ from B. This significantly weakens the claim that $\perp$ is given meaning only by its introduction rule; indeed, it seems to me to weaken the contrast Dummett makes between intuitionistic negation and classical, saying of the latter "there is no way of attaining an understanding of the classical negation operator if one does not have it already" [3, p. 299] Nonetheless, if we are to block the anomalies given by Counterexample 1, we must allow boundary rules with conclusion $\perp$.)

Alongside the technical difficulties there are philosophical ones. To use boundary rules in the manner envisioned makes the validity of inferences dependent on which boundary rules there are, and hence, in particular, on empirical claims about the connections of different empirical basic sentences. This is not consistent with the claim that the validity of the logical inferences comes only from the meaning of the logical connectives (as based on the introduction rules).

Finally, even if the latter difficulty is set aside, there is another disturbing consequence, namely, that it becomes impossible to put forth a link between atomic sentences as a supposition, and draw consequences from it. For either the link is taken as a boundary rule, and hence becomes part of the logical framework, usable in any argument anywhere and playing a role in the criterion of validity; or else there is no link, in which case having $A \to B$ as a supposition yields B as a valid conclusion, and therefore we can infer from the conditional everything that is yielded by its consequent alone. The irony here is that we have landed in a position akin to Frege's odd-sounding view that "Only true thoughts can be premises of inferences." [5, p. 335][3]

The true nature of the difficulty should be apparent, by now. The intuitionist reading of $F \to G$ is, roughly, "from any demonstration of F we can obtain a demonstration of G." In Brouwer and the early intuitionistic tradition, the notion of demonstration here is taken to be open-ended, identified not with any particular formal system, indeed, not with the entirety of means of demonstration we currently have at our disposal, but as anything that we might come to accept as a demonstration. In later studies, particularly those inspired by Kreisel's work of the 1950s, the generality in talking of "any demonstration" is expressed by speaking of the intuitionist $\to$ as being "impredicative": $F \to G$ implicitly quantifies over all demonstrations, including those that may contain the very demonstration of $F \to G$. Dummett, in contrast, wants to read "any demonstration" here as meaning "any valid canonical argument", where this notion is defined in an inductive and hence purely predicative way. It is this restriction that gives rise to the difficulties above, both in the case without boundary rules, and the peculiarities of trying to use a fixed set of boundary rules to block those difficulties.

It is I think far more natural to use the notion of boundary rule in a way not envisaged by Dummett, and in fact inconsistent with Dummett's aim. The definition of "valid" can be revised so that what counts as a valid inference is one that was valid in the old sense given *any* assumption of boundary rules.[4] This revision avoids both of the philosophical difficulties just canvassed. It does not restrict allowable arguments to a fixed set of accepted ones, but rather allows any collection of possible arguments from atomic sentences to atomic sentences. Since all sets of boundary rules are considered, there is no need for empirical input to determine which boundary rules should be adopted.

Technically, the consideration of all sets of boundary rules amounts to the consideration of different model-theoretic structures. There are two equivalent ways of formulating this. Given a set S of boundary rules, the relation $\Vdash$-relative-to-S, or $\Vdash_S$ as we shall write it, can be defined by appropriate changes in clauses (4) and (5), keeping clauses (1)–(3) as is. Alternatively, (1)–(5) can be kept as is, and the domain of sets altered to contain all and only sets α that are closed under all the boundary rules in S and do not contain $\perp$. For our purposes, the latter procedure is more convenient. For any set α of atomic sentences, let $cl_S(\alpha)$ be the closure of α

[3]For Dummett's appraisal of this view, see [4, p. 313].

[4]This is the idea in the work of Prawitz [7, p. 236].

under the rules in S, that is, the smallest set β such that $\alpha \subseteq \beta$ and if S contains a rule "infer B from $A_1, \ldots, A_n$" and $A_1, \ldots, A_n$ are in β, then B is in β.

It is easy to show that every inference that is valid in the revised sense is classically valid. Suppose the inference from premise F to conclusion G is valid in the revised sense. Let T be a (classical) truth-assignment to the atomic sentences in F and G under which F comes out true; we must show that G also comes out true under T. Let S be the set of boundary rules containing "from no premise infer A" for every atomic sentence A to which T assigns truth, and "from A infer $\perp$" for every other atomic sentence A. Obviously, there is only one set α that is closed under S and does not contain $\perp$, namely, the set of atomic sentences assigned truth by T. But then $\Vdash_S$ behaves classically on the connectives, so that $\alpha \Vdash_S F$. Since the inference is valid in the revised sense, $\alpha \Vdash_S G$. Hence G is true under T.

From this we can surmise that there will be no counterexamples of the alarming sort encountered above. However, validity in the revised sense still does not coincide with intuitionistic validity.

Counterexample 3 *Let A be an atomic sentence, and G and H any sentences. Then the inference from premise $A \to (G \vee H)$ to conclusion $(A \to G) \vee (A \to H)$ is valid in the revised sense.*

Proof Let S be a set of boundary rules, and suppose α is an S-closed set not containing $\perp$ such that $\alpha \Vdash_S A \to (G \vee H)$. Let β be the S-closure of $\alpha \cup \{A\}$. If $\perp \in \beta$, then $\alpha \Vdash_S A \to F$ for every F, so $\alpha \Vdash_S (A \to G) \vee (A \to H)$; hence we may suppose $\perp \notin \beta$. If $\beta \Vdash_S G$, then $\alpha \Vdash_S A \to G$, for if γ is any S-closed extension of α with $\gamma \Vdash_S A$ then $\beta \subseteq \gamma$, so that $\gamma \Vdash_S G$; similarly if $\beta \Vdash_S H$ then $\alpha \Vdash_S A \to H$; in either case $\alpha \Vdash_S (A \to G) \vee (A \to H)$. But if neither, then β is an S-closed extension of α such that $\beta \Vdash_S A$ while not $\beta \Vdash_S G \vee H$, which contradicts the hypothesis that $\alpha \Vdash_S A \to (G \vee H)$. $\square$

In the usual model-theory of intuitionistic logic, say via Kripke trees, one obtains a model of $A \to (G \vee H)$ that is not a model of $(A \to G) \vee (A \to H)$ by having two nodes v_1 and v_2, one of which models A and G but not H, the other models A and H but not G. For this it is essential that there be no u with $u \leq v_1$ and $u \leq v_2$ that models A; for if the root of the tree is to model $A \to (G \vee H)$ any such u would have to model $G \vee H$, and thus have to model G or model H, but every node above u would also model G or every node would also model H, thus defeating the example. The problem is that, using $\Vdash$ and boundary rules, these strictures cannot be met. For example, suppose G and H are also atomic. Using boundary rules one can insure that there is a closed set containing A and G and a distinct one containing A and H, but then there will also be a closed set containing A that is a subset of each of those, and in order to insure that $A \to (G \vee H)$ holds, that subset will have to contain either G or H.

It may be helpful to translate the situation back into Dummett's proof-theoretic language. Again suppose G and H, as well as A, are atomic. The counterexample shows that any valid canonical argument for $A \to (G \vee H)$ can be transformed into one for $(A \to G) \vee (A \to H)$. Suppose, then, there is a valid canonical argument

from premises α to conclusion $A \to (G \vee H)$. This is just to say that the inference from α and A to $G \vee H$ is valid, which in turn means that every valid canonical argument for α and A can be transformed into one for $G \vee H$. Since there is a valid canonical argument from α and A to α and A, there must be one from α and A to $G \vee H$. The last step of this must be an application of $\vee$-introduction. Hence there is either a valid canonical argument from α and A to G or one from α and A to H, and so there is a valid canonical argument from α to $(A \to G) \vee (A \to H)$. The idea is that there is only one way to demonstrate A, so to speak.

3 Schematic Inferences

The counterexamples I have presented are not *schematic* inferences, that is, inferences that rely only on the forms of the premises and conclusion. The inferences that I showed to be valid-by-Dummett's lights (although not valid in any ordinary sense) were further constrained, e.g., in Counterexample 2 the formulas could have no atomic constituent in common, and in Counterexample 3 the antecedent had to be atomic.

The question naturally arises as to how Dummett's definitions fare on schematic inferences. Let us call an inference *schematically* valid in Dummett's original sense iff the inference and all its instances are valid in Dummett's original sense. (An instance is simply any inference obtained from the original one by replacing atomic sentences with other sentences.)

Now any inference that is schematically valid is classically valid, since if F does not imply G in the classical sense, a truth-assignment T that makes F true and G false can be mimicked by the substitution instances of F and G in which sentence letters assigned truth by T are replaced with "$p \to p$" and those assigned falsity are replaced with "$\perp$". The resulting instances F^* and G^* are such that $\emptyset \Vdash F^*$ but not $\emptyset \Vdash G^*$ (since forcing will then just amount to two-valued truth-computation).

However, schematic validity outstrips intuitionistic logic.

Counterexample 4[5] *The inference from no premise to $\neg F \vee \neg\neg F$ is schematically valid (that is, for any F the inference from no premise to $\neg F \vee \neg\neg F$ is valid in Dummett's original sense).*

Proof If not $\alpha \Vdash \neg F$, then there exists β such that $\alpha \subseteq \beta$ and $\beta \Vdash F$. But then, for any γ such that $\alpha \subseteq \gamma$, there exists δ such that $\gamma \subseteq \delta$ and $\delta \Vdash F$, namely, $\delta = \gamma \cup \beta$. Thus, for any γ such that $\alpha \subseteq \gamma$, not $\gamma \Vdash \neg F$. That is, $\alpha \Vdash \neg\neg F$. □

However, we can obtain a positive result if we combine the notion of schematic inference with that of validity-in-the-revised-sense, that is, validity given any collection of boundary rules. That is, it is possible to prove the following:

[5] I owe this counterexample to Philip Kremer.

Theorem *If every instance of the inference from F to G is valid in the revised sense, then the inference from F to G is intuitionistically valid.*

For the proof, see the Appendix.

4 Assessment

Dummett can't take too much comfort in this positive result. Dummett is careful to point out, in framing his method, that the inferences treated are actual inferences, involving particular meaningful sentences, the atomic components of which are actual atomic sentences, not schematic parts [3, p. 254]. That is, he treats the language as fully interpreted. On the view he is propounding, an inference is justified by its validity; the justification of a schematic inference (an inference rule) can lie only in the fact that each of its instances is justified. There is simply no room, on his view, for a position to the effect that validity does not justify an inference unless all inferences like it in being subsumable under a particular rule are also justified.

Let us return to Counterexample 3. Intuitionistically, the inference from $A \to (G \vee H)$ to $(A \to G) \vee (A \to H)$ is incorrect, because the former means "any demonstration of A can be transformed either into one for G or into one for H" whereas the latter means "any demonstration of A can be transformed into one for G or any demonstration of A can be transformed into one for H." This inference turns out to be valid, in the revised sense, because—so to speak—the method treats an atomic formula as its own proof. That is, the criterion of the identity of a proof is just the atomic formulas it has as premises or are implied by its premises. Thus distinct ways of proving an atomic A will not register as distinct. Now this view of proofs arises because the whole set-up envisages all proofs as, ultimately from atomic formulas. And this is the upshot of the set-up's being part of, or the beginnings of, a verificationist meaning-theory.

I believe the basic views that lead to Dummett's difficulties are well exhibited in the following remark (he is speaking here only of mathematics, but presumably he would maintain the same for a language with empirical vocabulary as well):

> If the intuitionistic explanations of the logical constants and, more generally, of the meanings of mathematical statements are to be considered as constituting a coherent theory of meaning for the language of mathematics, then the notion of proof which is appealed to must be such that we can fully grasp the concept of a proof of any constituent of a given sentence in advance of grasping that of a proof of that sentence. It cannot, therefore, be identified with the notion of the sort of proof that we may, at some future time, come to consider valid ... [2, p. 402]

This remark expresses a fundamental view of Dummett's; and from it we can see three sources of the problems with his program for "proof-theoretic justification". Of course, most generally, his underlying concern to meld intuitionistic logic with theory of meaning impels him to differ with the Brouwerian tradition of the open-endedness of the notion of demonstration, and as we saw that was key to the anomalies

exemplified by Counterexamples 1 and 2. But the remark also expresses Dummett's commitment to molecularity: that what it is to prove a sentence is explained in terms of what it is to prove each constituent of the sentence. That, of course, is a denial of the impredicative nature of intuitionistic conditionals; and it signals his commitment to just the view of the proofs of atomic sentences as, if you like, logically unanalyzable, that engenders Counterexample 3.

The difference between Dummett and the treatments of intuitionism standard in mathematical logic on these two points has not been sufficiently explored. In classical truth-functional semantics, atomic sentences are the basic building blocks, and it is clear why. Dummett takes atomic sentences to be the basic building blocks of a proof-theoretic semantics—and on both the "basic" aspect and the "building block" aspect he differs from classical intuitionism. It is not clear why one should believe this, except perhaps for the conflation of the notion of verification and a mathematical notion of proof.

The third factor expressed in Dummett's remark is that there will be no definite meaning ascribed to a sentence unless it is fixed what a demonstration is for that sentence. That will presuppose that not just the logical rules but also the boundary rules are fixed. This tells us that, from Dummett's viewpoint, the revised notion of validity is not acceptable. For, in considering all possible boundary rules, it should be clear, the revised notion of validity treats the atomic components of sentences in abstraction from their actual content. It takes them to be schematic, in that their connections to one another (and hence also to complex sentences) are varied at will, but this is precisely what Dummett's insistence that he seeks to justify actual inferences, not schematic ones, would rule out.[6]

Finally, I suppose the following line might be taken. The claim that Dummett's method provides justifications of logical laws might be abandoned or weakened, while still it be pressed that the method does show *something*. That, under the revised notion of validity, the inference rules that yield valid inferences under all substitutions are not the classical ones but precisely the intuitionistic ones, surely supports the ascription of *some* advantageous status to intuitionistic logic. But here we should note at once that the method—in taking the introductory rules as definitory of the connectives—identifies the sense of $F \to G$ as "G can be validly inferred from F", and then goes on to define the latter as "every valid canonical argument for F can be transformed into a valid canonical argument for G". Thus, built into the method at the start is the intuitionistic construal of the conditional. As pointed out above, this is just what leads to condition (3) on the $\Vdash$-relation, and that in turn is the characteristic of the model theory of intuitionism. If (3) were to be replaced by $\alpha \Vdash F \to G$ iff (if $\alpha \Vdash F$ then $\alpha \Vdash G$), then what we obtain will be a classical

[6]Nor is Dummett's insistence ill-placed. The justificatory force of his method rests on what he calls the "fundamental assumption", that a logically complex sentence, if demonstrated, could have been demonstrated by a (valid) canonical argument. For this reason, Dummett spends an entire chapter of [3] investigating the exact sense and the plausibility of the fundamental assumption. Clearly, for the fundamental assumption to make sense at all requires that the sentences about which it speaks have content. If they are merely schemata, it is unclear what the assumption could mean, unless it is to be true by fiat.

notion of validity. In short, it should occasion no surprise, that the (revised) method yields intuitionistic validity, because the method is based on, or presupposes, an intuitionistic reading of the conditional. (Although attention is usually focused on negation as that which marks the difference between classical and intuitionistic, a case can be made that it's the conditional. A classical conditional combined with the definition of $\neg F$ as $F \to \bot$ would still yield the classical laws of negation.) And for this reason, that the method yields the intuitionistic inferences once it is applied schematically does not signal any greater virtue of intuitionistic logic, at least as framable from neutral ground. What is odd, and perhaps even undermining of the claims to virtue of intuitionistic logic, is that even when the intuitionistic reading of the conditional is built into the project at the start, it still takes lots of fussing here and jiggling there to get the method to yield just the intuitionistic laws.

Appendix

Theorem *Let F and G be sentential formulas such that the inference from F to G is not intuitionistically valid. Then there are instances F^* and G^* of F and G, a set S of boundary rules, and a set α of atomic sentences such that $\alpha \Vdash_S F^*$ but not $\alpha \Vdash_S G^*$.*

Proof Let Σ be the set of atomic sentences occurring in F or in G. Since the inference from F to G is not intuitionistically valid, there is a Kripke tree $(W, \leq, I)$ with root w such that $(W, \leq, I) \vDash_w F$ but not $(W, \leq, I) \vDash_w G$. Here $(W, \leq)$ is a tree (we take the root as being at the bottom), and I is a mapping from W to subsets of Σ: for each $u \in W$, $I(u)$ is the set of atomic sentences true at u. I is subject to the constraint that if $u \leq v$ then $I(u) \subseteq I(v)$. We wish to obtain sets of atomic sentences that "mimic" $(W, \leq, I)$. For this we shall need additional atomic sentences for there may be distinct nodes in W making the same atomic sentences of Σ true, but to these nodes we want to have correspond distinct sets of atomic sentences. For each u in W let u^* be a distinct atomic sentence not in Σ, and for each u in W let $\varphi(u) = \{v^* \mid v \in W \text{ and } v \leq u\}$. $\varphi(u)$ will be the set of atomic sentences corresponding to the node u. We now so formulate boundary rules that the only S-closed sets that do not contain $\bot$ are precisely the sets $\varphi(u)$ for $u \in W$. In fact, let S be the following set of boundary rules:

"Infer $\bot$ from $A_1, \ldots, A_n$ whenever $\{A_1, \ldots, A_n\} \subseteq \varphi(u)$ for no $u \in W$;"
"Infer v^* from u^* whenever $u, v \in W$ and $v < u$."

Now, for each $p \in \Sigma$, let $D(p)$ be the disjunction of all atomic sentences u^* such that p is in $I(u)$. Finally, for any sentence H constructed from members of Σ, let H^* be obtained from H by replacing each atomic sentence p with $D(p)$.

It is a routine matter to show by induction on the construction of formulas that, for any H and any $u \in W$, $(W, \leq, I) \vDash_u H$ iff $\varphi(u) \Vdash_S H^*$. It then follows from the supposition that $\varphi(u) \Vdash_S F^*$ but not $\varphi(u) \Vdash_S G^*$, so that the inference from F^* to G^* is not valid, in the revised sense. □

Author's Postscript, January 2015

From 1999 to 2007 I presented this paper at various universities and conferences. Often audience members raised stimulating points, particularly about my suggestions in Sect. 4, which I hoped to address in an expanded version, but I never managed to do so to my satisfaction. However, the basic issues still seem to me to be well-framed in the original version that is printed here.

The last presentation I gave was in September 2007 at the Oxford philosophy of mathematics seminar led by Daniel Isaacson. I was delighted that Michael Dummett was able to attend, despite infirmities of age. (Sir Michael and I had been on warm terms since his spring semester 1976 residence at Harvard, when he delivered the William James Lectures from which *The Logical Basis of Metaphysics* [3] evolved, and I was in my first year on the Harvard faculty.) It was particularly pleasing that at the 2007 seminar one of the younger Oxford philosophers, Ofra Magidor, raised the same objection that Sir Michael had framed nine years earlier in a letter to me, namely that Counterexamples 1 and 2 (from Sect. 1) are really not worrisome at all. In his 1998 letter he wrote, "I do not accept that your counter-examples are genuinely such." His point and Magidor's was that if G and F have no atomic sentences in common, and F has no occurrences of $\perp$, then the only reason to assert $F \to G$ would be that one had independent reason for thinking that F were false or G were true, and since the former is ruled out by F not containing $\perp$, it isn't at all surprising that we can infer G.

However this point seems to me mistaken: for if the rule (infer G from $F \to G$, when F does not contain $\perp$ and G and F contain no atomic parts in common) is justified, it is justified in application not just to assertions but also to *suppositions*. On the ordinary understanding of what it is to suppose $F \to G$, this is simply untenable. So if the rule is to be accepted, it would have to be argued that the ordinary understanding of supposition is incorrect, and in fact to suppose $F \to G$ is to suppose something like "every canonical argument for F can be transformed into a canonical argument for G". But this is clearly wrong if F and G are empirical. (It might be maintained for mathematical F and G, but in this case it would again appear that a bias in favor of intuitionistic logic were being built in at the ground level.)

Even though I was criticizing his position, Sir Michael clearly enjoyed my presentation at the seminar, no doubt because he thought the issues needed more discussion. Despite his infirmities, he maintained his famously cheerful humour as well as his robust sense of what philosophy could aspire to do. I dedicate this publication to his memory.

References

1. Dummett, M.: The Justification of Deduction. Oxford University Press, Oxford (1974). Reprinted in M. Dummett, Truth and Other Enigmas, Duckworth, London (1978), 290–318
2. Dummett, M.: Elements of Intuitionism. Oxford University Press, Oxford (1977)
3. Dummett, M.: The Logical Basis of Metaphysics. Harvard University Press, Cambridge (1991)
4. Dummett, M.: Frege: Philosophy of Language, 2nd edn. Harvard University Press, Cambridge (1993)
5. Frege, G.: Collected Papers on Mathematics, Logic and Philosophy. Basil Blackwell, New York. Translated by M. Black, V. Dudman, P. Geach, H. Kaal, E.-H.W. Kluge, B. McGuinness & R.H. Stoothoff(1984)
6. Gentzen, G.: Untersuchungen über das logische Schließen. Mathematische Zeitschrift 39, 176–210, 405–431 (1934/35). Translated in M.E. Szabo (ed.), The Collected Papers of Gerhard Gentian, 68–131. North-Holland, Amsterdam (1969)
7. Prawitz, D.: Towards a foundation of a general proof theory. In: Suppes, P., Henkin, L., et al. (eds.) Logic, Methodology, and Philosophy of Science IV, 225–250. North-Holland, Amsterdam (1973)

Self-contradictory Reasoning

Jan Ekman

Abstract This paper concerns the characterization of paradoxical reasoning in terms of structures of proofs. The starting point is the observation that many paradoxes use self-reference to give a statement a double meaning and that this double meaning results in a contradiction. Continuing by constraining the concept of meaning by the inferences of a derivation "self-contradictory reasoning" is formalized as reasoning with statements that have a double meaning, or equivalently, cannot be given any meaning. The "meanings" derived this way are global for the argument as a whole. That is, they are not only constraints for each separate inference step of the argument. It is shown that the basic examples of paradoxes, the liar paradox and Russell's paradox, are self-contradictory. Self-contradiction is not only a structure of paradoxes but is found also in proofs using self-reference. Self-contradiction is formalized in natural deduction systems for naïve set theory, and it is shown that self-contradiction is related to normalization. Non-normalizable deductions are self-contradictory.

Keywords Paradox · Proof structure · Self-contradiction · Proof theory · Russell's paradox

1 Introduction

Let us consider Russell's paradox:

> Let t be the set of all sets not containing themselves. Assume that t contains itself. Hence, by the definition of t, t does not contain itself. This contradicts the assumption that t contains itself and hence t does not contain itself. Since t does not contain itself, it follows from the definition of t that t contains itself. This is a contradiction.

This article is based on Chap. 5 of the author's PhD thesis; see Ekman (1994) [2]. Definitions of elementary notions can be found in the Appendix below.

J. Ekman (✉)
SICS Swedish ICT AB, Box 1263, SE-164 29 Kista, Sweden
e-mail: jan@sics.se

T. Piecha and P. Schroeder-Heister (eds.), *Advances in Proof-Theoretic Semantics*,
Trends in Logic 43, DOI 10.1007/978-3-319-22686-6_14

Let us take a closer look at the part of Russell's paradox that proves that *t does not contain itself*. Let $\mathscr{E}_{\mathscr{R}}$ be this part of Russell's paradox. We observe that the assumption that *t contains itself* is used twice in $\mathscr{E}_{\mathscr{R}}$. We shall now distinguish the use of an assumption from how it is used. Let us therefore, to express that an assumption is used in an argument, say that the assumption *occurs* in an argument. Thus there are two occurrences of the assumption that *t contains itself* in $\mathscr{E}_{\mathscr{R}}$. One of these two occurrences of the assumption that *t contains itself* is used together with the definition of *t* to derive that *t does not contain itself*. To contradict this last proposition the other occurrence of the assumption that *t contains itself* is used. Hence, there are two occurrences of the assumption that *t contains itself* in $\mathscr{E}_{\mathscr{R}}$, and they are used in such a way that they contradict each other. In the last step of $\mathscr{E}_{\mathscr{R}}$, the conclusion that *t does not contain itself* is drawn from the contradiction that the assumption that *t contains itself* leads to. In a sense the two occurrences of the assumption that *t contains itself* are identified in this step. Considering the two occurrences of the assumption that *t contains itself* as one and the same proposition, we have that there in $\mathscr{E}_{\mathscr{R}}$ is a proposition which is used in two ways and that the two ways of using the proposition are incompatible.

A *self-contradictory argument* is, informally, an argument, as $\mathscr{E}_{\mathscr{R}}$ above, in which there is a proposition which is used in two or more ways such that not all of the ways of using the proposition are compatible. In this article we aim to make those ideas more precise and formally express the notion of self-contradictory reasoning in some formal systems.

2 Meaning Conditions

The notion of a self-contradictory argument as introduced in the previous section is based on "the way in which a proposition is used in an argument." In this section we aim at making it more precise what we mean by this, and we will outline how the notion of a self-contradictory argument will be formally expressed in the succeeding sections. Given an argument and a proposition of this argument we shall in the following consider *the meaning forced on the proposition, by the steps of the argument*. The meaning forced on a proposition, by the steps of the argument, expresses precisely the way in which the proposition is used in the argument.

Let us consider an example. Let $\mathscr{D}$ be the following argument: *The wind is blowing* because *it's snowing and the wind is blowing*. Let A be the proposition *it's snowing and the wind is blowing* and let B be the proposition *the wind is blowing*. Thus $\mathscr{D}$ consists of one step and A and B are the premise and the conclusion, respectively, of this step. If we forget about which propositions A and B represent we still know something about them by remembering what kind of step the inference of $\mathscr{D}$ is. That is, knowing only that the inference of $\mathscr{D}$ is of the kind that informally corresponds to one of the &E inference schemata in natural deduction for naïve set theory **N** (see Appendix below), we know that since A is the premise of the step, A is A_1 *and* A_2 for some propositions A_1 and A_2. Moreover, if A is A_1 *and* A_2 then B is A_2. The

meaning forced on the propositions A and B by the inference of $\mathscr{D}$ is this knowledge about A and B given by the knowledge about what kind of step the inference of $\mathscr{D}$ is. Hence the meaning of *the meaning forced on the proposition, by the steps of the argument* depends on what is considered to be known, when knowing only what kind of steps the steps of the argument are.

In the previous section, "a self-contradictory argument" was explained to be an argument in which there is a proposition which is used in two or more ways such that not all of the ways of using the proposition are compatible. In this section "the meaning forced on a proposition, by the steps of the argument" expresses precisely the way in which the proposition is used in the argument. Hence, we can explain what "a self-contradictory argument" is by saying that it is an argument such that the steps of the argument force several meanings on one of the propositions of the argument and that not all of these meanings are compatible. Yet another way to put this is to say that an argument is self-contradictory if and only if the steps of the argument force an *ambiguous meaning* on one of the propositions of the argument. Note that, as is clear from the example above, the meaning forced on a proposition by an argument is not an interpretation of the proposition but a constraint on how it may be interpreted.

Now we change to how to formally express "a self-contradictory argument." Let us by *the meaning of a proposition* mean an interpretation of the proposition. For instance, *the wind is blowing* is the meaning of the proposition B in the example above. Let A be a formula occurrence in a deduction in some formal system. To denote that A has a certain meaning, m say, we decorate A with m. More precisely, we shall write $m : A$ to denote that A has the meaning m. We use these decorations to define *meaning conditions*. Meaning conditions are formal representations of the constraints given by the meaning forced on a proposition by an argument. For every formal system considered in this article we shall do the following. We shall define what the set of formal meanings is for decorating the formulas in deductions in the formal system and we shall give the meaning conditions associated with the formal system. Thus, through the meaning conditions we formally define what is informally described by "the way in which a proposition is used in an argument." By an *assignment* of meanings to the formulas in a deduction we mean a decoration of all of the formulas in the deduction. That a meaning is *assigned* to a formula means that the formula has been decorated with the meaning. The meaning conditions are given as constraints on the decorations, by formal meanings, of the formulas in the deductions. As an example let us consider, in the formal system **N**, a deduction consisting of an $\supset$E inference, α say. Let X, Y and Z be the major premise, the minor premise and the conclusion, respectively, of α. Let m_x, m_y and m_z denote some meanings assigned to X, Y and Z, respectively. We decorate the formulas in the deduction as follows.

$$\frac{m_x : X \qquad m_y : Y}{m_z : Z}\,\alpha$$

Reasoning in the same way as in the previous example, we know that since X is the major premise of an $\supset$E inference, X must be $X_1 \supset X_2$ for some propositions X_1 and X_2. We express this constraint by requiring the meaning m_x to be $m \Rightarrow n$ for some meanings m and n, where thus $\Rightarrow$ means "implies that." Moreover we require m_y to be m and m_z to be n. Thus, m_x may not be *it's snowing and the wind is blowing*. However m_y may be *it's snowing and the wind is blowing* and m_z may be *the wind is blowing*. In this case m_x must be *it's snowing and the wind is blowing implies that the wind is blowing*. We express meaning conditions given for any $\supset$E inference in any deduction in the formal system **N** by the schema

$$\frac{\begin{array}{c}\mathscr{D}\\ m \Rightarrow n : A\end{array} \qquad \begin{array}{c}\mathscr{E}\\ m : B\end{array}}{n : C}\ \supset\mathrm{E}$$

Hence the meaning condition for the major premise A of an $\supset$E inference is that A must have the meaning $m \Rightarrow n$ for some meanings m and n. Moreover, the meanings of the major premise, the minor premise and the conclusion respectively must have the relation to each other expressed by the schema. The notion of a self-contradictory deduction in a formal system is defined as follows.

Definition 1 Assume that **F** is a formal system. Assume that the set of formal meanings for decorating the formulas in the deductions in **F** are defined, and assume that the meaning conditions associated with the formal system are given in some way. Then a deduction $\mathscr{D}$ in **F** is *self-contradictory* if there is no assignment of formal meanings to the formulas in $\mathscr{D}$ such that this assignment satisfies the meaning conditions.

The meaning conditions, as we shall give them, are related to the *inversion principle* of Prawitz. In Prawitz (1965) [6] we can read the following.

> Observe that an elimination rule is, in a sense, the inverse of the corresponding introduction rule: by an application of an elimination rule one essentially only restores what had already been established if the major premise of the application was inferred by an application of an introduction rule.

We may say that, for a given deduction, the constraint expressed by the meaning conditions is an attempt to make the inversion principle global, in the deduction. But this attempt is successful if and only if the deduction is not self-contradictory, since otherwise there is no assignment of formal meanings to the formulas in the deduction such that this assignment satisfies the meaning conditions.

The Curry-Howard interpretation may resemble what designates meanings in the meaning conditions. However, the similarity is only superficial. In general, it is not the case that the assignment of Curry-Howard interpretations to the formula occurrences in a deduction satisfies the meaning conditions. Since the Curry-Howard interpretation is just a representation of an argument, there are always Curry-Howard interpretations of the formula occurrences in a deduction, but there need not be an assignment of formal meanings to the formulas in the deduction such that this assignment satisfies the meaning conditions.

3 The Liar Paradox

In this section we shall study the liar paradox as an example of a self-contradictory argument. The liar paradox is the following.

> Let P be the sentence "This sentence is false." That is, P is the sentence "P is false." Assume P. Hence, by the definition of P, P is false. This contradicts the assumption P, and hence P is false. Since P is false, P follows from the definition of P. This is a contradiction.

This argument is very similar to Russell's paradox. Below we present the formal system **FP**, specially designed for a formal presentation of the liar paradox. The language of **FP** is the set of formulas, where $\bot$ and P are formulas, and if A and B are formulas, then $A \supset B$ is a formula; $\neg A$ is defined to be $A \supset \bot$. The inference schemata of **FP** are the following.

$$\dfrac{\begin{array}{c}\mathscr{D}\\ \neg P\end{array}}{P}\,\mathrm{PI} \qquad\qquad \dfrac{\begin{array}{c}\mathscr{D}\\ P\end{array}}{\neg P}\,\mathrm{PE}$$

$$\dfrac{\begin{array}{c}[A]\\ \mathscr{D}\\ B\end{array}}{A \supset B}\,\supset\mathrm{I} \qquad\qquad \dfrac{\begin{array}{c}\mathscr{D}\\ A \supset B\end{array} \quad \begin{array}{c}\mathscr{E}\\ A\end{array}}{B}\,\supset\mathrm{E}$$

The liar paradox is formally represented by the following deduction $\mathscr{G}$,

$$\left.\dfrac{\begin{array}{c}\mathscr{F}\\ \neg P\end{array} \quad \dfrac{\begin{array}{c}\mathscr{F}\\ \neg P\end{array}}{P}\,\mathrm{PI}}{\bot}\,\supset\mathrm{E}\right\}\mathscr{G} \quad \text{where} \quad \mathscr{F}\left\{\dfrac{\dfrac{[P]}{\neg P}\,\mathrm{PE} \quad [P]}{\dfrac{\bot}{\neg P}\,\supset\mathrm{I}}\,\supset\mathrm{E}\right.$$

The set of formal meanings to be assigned to formulas in deductions in the formal system **FP** is inductively defined as follows. The meaning variable x is a meaning, and if m and n are meanings, then pm and $m \Rightarrow n$ are meanings. We may interpret the meanings as follows: $m \Rightarrow n$ means "m implies that n," and pm means "This sentence is false," where "This" refers to the sentence expressed by m. The meaning conditions associated with the formal system **FP** are the following.

$$\dfrac{\begin{array}{c}\mathscr{D}\\ m : \neg P\end{array}}{pm : P}\,\mathrm{PI} \qquad\qquad \dfrac{\begin{array}{c}\mathscr{D}\\ pm : P\end{array}}{m : \neg P}\,\mathrm{PE}$$

$$\dfrac{\begin{array}{c}[m : A]\\ \mathscr{D}\\ n : B\end{array}}{m \Rightarrow n : A \supset B}\,\supset\mathrm{I} \qquad\qquad \dfrac{\begin{array}{c}\mathscr{D}\\ m \Rightarrow n : A \supset B\end{array} \quad \begin{array}{c}\mathscr{E}\\ m : A\end{array}}{n : B}\,\supset\mathrm{E}$$

Now assume that there is an assignment of formal meanings to the formulas in the deduction $\mathscr{F}$ above such that this assignment satisfies the meaning conditions.

Assume that m is the meaning of the minor premise P of the $\supset$E inference and that n is the meaning of the conclusion $\perp$ of the $\supset$E inference. Then, by the conditions above we conclude that the meaning of the premise P of the PE inference must be $p(m \Rightarrow n)$.

$$\left.\begin{array}{c}\dfrac{\dfrac{[p(m \Rightarrow n) : P]}{m \Rightarrow n : \neg P}\,\mathrm{PE} \qquad [m : P]}{n : \perp}\,\supset E \\ \dfrac{}{? : \neg P}\,\supset\mathrm{I}\end{array}\right\}\mathscr{F}$$

The condition given for the $\supset$I inference schema requires both of the formulas cancelled at the $\supset$I inference in $\mathscr{F}$ to have the same meaning. However, no matter how we choose m and n the meanings m and $p(m \Rightarrow n)$ are not the same. Hence, there is no assignment of formal meanings to the formulas in $\mathscr{F}$ such that this assignment satisfies the meaning conditions. Hence, $\mathscr{F}$ is self-contradictory.

4 Self-contradictory Reasoning in $\mathbf{N}_{-\forall\exists=}$

Let $\mathbf{N}_{-\forall\exists=}$ be the fragment of $\mathbf{N}$ obtained by removing the symbols $\forall$, $\exists$ and $=$ and the inference schemata corresponding to these symbols from $\mathbf{N}$. In this section we shall study the notion of self-contradictory deductions in the formal system $\mathbf{N}_{-\forall\exists=}$. We shall also prove the following theorem.

Theorem 1 *Every non-self-contradictory deduction in* $\mathbf{N}_{-\forall\exists=}$ *is normalizable.*

In this section and the two succeeding ones we shall use the terminology of Ekman (1994) [2, Sect. 3.1], see Appendix below. Hence, by "normalizable" in Theorem 1 we mean normalizable as defined in Ekman (1994) [2, Sect. 3.1], see Appendix below. As in the formal system **FP**, m and n denote meanings.

Assume that A is a formula such that there is no normal proof of A in $\mathbf{N}_{-\forall\exists=}$. Then, by Proposition 3.1.4 in Ekman (1994) [2] there is no normalizable proof of A in $\mathbf{N}_{-\forall\exists=}$. Hence by Theorem 1 every proof of A is self-contradictory. Since there is no normal proof of $\perp$ in $\mathbf{N}_{-\forall\exists=}$ it follows that every paradox in $\mathbf{N}_{-\forall\exists=}$ is self-contradictory, if by paradox we mean a proof of $\perp$. In Ekman (1994) [2, Sect. 2.1] it is shown that there is no normal proof of the formula $t \notin u$, where t is the term defined by

$$t \equiv \{x \mid x \in u \ \&\ x \notin x\}$$

Hence, every proof of $t \notin u$ in $\mathbf{N}_{-\forall\exists=}$ is self-contradictory. In Ekman (1994) [2, Sect. 2.1] also a proof, named Crabbe's counterexample (see Crabbé (1974) [1]), of the formula $t \notin u$ is presented. This proof is a proof in $\mathbf{N}_{-\forall\exists=}$ and hence Crabbe's counterexample is a self-contradictory proof. It is also argued in Ekman (1994) [2, Sect. 2.1] that Crabbe's counterexample expresses a correct argument in **ZF**. Hence

the formula $t \notin u$, or the proposition that informally corresponds to $t \notin u$, serves as an example of a proposition provable in **ZF**, but only by self-contradictory proofs, unless we use proof principles not expressible in $\mathbf{N}_{-\forall\exists=}$.

The variables of the language of $\mathbf{N}_{-\forall\exists=}$ will also be used to denote *meaning variables*. The set of formal meanings to be assigned to formulas in deductions in the formal system $\mathbf{N}_{-\forall\exists=}$ is inductively defined as follows. The meaning variable x and *false* are meanings, and if m and n are meanings, then ϵm, $m \Rightarrow n$, $m \wedge n$ and $m + n$ are meanings. The meaning conditions associated with the formal system $\mathbf{N}_{-\forall\exists=}$ are the following.

$$\dfrac{\begin{array}{c}\mathscr{D}\\ m : A[t/x]\end{array}}{\epsilon m : t \in \{x \mid A\}}\ \in\mathrm{I} \qquad \dfrac{\begin{array}{c}\mathscr{D}\\ \epsilon m : t \in \{x \mid A\}\end{array}}{m : A[t/x]}\ \in\mathrm{E}$$

$$\dfrac{\begin{array}{c}\mathscr{D}\\ \mathit{false} : \bot\end{array}}{m : A}\ \bot\mathrm{E}$$

$$\dfrac{\begin{array}{c}[m : A]\\ \mathscr{D}\\ n : B\end{array}}{m \Rightarrow n : A \supset B}\ \supset\mathrm{I} \qquad \dfrac{\begin{array}{c}\mathscr{D}\\ m \Rightarrow n : A \supset B\end{array} \quad \begin{array}{c}\mathscr{E}\\ m : A\end{array}}{n : B}\ \supset\mathrm{E}$$

$$\dfrac{\begin{array}{c}\mathscr{D}\\ m : A\end{array} \quad \begin{array}{c}\mathscr{E}\\ n : B\end{array}}{m \wedge n : A \,\&\, B}\ \&\mathrm{I} \qquad \dfrac{\begin{array}{c}\mathscr{D}\\ m \wedge n : A \,\&\, B\end{array}}{m : A}\ \&\mathrm{E1} \qquad \dfrac{\begin{array}{c}\mathscr{D}\\ m \wedge n : A \,\&\, B\end{array}}{n : B}\ \&\mathrm{E2}$$

$$\dfrac{\begin{array}{c}\mathscr{D}\\ m : A\end{array}}{m + n : A \vee B}\ \vee\mathrm{I} \qquad \dfrac{\begin{array}{c}\mathscr{D}\\ n : B\end{array}}{m + n : A \vee B}\ \vee\mathrm{I} \qquad \dfrac{\begin{array}{c}\mathscr{D}\\ m_1 + m_2 : A_1 \vee A_2\end{array} \quad \begin{array}{c}[m_1 : A_1]\\ \mathscr{E}_1\\ n : C\end{array} \quad \begin{array}{c}[m_2 : A_2]\\ \mathscr{E}_2\\ n : C\end{array}}{n : C}\ \vee\mathrm{E}$$

Let $\mathscr{D}$ and $\mathscr{E}$ be two deductions in $\mathbf{N}_{-\forall\exists=}$ such that $\mathscr{D}$ is non-self-contradictory, θ is an assignment of formal meanings to the formula occurrences in $\mathscr{D}$ such that this assignment satisfies the meaning conditions and $\mathscr{D} \Longrightarrow \mathscr{E}$ (i.e., $\mathscr{D}$ reduces to $\mathscr{E}$; see Appendix for the definition of reductions of deductions). Then we let $a(\theta, \mathscr{E}, \mathscr{D})$ denote the assignment of formal meanings to the formula occurrences in $\mathscr{E}$ given by considering every formula occurrence in $\mathscr{E}$ to correspond to a formula occurrence in $\mathscr{D}$ and assigning the same meaning to the formula occurrence in $\mathscr{E}$ as the meaning assigned to the corresponding formula occurrence in $\mathscr{D}$. If $\mathscr{D}$ reduces to $\mathscr{E}$ via an epsilon reduction, then the deduction $\mathscr{D}$, with its formula occurrences decorated by θ has the form

$$\left.\begin{array}{c}\mathscr{F}\\ \dfrac{\dfrac{m : A[t/x]}{\epsilon m : t \in \{x \mid A\}}\ \in\mathrm{I}}{m : A[t/x]}\ \in\mathrm{E}\\ \mathscr{G}\\ C\end{array}\right\}\mathscr{D}$$

In this case $\mathscr{E}$, with its formula occurrences decorated by $a(\theta, \mathscr{E}, \mathscr{D})$, is the following deduction

$$\left.\begin{array}{c} \mathscr{F} \\ m : A[t/x] \\ \mathscr{G} \\ C \end{array}\right\} \mathscr{E}$$

If $\mathscr{D}$ reduces to $\mathscr{E}$ via an imply reduction, then $\mathscr{D}$, with its formula occurrences decorated by θ, has the form

$$\left.\begin{array}{c} \dfrac{\dfrac{\begin{array}{c}[m : A] \\ \mathscr{F} \\ n : B\end{array}}{m \Rightarrow n : A \supset B}\supset\mathrm{I} \qquad \begin{array}{c}\mathscr{G} \\ m : A\end{array}}{n : B}\supset\mathrm{E} \\ \mathscr{H} \\ C \end{array}\right\} \mathscr{D}$$

In this case $\mathscr{E}$, with its formula occurrences decorated by $a(\theta, \mathscr{E}, \mathscr{D})$, is the deduction

$$\left.\begin{array}{c} \mathscr{G} \\ m : A \\ \mathscr{F} \\ n : B \\ \mathscr{H} \\ C \end{array}\right\} \mathscr{E}$$

For all other cases of the kind of reduction that takes $\mathscr{D}$ to $\mathscr{E}$, $a(\theta, \mathscr{E}, \mathscr{D})$ is defined similarly.

Lemma 1 *If a deduction $\mathscr{D}$ is non-self-contradictory and $\mathscr{D}$ reduces to $\mathscr{E}$, then also the deduction $\mathscr{E}$ is non-self-contradictory.*

Proof Let θ be an assignment of formal meanings to the formula occurrences in $\mathscr{D}$ such that this assignment satisfies the meaning conditions. Then $a(\theta, \mathscr{E}, \mathscr{D})$ is an assignment of formal meanings to the formula occurrences in $\mathscr{E}$ such that this assignment satisfies the meaning conditions. □

Let the formal system **P** of propositional logic be given as in the Appendix below. We assume that there is at least one propositional variable P in the language of **P**. Let * be the function from the set of meanings to be assigned to formulas in deductions in the formal system $\mathbf{N}_{-\forall\exists=}$ onto the set of formulas of **P**, defined as follows.

$$
\begin{aligned}
x^* &\equiv P \\
\mathit{false} &\equiv \bot \\
(\epsilon m)^* &\equiv (\bot \supset \bot) \supset m^* \\
(m \Rightarrow n)^* &\equiv m^* \supset n^* \\
(m \wedge n)^* &\equiv m^* \,\&\, n^* \\
(m + n)^* &\equiv m^* \vee n^*
\end{aligned}
$$

We extend * to a function from the set of sets of meanings to be assigned to formulas in deductions in the formal system $\mathbf{N}_{-\forall\exists=}$ onto the set of sets of formulas of $\mathbf{P}$ by letting Γ^* denote the set of formulas A^* such that A belongs to Γ, for all sets of meanings to be assigned to formulas in deductions in the formal system $\mathbf{N}_{-\forall\exists=}$.

We extend * once more, to a function from the set of non-self-contradictory deductions in $\mathbf{N}_{-\forall\exists=}$ to the set of deductions in $\mathbf{P}$. If $\mathscr{D}$ is a deduction in $\mathbf{N}_{-\forall\exists=}$ consisting of the open assumption $m : A$, then $\mathscr{D}^*$ is the open assumption m^*:

$$
\left(\dfrac{\begin{array}{c}\mathscr{D}\\ m : A[t/x]\end{array}}{\epsilon m : t \in \{x \mid A\}}\,{\in}\mathrm{I}\right)^* \equiv \dfrac{\begin{array}{c}\mathscr{D}^*\\ m^*\end{array}}{(\bot \supset \bot) \supset m^*}\,{\supset}\mathrm{I}
$$

Observe that there is no open assumption of the form $\bot \supset \bot$ in $\mathscr{D}^*$, cancelled at the $\supset$I inference, in the deduction to the right above.

$$
\left(\dfrac{\begin{array}{c}\mathscr{D}\\ \epsilon m : t \in \{x \mid A\}\end{array}}{m : A[t/x]}\,{\in}\mathrm{E}\right)^* \equiv \dfrac{\begin{array}{c}\mathscr{D}^*\\ (\bot \supset \bot) \supset m^*\end{array} \qquad \dfrac{[\bot]}{\bot \supset \bot}\,{\supset}\mathrm{I}}{m^*}\,{\supset}\mathrm{E}
$$

For all other cases of the end inference of a deduction $\mathscr{D}$, the definition of $\mathscr{D}^*$ commutes with the definition of deduction. For instance, for the case that an $\supset$I is the last inference of a deduction, we have the following clause defining the image under * of this deduction:

$$
\left(\dfrac{\begin{array}{c}[m : A]\\ \mathscr{D}\\ n : B\end{array}}{m \Rightarrow n : A \supset B}\,{\supset}\mathrm{I}\right)^* \equiv \dfrac{\begin{array}{c}[m^*]\\ \mathscr{D}^*\\ n^*\end{array}}{m^* \supset n^*}\,{\supset}\mathrm{I}
$$

Proposition 1 *Assume that $\mathscr{D}$ is a non-self-contradictory deduction, θ is an assignment of formal meanings to the formula occurrences in $\mathscr{D}$ such that this assignment satisfies the meaning conditions and $\mathscr{D} \Longrightarrow \mathscr{E}$. Let $\mathscr{D}$ also denote the deduction obtained from $\mathscr{D}$ by decorating the formula occurrences in $\mathscr{D}$ with θ. Let $\mathscr{E}$ also denote the deduction obtained from $\mathscr{E}$ by decorating the formula occurrences in $\mathscr{E}$ with $a(\theta, \mathscr{E}, \mathscr{D})$. Then $\mathscr{D}^* \Longrightarrow \mathscr{E}^*$.*

Since $\mathbf{P}$ is strongly normalizable (see Prawitz (1965) [6]), we have Theorem 1 as a consequence of Proposition 1.

5 Self-contradictory Reasoning in $\mathbf{N}_{-\exists=}$

Under the assumption that meaning conditions formally express the way in which a proposition is used, as outlined in Sect. 2, it is a bit more complicated to define the meaning conditions associated with a formal system with quantifiers than it is to define the meaning conditions associated with a quantifier-free formal system. In this section we shall study the notion of self-contradictory deductions in the formal system $\mathbf{N}_{-\exists=}$, which is the fragment of $\mathbf{N}$ obtained by removing the symbols $\exists$ and $=$ and the inference schemata corresponding to these symbols from $\mathbf{N}$. We shall also prove the following theorem.

Theorem 2 *Every non-self-contradictory deduction in* $\mathbf{N}_{-\exists=}$ *is normalizable.*

Let A be any formula. To define the meaning conditions associated with the formal system $\mathbf{N}_{-\exists=}$ we shall informally consider $\forall x A$ to represent the informally given infinitely long formula $A[t_1/x]\ \&\ (A[t_2/x]\ \&\ (A[t_3/x]\ \&\ \ldots))$, where $t_1, t_2, t_3, \ldots$ are all terms of the formal system $\mathbf{N}_{-\exists=}$.

A naïve way to give the meaning conditions associated with $\mathbf{N}_{-\exists=}$ is to add the following meaning conditions to the meaning conditions associated with $\mathbf{N}_{-\forall\exists=}$, where λ is assumed to have been added to the constructors of the syntax defining what the set of formal meanings to be assigned to formulas in deductions is, such that λm is a meaning for any meaning m.

$$\dfrac{\begin{array}{c}\mathscr{D}\\ m : A\end{array}}{\lambda m : \forall x A}\ \forall\mathrm{I} \qquad\qquad \dfrac{\begin{array}{c}\mathscr{D}\\ \lambda m : \forall x A\end{array}}{m : A[t/x]}\ \forall\mathrm{E}$$

With meaning conditions given this way, we require that there is a one to one correspondence between the meaning of the premise and the conclusion both for $\forall$I inferences and for $\forall$E inferences. This condition is however too strong, if we consider $\forall x A$ to represent the informally given infinitely long formula above, since the meaning conditions given for &E inferences does not require that there is a one to one correspondence between the meaning of the premise and the conclusion of an &E inference. As an example, consider the following deduction.

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{[r \in \{y \mid A\}\ \&\ (r \in \{y \mid \neg A\}\ \&\ C)]}{r \in \{y \mid \neg A\}\ \&\ C}\ \&\mathrm{E}}{r \in \{y \mid \neg A\}}\ \in\mathrm{E}}{\neg A[r/y]} \qquad \dfrac{\dfrac{[r \in \{y \mid A\}\ \&\ (r \in \{y \mid \neg A\}\ \&\ C)]}{r \in \{y \mid A\}}\ \&\mathrm{E}}{A[r/y]}\ \in\mathrm{E}}{\bot}\ \supset\mathrm{E}}{\neg(r \in \{y \mid A\}\ \&\ (r \in \{y \mid \neg A\}\ \&\ C))}\ \supset\mathrm{I}$$

This deduction is non-self-contradictory independently of which formulas A and C are. It is straightforward to assign meanings to the formula occurrences of the deduction above such that this assignment satisfies the meaning conditions. Assume that C is $\forall x(r \in x)$ and let us consider C to represent the informally given formula $(r \in t_1)\ \&\ ((r \in t_2)\ \&\ ((r \in t_3)\ \&\ \ldots))$, where $t_1, t_2, t_3, \ldots$ are all terms of the formal

system $\mathbf{N}_{-\exists=}$. Then $r \in \{y \mid A\}$ & $(r \in \{y \mid \neg A\}$ & $C)$ and C informally represent the same formula. We have the following proof of $\neg C$, which from an informal point of view is another presentation of the deduction above.

$$\dfrac{\dfrac{\dfrac{[\forall x(r \in x)]}{r \in \{y \mid \neg A\}}\,\forall\mathrm{E}}{\neg A[r/y]}\,\in\mathrm{E} \qquad \dfrac{\dfrac{[\forall x(r \in x)]}{r \in \{y \mid A\}}\,\forall\mathrm{E}}{A[r/y]}\,\in\mathrm{E}}{\dfrac{\bot}{\neg\forall x(r \in x)}\,\supset\mathrm{I}}\,\supset\mathrm{E}$$

This deduction is self-contradictory if the meaning conditions are given as above.

We suggest the following definition of meaning conditions associated with the formal system $\mathbf{N}_{-\exists=}$. The set of formal meanings to be assigned to formulas in deductions in the formal system $\mathbf{N}_{-\exists=}$ is inductively defined as follows. The meaning variable x and *false* are meanings, and if m and n are meanings, then ϵm, $m \Rightarrow n$, $m \wedge n$, $m + n$ and $\lambda x.m$ are meanings. The meaning conditions are the following and in addition the meaning conditions associated with the formal system $\mathbf{N}_{-\forall\exists=}$.

$$\dfrac{\begin{matrix}\mathscr{D}\\ m : A\end{matrix}}{\lambda x.m : \forall x A}\,\forall\mathrm{I} \qquad\qquad \dfrac{\begin{matrix}\mathscr{D}\\ \lambda x.m : \forall x A\end{matrix}}{m[n/x] : A[t/x]}\,\forall\mathrm{E}$$

We have the restriction on the meaning variable, designated x, in the $\forall$I meaning condition schema that it may not occur free in any meaning assigned to an open assumption in $\mathscr{D}$. This restriction excludes, for instance, the following decoration of a deduction.

$$\dfrac{\dfrac{\lambda y.x : \forall y(r \in y)}{x : r \in x}\,\forall\mathrm{E}}{\lambda x.x : \forall x(r \in x)}\,\forall\mathrm{I}$$

Remember that the aim is to define the meaning conditions so that the meaning conditions express a constraint given by the meaning forced on a proposition given by an argument, in the sense of Sect. 2. Remember also that the meaning forced on a proposition given by an argument is arbitrary so far as what is considered to be known is arbitrary, when knowing only what kind of steps the steps of the argument are. We do not claim that the meaning conditions given are the only possible. The given meaning conditions express constraints which we judge as accurate.

We have chosen the constraint defined by the meaning conditions to be no more restrictive than what is necessary to prove Theorem 2. There are however reasons to consider further restrictions on the meaning conditions. Consider the deduction

$$\dfrac{\dfrac{m_1 : A}{\lambda x.x : \forall x A}\,\forall\mathrm{I}}{m_2 : A}\,\forall\mathrm{E}$$

Assume that x does not occur free in A. Then the constraint defined by the meaning conditions can be strengthened so that m_2 and m_1 are required to be syntactically equal. More generally, if x occurs free in A we can strengthen the constraint defined by the meaning conditions so that, in an informal sense, if one "submeaning" of m_2 and one "submeaning" of m_1 "correspond" to the same subformula of A, and x does not occur free in this subformula, then these "submeanings" of m_1 and m_2, respectively, are required to be syntactically equal.

In the following we shall not assume this last restriction to be added. Of course, if Theorem 2 holds without this restriction added to the restrictions of the meaning conditions, then this theorem also is true with this restriction added.

All meaning condition schemata except the $\bot$E meaning condition schema define a relation between the meanings assigned to the premises and the conclusion of the inference. We can interpret this as follows: use of the $\bot$E inference schema says that nothing more is known about how the premise of an $\bot$E inference is derived other than that it is the premise of an $\bot$E inference. Instead of having $\bot$ primitively given in **N** we can define it by $\forall x(r \in x)$, where r is an arbitrary term. We then have the $\bot$E inference schema as a derived schema, derived as follows, where x is supposed to be chosen so that x does not occur free in A.

$$\dfrac{\dfrac{\lambda x.\epsilon x : \forall x(r \in x)}{\epsilon m : r \in \{x \mid A\}}\ \forall\mathrm{E}}{m : A}\ \in\mathrm{E}$$

Then if we also take *false* to be defined by $\lambda x.\epsilon x$ we have the $\bot$E meaning condition schema as a derived meaning condition schema, derived from the meaning condition schemata $\forall$E and $\in$E.

Lemma 2 *If a deduction $\mathscr{D}$ is non-self-contradictory and $\mathscr{D}$ reduces to $\mathscr{E}$ then also the deduction $\mathscr{E}$ is non-self-contradictory.*

The proof of Theorem 2 is similar to the proof of Theorem 1. To prove Theorem 1 we define a function * from the set of non-self-contradictory deductions in $\mathbf{N}_{-\forall\exists=}$ to the set of deductions in **P**. To prove Theorem 2 we shall instead defined a function * from the set of non-self-contradictory deductions in $\mathbf{N}_{-\exists=}$ to the set of deductions in $\mathbf{P^2}$, where $\mathbf{P^2}$ denotes the formal system of second order propositional logic. The language of $\mathbf{P^2}$ is the set of formulas, inductively defined as follows. The propositional variables $X, X_1, X_2, \ldots$ and $\bot$ are formulas, and if A and B are formulas, then $A \supset B$, $A \mathbin{\&} B$, $A \vee B$ and $\forall X A$ are formulas. The $\bot$, $\supset$, & and $\vee$ inference schemata are the same for $\mathbf{P^2}$ as for the formal system $\mathbf{N}_{-\forall\exists=}$. The $\forall$ inference schemata for $\mathbf{P^2}$ are the following.

$$\dfrac{\begin{matrix}\mathscr{D}\\ A\end{matrix}}{\forall X A}\ \forall\mathrm{I} \qquad\qquad \dfrac{\begin{matrix}\mathscr{D}\\ \forall X A\end{matrix}}{A[B/X]}\ \forall\mathrm{E}$$

We have the restriction on deductions in $\mathbf{P^2}$ that the variable designated X in the $\forall$I schema may not occur free in any open assumption in the deduction designated $\mathscr{D}$. The reduction rules for deductions in $\mathbf{P^2}$ are the same as the reduction

rules for deductions in $\mathbf{N}_{-\exists=}$ except that the substitution of a term for a variable in the $\forall$-reduction in $\mathbf{N}_{-\exists=}$ corresponds, in $\mathbf{P^2}$, to a substitution of a proposition for a propositional variable. We presuppose that the set of variables of $\mathbf{N}_{-\exists=}$ and the set of propositional variables of $\mathbf{P^2}$ have the same cardinality. Hence there is a one to one correspondence, * say, between the set of variables of $\mathbf{N}_{-\exists=}$ and the set of propositional variables of $\mathbf{P^2}$. For any variable x of $\mathbf{N}_{-\exists=}$ we let the propositional variable X of $\mathbf{P^2}$ denote x^*. The function * from the set of meanings to be assigned to formulas in deductions in the formal system $\mathbf{N}_{-\exists=}$ onto the set of formulas of $\mathbf{P^2}$ is defined as follows.

$$\begin{aligned}
x^* &\equiv X \\
\mathit{false} &\equiv \bot \\
(\epsilon m)^* &\equiv (\bot \supset \bot) \supset m^* \\
(m \Rightarrow n)^* &\equiv m^* \supset n^* \\
(m \wedge n)^* &\equiv m^* \,\&\, n^* \\
(m + n)^* &\equiv m^* \vee n^* \\
(\lambda x.m)^* &\equiv \forall X m^*
\end{aligned}$$

The function * is extended to a function from the set of sets of meanings to be assigned to formulas in deductions in the formal system $\mathbf{N}_{-\exists=}$ onto the set of sets of formulas of $\mathbf{P^2}$ by letting Γ^* denote the set of formulas A^* such that A belongs to Γ, for all sets Γ of meanings to be assigned to formulas in deductions in the formal system $\mathbf{N}_{-\exists=}$. In a similar way as in Sect. 4 we extend * once more, to a function from the set of non-self-contradictory deductions in $\mathbf{N}_{-\exists=}$ to the set of deductions in $\mathbf{P^2}$. To define this function we add the following clauses to the definition of the function * in Sect. 4.

$$\left(\dfrac{\begin{matrix}\mathscr{D}\\ m : A\end{matrix}}{\lambda x.m : \forall x A}\,\forall\mathrm{I}\right)^* \equiv \dfrac{\begin{matrix}\mathscr{D}^*\\ m^*\end{matrix}}{\forall X m^*}\,\forall\mathrm{I}$$

$$\left(\dfrac{\begin{matrix}\mathscr{D}\\ \lambda x.m : \forall x A\end{matrix}}{m[n/x] : A[t/x]}\,\forall\mathrm{E}\right)^* \equiv \dfrac{\begin{matrix}\mathscr{D}^*\\ \forall X m^*\end{matrix}}{m^*[n^*/X]}\,\forall\mathrm{E}$$

The definition of $a(\theta, \mathscr{E}, \mathscr{D})$, given in Sect. 4, extends from deductions in $\mathbf{N}_{-\forall\exists=}$ to deductions in $\mathbf{N}_{-\exists=}$ by defining $a(\theta, \mathscr{E}, \mathscr{D})$ also in the case $\mathscr{D}$ reduces to $\mathscr{E}$ via an $\forall$ reduction. This is done in a similar way as for the other cases of the kind of reduction that takes $\mathscr{D}$ to $\mathscr{E}$.

Proposition 2 *Assume that $\mathscr{D}$ is a non-self-contradictory deduction, θ is an assignment of formal meanings to the formula occurrences in $\mathscr{D}$ such that this assignment satisfies the meaning conditions and $\mathscr{D} \Longrightarrow \mathscr{E}$. Let $\mathscr{D}$ also denote the deduction obtained from $\mathscr{D}$ by decorating the formula occurrences in $\mathscr{D}$ with θ. Let $\mathscr{E}$ also*

denote the deduction obtained from $\mathscr{E}$ by decorating the formula occurrences in $\mathscr{E}$ with $a(\theta, \mathscr{E}, \mathscr{D})$. Then $\mathscr{D}^* \Longrightarrow \mathscr{E}^*$.

From Girard (1971) [4] it is known that deductions in $\mathbf{P^2}$ are strongly normalizable; see also Martin-Löf (1971) [5]. From this together with Proposition 2, Theorem 2 follows.

6 Self-contradictory Reasoning in $\mathbf{N}_{-=}$

The meaning conditions associated with $\mathbf{N}_{-=}$ are defined by adding to the meaning conditions associated with $\mathbf{N}_{-\exists=}$ some constraints given by informally considering $\exists x A$ to represent the informally given infinitely long formula $A[t_1/x] \vee (A[t_2/x] \vee (A[t_3/x] \vee \ldots))$, where $t_1, t_2, t_3, \ldots$ are all terms of the formal system $\mathbf{N}_{-=}$. The set of formal meanings to be assigned to formulas in deductions in the formal system $\mathbf{N}_{-=}$ is inductively defined as follows. The meaning variable x and *false* are meanings, and if m and n are meanings, then ϵm, $m \Rightarrow n$, $m \wedge n$, $\lambda x.m$ and $\mu x.m$ are meanings. The meaning conditions associated with the formal system $\mathbf{N}_{-=}$ are the following, and in addition the meaning conditions associated with the formal system $\mathbf{N}_{-\exists=}$.

$$\dfrac{\begin{array}{c}\mathscr{D}\\ m[n/x] : A[t/x]\end{array}}{\mu x.m : \exists x A}\ \exists\mathrm{I} \qquad\qquad \dfrac{\begin{array}{c}\mathscr{D}\\ \mu x.m : \exists x A\end{array} \quad \begin{array}{c}[m : A]\\ \mathscr{E}\\ n : C\end{array}}{n : C}\ \exists\mathrm{E}$$

We have the restriction on the meaning variable designated x in the $\exists E$ meaning condition schema that neither may it occur free in the meaning designated n assigned to the subsequent premise of the $\exists E$ nor may it occur free in any meaning assigned to an open assumption of the deduction of the subsequent premise $\mathscr{E}$ other than the open assumption designated A.

Theorem 3 *Every non-self-contradictory deduction in $\mathbf{N}_{-=}$ is normalizable.*

Let **PR** be the formal system with the same language as $\mathbf{N}_{-=}$, obtained by removing the $\in$-inferences from **N**. We have the following result concerning **PR**.

Proposition 3 *Every deduction in* **PR** *is non-self-contradictory.*

Proof Let $\mathscr{D}$ be any given deduction in **PR**. We shall define an assignment of formal meanings to the formulas in $\mathscr{D}$ such that this assignment satisfies the meaning conditions. This assignment is defined by decorating every formula occurrence A in $\mathscr{D}$ with the formal meaning A°, where $^\circ$ is a function from the set of formulas of **PR** to the set of formal meanings to be assigned to formulas in deductions in the formal system $\mathbf{N}_{-=}$. The bijection $^\circ$ is defined as follows.

$$(r \in x)^\circ \equiv \epsilon x$$

$$
\begin{aligned}
(r \in \{x \mid A\})^\circ &\equiv \epsilon A^\circ \\
\bot^\circ &\equiv \mathit{false} \\
(A \supset B)^\circ &\equiv A^\circ \Rightarrow B^\circ \\
(A \,\&\, B)^\circ &\equiv A^\circ \wedge B^\circ \\
(A \vee B)^\circ &\equiv A^\circ + B^\circ \\
(\forall x A)^\circ &\equiv \lambda x.A^\circ \\
(\exists x A)^\circ &\equiv \mu x.A^\circ
\end{aligned}
$$

□

Appendix

Naïve Set Theory

We present the system **N** of natural deduction for naïve set theory. The syntactic categories of the language of **N** are

1. *Variables*, x, y, z
2. *Terms*, r, s, t, u, v, w
3. *Formulas*, $A, B, C, \ldots$

The *language* of **N** is the set of terms and formulas, inductively defined as follows. Variables x, y and z are terms, and if A is a formula, then $\{x \mid A\}$ is a term. If r and s are terms, then $r = s$ and $r \in s$ are formulas, and if A and B are formulas, then $A \supset B$, $A \,\&\, B$, $\forall x A$, $A \vee B$ and $\exists x A$ are formulas; $\bot$ is also a formula. The symbols used in the language of a formal system are the *primitive symbols* of that formal system. In addition to the primitive symbols of **N**, we shall use the following defined symbols

$$
\begin{aligned}
\neg A &\equiv A \supset \bot \\
r \notin s &\equiv \neg(r \in s)
\end{aligned}
$$

We use $\mathscr{D}, \mathscr{E}, \mathscr{F}, \ldots$ to denote deductions. The deductions in **N** are defined by the following inference schemata.

$$\dfrac{\begin{array}{c}\mathscr{D}\\ A[t/x]\end{array}}{t \in \{x \mid A\}}\ \in\mathrm{I} \qquad \dfrac{\begin{array}{c}\mathscr{D}\\ t \in \{x \mid A\}\end{array}}{A[t/x]}\ \in\mathrm{E}$$

$$\dfrac{\begin{array}{c}\mathscr{D}\\ \bot\end{array}}{A}\ \bot\mathrm{E}$$

$$\dfrac{\begin{array}{c}[A]\\ \mathscr{D}\\ B\end{array}}{A \supset B}\ \supset\mathrm{I} \qquad \dfrac{\begin{array}{c}\mathscr{D}\\ A \supset B\end{array} \quad \begin{array}{c}\mathscr{E}\\ A\end{array}}{B}\ \supset\mathrm{E}$$

$$\dfrac{\begin{array}{c}\mathscr{D}\\ A\end{array} \quad \begin{array}{c}\mathscr{E}\\ B\end{array}}{A \mathbin{\&} B}\ \&\mathrm{I} \qquad \dfrac{\begin{array}{c}\mathscr{D}\\ A \mathbin{\&} B\end{array}}{A}\ \&\mathrm{E} \qquad \dfrac{\begin{array}{c}\mathscr{D}\\ A \mathbin{\&} B\end{array}}{B}\ \&\mathrm{E}$$

$$\dfrac{\begin{array}{c}\mathscr{D}\\ A\end{array}}{\forall x\, A}\ \forall\mathrm{I} \qquad \dfrac{\begin{array}{c}\mathscr{D}\\ \forall x\, A\end{array}}{A[t/x]}\ \forall\mathrm{E}$$

$$\dfrac{\begin{array}{c}\mathscr{D}\\ A[t/x]\end{array}}{\exists x\, A}\ \exists\mathrm{I} \qquad \dfrac{\begin{array}{c}\mathscr{D}\\ \exists x\, A\end{array} \quad \begin{array}{c}[A]\\ \mathscr{E}\\ C\end{array}}{C}\ \exists\mathrm{E}$$

$$\dfrac{\begin{array}{c}\mathscr{D}\\ A\end{array}}{A \vee B}\ \vee\mathrm{I} \qquad \dfrac{\begin{array}{c}\mathscr{D}\\ B\end{array}}{A \vee B}\ \vee\mathrm{I} \qquad \dfrac{\begin{array}{c}\mathscr{D}\\ A_1 \vee A_2\end{array} \quad \begin{array}{c}[A_1]\\ \mathscr{E}\\ C\end{array} \quad \begin{array}{c}[A_2]\\ \mathscr{F}\\ C\end{array}}{C}\ \vee\mathrm{E}$$

$$\dfrac{\begin{array}{c}[x \in r]\\ \mathscr{D}\\ x \in t\end{array} \quad \begin{array}{c}[y \in t]\\ \mathscr{E}\\ y \in r\end{array}}{r = t}\ =\mathrm{I} \qquad \dfrac{\begin{array}{c}\mathscr{D}\\ r = t\end{array} \quad \begin{array}{c}\mathscr{E}\\ A[r/x]\end{array}}{A[t/x]}\ =\mathrm{E} \qquad \dfrac{\begin{array}{c}\mathscr{D}\\ r = t\end{array} \quad \begin{array}{c}\mathscr{E}\\ A[t/x]\end{array}}{A[r/x]}\ =\mathrm{E}$$

An *inference* is an application of an inference schema. An *atomic formula* is a formula that cannot be the conclusion of an introduction inference. In an elimination inference the leftmost premise is the *major premise* and all other premises, if there are any, are the *minor premises*. A *proof* is a deduction without open assumptions. A *subdeduction* is defined to be an occurrence of a subdeduction in a deduction.

The variable x in the $\forall$I and $\exists$E schemata and the variables x and y in the $=$I schema designate *eigenvariables* of inferences. We require that the eigenvariables occurring in a deduction $\mathscr{D}$ are syntactically distinguished from each other and from variables with free non-eigenvariable occurrences in $\mathscr{D}$.

For a treatment of the basic concepts of natural deduction the reader is referred to Gentzen (1969) [3] and Prawitz (1965, 1971) [6, 7].

Normal Deductions in a Fragment of **N**

Let **F** be a formal system. We consider a *fragment* of **F** to be a formal system obtained from **F** by removing some primitive symbols and the corresponding inference schemata. To begin with we look at normal deductions in the formal system obtained by removing the symbols $\exists$, $\vee$ and $=$ and the inference schemata corresponding to these symbols from **N**. We let $\mathbf{N}_{-\exists\vee=}$ denote this formal system.

In addition to the uniqueness of names of eigenvariables we have restrictions on deductions concerning the *scopes of eigenvariables*. The scope of an eigenvariable in a deduction $\mathscr{D}$ is the subdeduction of $\mathscr{D}$ in which the eigenvariable is defined. The scope of an eigenvariable of an $\forall$I inference is the premise deduction of the inference. We have the restriction on deductions that an eigenvariable of an inference may not occur free in any open assumption in the scope of the eigenvariable other than assumptions cancelled at the inference.

Definition 2 In $\mathbf{N}_{-\exists\vee=}$ a *cut* is formula occurrence which is both the conclusion of an introduction inference and the major premise of an elimination inference. A *normal* deduction is a deduction containing no cut.

Definition 3 A *branch* in a deduction $\mathscr{D}$ is a sequence $A_1, A_2, \ldots, A_n$ of formula occurrences in $\mathscr{D}$ such that: (1) A_1 is an assumption. (2) For each i such that $1 \leq i < n$, A_i stands immediately above A_{i+1} and A_i is not the minor premise of an elimination inference. (3) A_n is the end formula of the deduction or the minor premise of an elimination inference. An *E-part of a branch* is a sequence of consecutive formulas of the branch, none of which is the conclusion of an introduction inference. An *I-part of a branch* is a sequence of consecutive formulas of the branch, all of which are the conclusions of introduction inferences. A *main branch* is a branch $A_1, A_2, \ldots, A_n$, with A_n as the end formula of the deduction. An *E-main branch* is a main branch consisting only of an E-part. Note that there cannot be more than one E-main branch in a deduction.

If a formula occurrence in a deduction in $\mathbf{N}_{-\exists\vee=}$ is a minor premise of an elimination inference then this formula occurrence is the minor premise of an $\supset$E inference. The reason that the phrase *the minor premise of an elimination inference* is used in the definition of a branch above is to make the definition applicable to deductions in other formal systems, where it is not the case that a minor premise of an elimination inference always is the minor premise of an $\supset$E inference.

Proposition 4 *Every branch in a normal deduction in* $\mathbf{N}_{-\exists\vee=}$ *consists of an E-part followed by a (possibly empty) I-part.*

Proposition 5 *A normal proof in* $\mathbf{N}_{-\exists\vee=}$ *has an introduction inference as its last inference.*

Reductions of Deductions in $\mathbf{N}_{-=}$

We use $\mathscr{D} \Longrightarrow \mathscr{E}$ to denote that $\mathscr{D}$ *reduces* to the deduction $\mathscr{E}$. If there is a deduction $\mathscr{E}$ such that $\mathscr{D} \Longrightarrow \mathscr{E}$ then $\mathscr{D}$ is *reducible*. $\mathscr{D}$ reduces in zero steps to itself. If there are deductions $\mathscr{E}_1, \ldots, \mathscr{E}_n$, where $n \geq 1$, such that

$$\mathscr{D} \Longrightarrow \mathscr{E}_1 \Longrightarrow \ldots \Longrightarrow \mathscr{E}_n$$

then $\mathscr{D}$ *reduces in n steps* to the deduction $\mathscr{E}_n$. Hence, the two phrases $\mathscr{D}$ *reduces in one step to* $\mathscr{E}$ and $\mathscr{D}$ *reduces to* $\mathscr{E}$ have the same meaning. If there is an $n \geq 0$ and a deduction $\mathscr{E}$ such that $\mathscr{D}$ reduces in n steps to $\mathscr{E}$, and $\mathscr{E}$ is not reducible, then $\mathscr{D}$ is *normalizable*. If there is no infinite family $\{\mathscr{E}_i\}$, $i = 1, 2, 3, \ldots$ of deductions such that $\mathscr{D} \Longrightarrow \mathscr{E}_1$ and $\mathscr{E}_i \Longrightarrow \mathscr{E}_{i+1}$, for $i \geq 1$, then $\mathscr{D}$ is *strongly normalizable*.

The relation $\Longrightarrow$ is defined inductively, by the schemata below. Notice that a deduction is reducible only if it has a cut and that the reduction defined removes the cut.

Epsilon reduction

$$\dfrac{\dfrac{\begin{matrix}\mathscr{D}\\ A[t/x]\end{matrix}}{t \in \{x \mid A\}}\,\in\mathrm{I}}{A[t/x]}\,\in\mathrm{E} \quad \Longrightarrow \quad \begin{matrix}\mathscr{D}\\ A[t/x]\end{matrix}$$

Imply reduction

$$\dfrac{\dfrac{\begin{matrix}[A]\\ \mathscr{D}\\ B\end{matrix}}{A \supset B}\,\supset\mathrm{I} \qquad \begin{matrix}\mathscr{E}\\ A\end{matrix}}{B}\,\supset\mathrm{E} \quad \Longrightarrow \quad \begin{matrix}\mathscr{E}\\ A\\ \mathscr{D}\\ B\end{matrix}$$

And reduction

$$\dfrac{\dfrac{\begin{matrix}\mathscr{D}\\ A\end{matrix} \qquad \begin{matrix}\mathscr{E}\\ B\end{matrix}}{A \,\&\, B}\,\&\mathrm{I}}{A}\,\&\mathrm{E} \quad \Longrightarrow \quad \begin{matrix}\mathscr{D}\\ A\end{matrix}$$

$$\dfrac{\dfrac{\begin{matrix}\mathscr{D}\\ A\end{matrix} \qquad \begin{matrix}\mathscr{E}\\ B\end{matrix}}{A \,\&\, B}\,\&\mathrm{I}}{B}\,\&\mathrm{E} \quad \Longrightarrow \quad \begin{matrix}\mathscr{E}\\ B\end{matrix}$$

Or reduction

$$\dfrac{\dfrac{\begin{matrix}\mathscr{D}\\ A\end{matrix}}{A \vee B}\,\vee\mathrm{I} \qquad \begin{matrix}[A]\\ \mathscr{E}\\ C\end{matrix} \qquad \begin{matrix}[B]\\ \mathscr{F}\\ C\end{matrix}}{C}\,\vee\mathrm{E} \quad \Longrightarrow \quad \begin{matrix}\mathscr{D}\\ A\\ \mathscr{E}\\ C\end{matrix}$$

$$\dfrac{\dfrac{\begin{matrix}\mathscr{D}\\ B\end{matrix}}{A \vee B}\,\vee\mathrm{I} \qquad \begin{matrix}[A]\\ \mathscr{E}\\ C\end{matrix} \qquad \begin{matrix}[B]\\ \mathscr{F}\\ C\end{matrix}}{C}\,\vee\mathrm{E} \quad \Longrightarrow \quad \begin{matrix}\mathscr{D}\\ B\\ \mathscr{F}\\ C\end{matrix}$$

Exist reduction

$$\dfrac{\dfrac{\begin{matrix}\mathscr{D}\\ A[t/x]\end{matrix}}{\exists x A}\,\exists\mathrm{I} \qquad \begin{matrix}[A]\\ \mathscr{E}\\ C\end{matrix}}{C}\,\exists\mathrm{E} \quad \Longrightarrow \quad \begin{matrix}\mathscr{D}\\ A[t/x]\\ \mathscr{E}[t/x]\\ C\end{matrix}$$

For all reduction

$$\dfrac{\dfrac{\begin{matrix}\mathscr{D}\\ A\end{matrix}}{\forall x A}\,\forall\mathrm{I}}{A[t/x]}\,\forall\mathrm{E} \quad \Longrightarrow \quad \begin{matrix}\mathscr{D}[t/x]\\ A[t/x]\end{matrix}$$

For two further reductions (*Left Compose* and *Subderivation*) see Ekman (1994) [2, Sect. 4.1].

Propositional Logic

Propositional logic is the formal system **P** obtained by removing the symbols $\in$, $\forall$, $\exists$ and $=$ and the inference schemata corresponding to these symbols from **N**. The formal system **P** does not have any term variables but instead propositional variables. The *language* of **P** is the set of formulas, inductively defined as follows. The *propositional variables* P, Q, R and $\perp$ are *formulas*, and if A and B are formulas, then $A \supset B$, $A \,\&\, B$ and $A \vee B$ are *formulas*.

A branch in a deduction in **P** is defined as a branch in a deduction in $\mathbf{N}_{-\exists\vee=}$. The notion of a cut in a deduction in **P** and the notion of a normal deduction in **P** is defined as in $\mathbf{N}_{-=}$. The definitions of an E-part of a branch, an I-part of a branch, a main branch and an E-main branch are the same for a branch in a deduction in **N** as for a branch in a deduction in $\mathbf{N}_{-\exists\vee=}$.

References

1. Crabbé, M.: Non-normalisation de la théorie de Zermelo. http://www.logic-center.be/Publications/Bibliotheque/contreexemple.pdf (sketch of the result presented at the Proof Theory Symposium held in Kiel in 1974)
2. Ekman, J.: Normal proofs in set theory. Ph.D. thesis, Department of Computing Science, University of Göteborg (1994)
3. Gentzen, G.: Untersuchungen über das logische Schließen. Mathematische Zeitschrift, 39, 176-210, 405-431 (1934/35) (English translation in: The Collected Papers of Gerhard Gentzen (Szabo, M.E. (ed.). Amsterdam, North Holland (1969), pp. 68-131)
4. Girard, J.-Y.: Une extension de l'interprétation de Gödel à l'analyse, et son application à l'élimination des coupures dans l'analyse et la théorie des types. In: Fenstad, J.E. (ed.) Proceedings of the Second Scandinavian Logic Symposium, pp. 63-92. North-Holland, Amsterdam (1971)
5. Martin-Löf, P.: Hauptsatz for the theory of species. In: Fenstad, J.E. (ed.) Proceedings of the Second Scandinavian Logic Symposium, pp. 217-233. North-Holland, Amsterdam (1971)
6. Prawitz, D.: Natural Deduction: A Proof-Theoretical Study. Almqvist & Wiksell, Stockholm (1965). Reprinted Dover Publ., Mineola 2006
7. Prawitz, D.: Ideas and results in proof theory. In: Fenstad, J.E. (ed.) Proceedings of the Second Scandinavian Logic Symposium (Oslo 1970), pp. 235–307. North-Holland, Amsterdam (1971)

Completeness in Proof-Theoretic Semantics

Thomas Piecha

Abstract We give an overview of completeness and incompleteness results within proof-theoretic semantics. Completeness of intuitionistic first-order logic for certain notions of validity in proof-theoretic semantics has been conjectured by Prawitz. For the kind of semantics proposed by him, this conjecture is still undecided. For certain variants of proof-theoretic semantics the completeness question is settled, including a positive result for classical logic. For intuitionistic logic there are positive as well as negative completeness results, depending on which variant of semantics is considered. Further results have been obtained for certain fragments of first-order languages.

Keywords Completeness · Proof-theoretic validity · Intuitionistic logic · Classical logic · Atomic systems

1 Introduction

In proof-theoretic semantics (see Schroeder-Heister [34]; cf. Wansing [36]) for logical constants several related notions of validity have been proposed. We mention Kreisel (cf. Gabbay [6]), Prawitz [18–22], Dummett [3] and Sandqvist [26]. Overviews and discussions of such proof-theoretic notions of validity can be found in Schroeder-Heister [31] and Read [24].

What these notions of validity have in common is that the validity of an atomic formula, or atom, is defined in terms of the derivability of that atom in a given system of atomic rules, that is, of rules which can only contain atoms. Let $a, b, \ldots, a_1, a_2, \ldots$ be atoms. Then

$$\frac{}{a} \qquad \frac{}{b} \qquad \frac{a \qquad b}{c}$$

T. Piecha (✉)
Department of Computer Science, University of Tübingen, Tübingen, Germany
e-mail: thomas.piecha@uni-tuebingen.de

T. Piecha and P. Schroeder-Heister (eds.), *Advances in Proof-Theoretic Semantics*,
Trends in Logic 43, DOI 10.1007/978-3-319-22686-6_15

is an example of a system S of atomic rules (the first two having the form of atomic axioms), in which c is derivable by

$$\frac{\overline{a} \qquad \overline{b}}{c}$$

and therefore valid with respect to S. Atomic rules are also called boundary rules (cf. Dummett [3]) or production rules. Atomic systems S are also called bases; they can have the form of Post systems, definite Horn clause logic programs etc.

The validity of complex formulas $A, B, \ldots, A_1, A_2, \ldots$ (constructed as usual from atoms with logical constants) with respect to an atomic system S can then be defined inductively by giving semantic clauses for the logical constants. The validity of implications $A \to B$ with respect to an atomic system S is usually defined by taking into account arbitrary atomic extensions S' of S. Let $\vDash_S$ stand for 'valid with respect to S'; then the semantic clause for implication has the form

$$\vDash_S A \to B \;:\Longleftrightarrow\; \forall S' \supseteq S : (\vDash_{S'} A \implies \vDash_{S'} B)$$

where in the definiens all extensions S' of S have to be considered. This ensures that implications $A \to B$ cannot become valid with respect to S just because some atom on which the validity of A depends is not derivable in S. Considering extensions thus guarantees monotonicity for validity with respect to S.

It was conjectured by Prawitz [19, 22] that intuitionistic first-order logic is complete with respect to certain notions of validity for inference rules. This conjecture is still undecided. There are, however, several negative as well as positive results about completeness for certain plausible variants of this notion of validity, formulated not for inference rules but for formulas. One kind of variants considers only certain fragments of first-order languages. Other variants are based on different kinds of atomic systems which allow for atomic rules of a more general form than production rules only. Further variants are given through different treatments of negation or absurdity, and by different notions of what an extension of an atomic system is.

In the following, we present several of these variants together with their respective completeness or incompleteness results.

2 Prawitz's Conjecture

Prawitz has given several definitions of proof-theoretic validity (see Prawitz [18–22]), and he has conjectured completeness of intuitionistic first-order logic for some of them. We here present a formulation for the fragment $\{\to, \vee, \wedge\}$ as given by Schroeder-Heister [33], which captures the main ideas underlying Prawitz's definitions. The restriction to the fragment $\{\to, \vee, \wedge\}$ is only made to keep the exposition simple; the definitions can be extended to the first-order case in a more or less straightforward way.

We first define some preliminary notions:

Definition 1 A *(first-level) atomic system S* is a (possibly empty) set of atomic rules of the form

$$\frac{a_1 \qquad \dots \qquad a_n}{b}$$

where the a_i and b are atoms. The set of premises $\{a_1, \dots, a_n\}$ in a rule can be empty; in this case the rule is an *atomic axiom* and of *level* 0. First-level atomic systems that do not contain atomic axioms are called *proper first-level atomic systems*.

Definition 2 An *arbitrary inference rule* has the form

$$\frac{\begin{array}{c}[A_{11}, \dots, A_{1m_1}]\\ B_1\end{array} \qquad \dots \qquad \begin{array}{c}[A_{n1}, \dots, A_{nm_n}]\\ B_n\end{array}}{C}$$

The notation is the same as the one used for the logical rules of natural deduction (see Gentzen [7]). That is, rules of this form allow one to conclude C from the set of premises $\{B_1, \dots, B_n\}$ and to discharge any of the assumptions A_{ij}, written in square brackets [ ], on which premises B_i might depend.

Definition 3 A *derivation structure* is a derivation tree composed of arbitrary inference rules. (Derivation structures correspond to what Prawitz calls '(argument or proof) schemata' or '(argument or proof) skeletons'.)

The notions *open/closed* and *canonical/non-canonical* as used for derivations in natural deduction are carried over to derivation structures. That is, a derivation structure with no open assumptions is *closed*, otherwise *open*. It is *canonical*, if it ends with one of the introduction rules

$$\frac{\begin{array}{c}[A]\\ B\end{array}}{A \to B} \qquad \frac{A_i}{A_1 \vee A_2}\ (i = 1 \text{ or } 2) \qquad \frac{A \qquad B}{A \wedge B}$$

It is *non-canonical*, if it does not.

Definition 4 A *reduction procedure* transforms a given derivation structure into another derivation structure.

A *justification* J of an arbitrary inference rule R, excluding introduction rules, is a set of reduction procedures which transform derivation structures $\mathscr{D}$ ending with an application of R into another derivation structure with the same end formula as $\mathscr{D}$ and having no more open assumptions than $\mathscr{D}$ (see Prawitz [22]).

Now *validity with respect to atomic systems S and justifications J* (short: (S, J)-*validity*) can be defined as follows:

Definition 5 (i) Every closed derivation in an atomic system S is (S, J)-*valid* (for every justification J).

(ii) A closed canonical derivation structure is (S, J)-*valid*, if all its immediate substructures are (S, J)-valid.

(iii) A closed non-canonical derivation structure is (S, J)-*valid*, if it reduces, with respect to J, to a canonical derivation structure, which is (S, J)-valid.

(iv) An open derivation structure

$$\begin{array}{ccc} A_1 & \dots & A_n \\ & \mathscr{D} & \\ & B & \end{array}$$

where all open assumptions of $\mathscr{D}$ are in $\{A_1, \dots, A_n\}$, is (S, J)-*valid*, if for every extension S' of S and every extension J' of J, and for every list of closed derivation structures $\begin{array}{c}\mathscr{D}_i \\ A_i\end{array}$ (for $1 \leq i \leq n$) which are (S', J')-valid, the derivation structure

$$\begin{array}{ccc} \mathscr{D}_1 & & \mathscr{D}_n \\ A_1 & \dots & A_n \\ & \mathscr{D} & \\ & B & \end{array}$$

is (S', J')-valid.

Extensions S' of S and J' of J are here understood in the set-theoretic sense as $S' \supseteq S$ and $J' \supseteq J$. Taking extensions into account ensures that (S, J)-validity of derivation structures is monotone with respect to extensions of S and J. This is an important constraint, if atomic systems S and justifications J are understood to represent, for example, states of knowledge.

In [18, Appendix A.1], Prawitz gave a definition of 'valid derivation', which makes use of extensions of atomic systems. However, in definitions of the more general notion of 'valid derivation structure' (i.e., of 'valid argument schema' or 'valid argument') he uses (consistent) extensions of justifications, but no extensions of atomic systems. Completeness of minimal logic for one such notion was conjectured in Prawitz [19]. A completeness conjecture for intuitionistic logic and a similar notion of validity is made in Prawitz [22]:

Conjecture 1 (Prawitz [22, p. 274]) *Every valid inference rule that can be formulated within first-order languages holds as a derivable inference rule within the system of natural deduction for intuitionistic logic.*

Prawitz's motivation for considering proof-theoretic notions of validity is to give an answer to the question of whether the elimination rules of Gentzen's intuitionistic system of natural deduction are the strongest possible ones justifiable in terms of the introduction rules of that system. Gentzen's idea that the introduction rules define the logical constants and that the elimination rules have to be justified on the basis of

the introduction rules (see Gentzen [7]; cf. [1]) is reflected in the notion of validity by the fact that priority is given to canonical derivation structures, that is, to derivation structures ending with an introduction rule, to which non-canonical derivation structures have to be reduced. The (as yet unsettled) completeness conjecture implies a positive answer to that question.

In [22], Prawitz also gives a further modification of the notion of validity with respect to the role played by justifications. We will not discuss this modification here. Moreover, in what follows we will focus on proof-theoretic notions of validity for *formulas* instead of validity for derivation structures or inference rules. This approach has the advantage that justifications J (i.e., sets of reduction procedures for derivation structures) do not need to be considered at all. We here only mention that certain notions of validity for inference rules were given in Schroeder-Heister [28, 30], and that intuitionistic logic was claimed to be complete with respect to them there.

3 Failure of Completeness for Intuitionistic Logic

Our first example of a notion of validity for formulas is due to Kreisel [10]. We follow the expositions given by Gabbay in [5] and [6, Chap. 13], adjust the notation and speak of 'Kreisel validity'.

Let $\mathscr{A}$ be a fixed alphabet and S a Post system on $\mathscr{A}$. If a word w over $\mathscr{A}$ is derivable in S, we write $\vdash_S w$. Let h be any function which assigns words over $\mathscr{A}$ to all variables $x, y, x_1, x_2, \ldots$ and relation symbols R of a first-order language, and let $h_1 =_x h_2 := h_1(y) = h_2(y)$ for all $y \neq x$.

Definition 6 *Kreisel S-validity* ($\vDash^h_S$) and *Kreisel validity* ($\vDash$) are defined as follows:

(K1) $\vDash^h_S R(x_1, \ldots, x_n) \;:\Longleftrightarrow\; \vdash_S h(R)h(x_1)\ldots h(x_n)$ (where $R(x_1, \ldots, x_n)$ is an atom),
(K2) $\vDash^h_S A \to B \;:\Longleftrightarrow\; \forall S' \supseteq S : (\vDash^h_{S'} A \implies \vDash^h_{S'} B)$,
(K3) $\vDash^h_S A \vee B \;:\Longleftrightarrow\; \vDash^h_S A$ or $\vDash^h_S B$,
(K4) $\vDash^h_S A \wedge B \;:\Longleftrightarrow\; \vDash^h_S A$ and $\vDash^h_S B$,
(K5) $\vDash^h_S \neg A \;:\Longleftrightarrow\;$ for all consistent $S' \supseteq S : \nvDash^h_{S'} A$ (where S' is consistent iff $\nvdash_{S'} w$ for some word w),
(K6) $\vDash^h_S \exists x A(x) \;:\Longleftrightarrow\;$ for some $h_1 =_x h : \vDash^{h_1}_S A(x)$,
(K7) $\vDash^h_S \forall x A(x) \;:\Longleftrightarrow\;$ for all $h_1 =_x h : \vDash^{h_1}_S A(x)$,
(K8) $\vDash A \;:\Longleftrightarrow\; \forall \mathscr{A}, S, h : \vDash^h_S A$.
(K9) A is *substitution-Kreisel-valid* $:\Longleftrightarrow$ all substitution instances of A are Kreisel valid (where substitutions are uniform substitutions of formulas for atoms in A).

Note that clause (K5) for negation is restricted to *consistent* extensions, and that extensions $S' \supseteq S$ are understood in the normal set-theoretic sense, that is, the Post system S' contains at least all the rules of the Post system S. Alternatively, extensions

S' of S can be understood to mean that the implication $\vdash_S w \implies \vdash_{S'} w$ holds for all words w over $\mathscr{A}$. In this latter case, Gabbay speaks of *weak validity*.

Intuitionistic first-order logic is neither complete for weak validity nor for Kreisel validity. Completeness already fails in the propositional case for both notions (we now consider weak validity and Kreisel validity restricted to the propositional fragment):

Theorem 1 (Gabbay [6, p. 224]) *Intuitionistic propositional logic is not complete for weak validity. The formula* $((\neg\neg A \to A) \to (\neg A \vee \neg\neg A)) \to (\neg A \vee \neg\neg A)$ *is a counterexample.*

Theorem 2 (Gabbay [6, p. 225]) *Intuitionistic propositional logic is not complete for Kreisel validity. The set of Kreisel valid sentences is not closed under substitution. The formula* $(a \to (b \vee c)) \to ((a \to b) \vee (a \to c))$*, for propositional atoms* a, b, c*, is a counterexample.*

Considering only the propositional fragment, completeness has been conjectured for substitution-Kreisel-validity:

Conjecture 2 (Gabbay [6, p. 226]) *Intuitionistic propositional logic is complete for substitution-Kreisel-validity (restricted to the propositional fragment).*

4 Goldfarb's Account of Dummett's Approach

Dummett [3, Chaps. 11–13] made an approach to proof-theoretic validity for inference rules (or arguments) which is similar to Prawitz's (cf. Sect. 2). It is supposed to yield a justification of intuitionistic first-order logic. Goldfarb [8] (this volume) has given an analysis of the propositional part of Dummett's approach, resulting in a notion of validity for formulas (instead of inference rules).

Goldfarb first gives a formulation for atomic systems of axioms only, that is, for sets of atoms. It is presumed that there are infinitely many atoms available and that only finite sets of atoms α, β are ever considered. We follow his notation in writing α, β for such sets but adjust it to ours otherwise:

Definition 7

(G1) $\alpha \vDash a \;:\Longleftrightarrow\; a \in \alpha$,
(G2) $\alpha \vDash A \to B \;:\Longleftrightarrow\; \forall \beta \supseteq \alpha : (\beta \vDash A \implies \beta \vDash B)$,
(G3) $\alpha \vDash A \vee B \;:\Longleftrightarrow\; \alpha \vDash A$ or $\alpha \vDash B$,
(G4) $\alpha \vDash A \wedge B \;:\Longleftrightarrow\; \alpha \vDash A$ and $\alpha \vDash B$,
(G5) There is no α such that $\alpha \vDash \bot$.

This notion of validity ($\vDash$) can be discarded right away, since it validates formulas which are not even derivable in classical logic (see Goldfarb [8]):

Lemma 1 *(i) Suppose A does not contain $\bot$. Then $\alpha \vDash (A \to \bot) \to \bot$.*
(ii) Let a and b be two distinct atoms. Then $\alpha \vDash (a \to b) \to b$.

Goldfarb then modifies this notion of validity by relativizing the relation $\vDash$ to proper first-level atomic systems S (i.e., in Dummett's terminology, to sets of boundary rules) as in Dummett's approach. He points out that in order to avoid cases like Lemma 1 (i), atomic rules with conclusion $\bot$ have to be allowed as well. The modified notion can be given by rewriting clauses (G1)–(G5) with $\vDash_S$ instead of $\vDash$, together with the condition that sets α, β have now to be closed under the rules in S and do not contain $\bot$:

Definition 8 Let S be a proper first-level atomic system. Let the sets α, β be closed under the rules in S, and $\bot \notin \alpha, \beta$.

(G1′) $\alpha \vDash_S a \;:\Longleftrightarrow\; a \in \alpha$,
(G2′) $\alpha \vDash_S A \to B \;:\Longleftrightarrow\; \forall \beta \supseteq \alpha : (\beta \vDash_S A \implies \beta \vDash_S B)$,
(G3′) $\alpha \vDash_S A \vee B \;:\Longleftrightarrow\; \alpha \vDash_S A$ or $\alpha \vDash_S B$,
(G4′) $\alpha \vDash_S A \wedge B \;:\Longleftrightarrow\; \alpha \vDash_S A$ and $\alpha \vDash_S B$,
(G5′) There is no α such that $\alpha \vDash_S \bot$.

According to Goldfarb, this notion of validity is a revision of Dummett's approach in that it considers in principle all atomic systems S instead of only a fixed one.

For this revised notion of validity all valid formulas are classically valid. Completeness for intuitionistic logic does not hold (see Goldfarb [8]):

Lemma 2 *(i) Every valid formula is derivable in classical logic.*
(ii) The formula $(a \to (B \vee C)) \to ((a \to B) \vee (a \to C))$ is valid for any atom a and any formulas B and C, but it is not intuitionistically derivable for all B, C.

The counterexamples to completeness given in Lemmas 1 and 2 are not schematic in the sense that all substitution instances of the valid formulas presented there are valid too. Goldfarb introduces the relation of *schematic validity*, which holds for a formula A if and only if all instances of A resulting from uniform substitutions of formulas for atoms in A are valid (cp. substitution-Kreisel-validity). He shows that the intuitionistically non-derivable formula $\neg A \vee \neg\neg A$ is schematically valid for atomic systems which do only contain atoms (i.e., for atomic systems of level 0). In other words:

Theorem 3 (Goldfarb [8]) *Intuitionistic logic is not complete for schematic validity for sets of atoms α (i.e., for the notion of schematic validity based on validity ($\vDash$) according to Definition 7).*

However, for the schematically understood revised notion of validity the following completeness result holds:

Theorem 4 (Goldfarb [8]) *Intuitionistic propositional logic is complete for schematic validity based on the revised notion of validity (i.e., for the notion of schematic validity based on validity ($\vDash_S$) according to Definition 8).*

We note that this completeness result depends on the restriction to consistent sets of atoms α, β in the sense that $\bot \notin \alpha, \beta$. A restriction to consistent extensions is also made in Definition 6 of (substitution-) Kreisel validity, namely in clause (K5) for negation. If negation is understood as $\neg A := A \to \bot$, and $\bot$ is explained by $\alpha \vDash_S \bot \;:\Longleftrightarrow\; \forall a : \alpha \vDash_S a$, then

$$\alpha \vDash_S \neg A \iff \forall \beta \supseteq \alpha : \beta \nvDash_S A.$$

Since α, β are consistent, this is equivalent to clause (K5), where $\bot$ is a word w such that $\nvdash_{S'} w$. However, in the case of (substitution-) Kreisel validity this is the only clause where a restriction to consistent atomic systems (resp. Post systems) S, S' is made, whereas such a restriction applies in general in the case of (schematic) validity according to Definition 8. Assuming consistent extensions in general also in the case of Kreisel validity implies completeness for substitution-Kreisel-validity. That is, Conjecture 2 is decided positively in this case.

5 Proof-Theoretic Validity for Generalized Atomic Systems

We now consider atomic systems which are not restricted to first-level atomic rules but which can contain atomic rules that can also discharge assumptions of a certain kind. One can show that intuitionistic logic is not complete for a notion of proof-theoretic validity based on such generalized atomic systems (see [16]).

To motivate such a generalization one might argue that since the device of assumption discharge is available at the level of logical rules (e.g., in the rules of implication introduction and disjunction elimination of natural deduction), it should be available at the level of atomic rules, too. However, from the point of view of attempting a justification of a certain logic by giving a semantics based on atomic systems, such a generalization might be conceived as being counterproductive, as it introduces a feature of implication already at the level of atomic rules.

5.1 Generalized Atomic Systems

We generalize the notion of first-level atomic system to higher-level atomic systems by allowing for atomic rules that can discharge atomic assumptions (cf. [16]).

Definition 9 A *second-level atomic system* S is a (possibly empty) set of atomic rules of the form

$$\frac{\begin{array}{c}[\Gamma_1]\\ a_1\end{array} \quad \dots \quad \begin{array}{c}[\Gamma_n]\\ a_n\end{array}}{b}$$

where the a_i and b are atoms, and the Γ_i are finite sets of atoms. The sets Γ_i may be empty, in which case the rule is a *first-level rule*. The set of premises $\{a_1, \ldots, a_n\}$ can be empty as well; in this case the rule is an axiom.

Such a rule can be applied as follows: If the premises $a_1, \ldots, a_n$ have been derived in S from certain assumptions, then one may conclude b, where, for each i, in the branch of the subderivation leading to a_i assumptions belonging to Γ_i may be discharged.

Second-level atomic systems are now further generalized to the higher-level case by allowing for atomic rules which can discharge not only atoms but atomic rules as assumptions (see Schroeder-Heister [29, 32] and Olkhovikov and Schroeder-Heister [15]; cf. [16]). We use the following linear notation for atomic higher-level rules:

Definition 10 (i) Every atom a is a rule of level 0.
(ii) If $R_1, \ldots, R_n$ are rules ($n \geq 1$), whose maximal level is ℓ, and a is an atom, then $(R_1, \ldots, R_n \rhd a)$ is a rule of level $\ell + 1$.

Definition 11 A *higher-level atomic system* S is a (possibly empty) set of atomic rules of the form

$$\frac{\begin{array}{c}[\Gamma_1]\\ a_1\end{array} \quad \ldots \quad \begin{array}{c}[\Gamma_n]\\ a_n\end{array}}{b}$$

(in linear notation: $(\Gamma_1 \rhd a_1), \ldots, (\Gamma_n \rhd a_n) \rhd b$), where the a_i and b are atoms, and the Γ_i are now finite sets $\{R_1^i, \ldots, R_k^i\}$ of rules, which may be empty. The set of premises $\{a_1, \ldots, a_n\}$ of such a rule can also be empty, in which case the rule is an axiom.

In the higher-level case atomic *rules* can be used as (dischargeable) assumptions, whereas in the second-level case only atoms could be used in that way. This difference requires a definition of the notion of *derivation* of an atom a from rules $R_1, \ldots, R_n$:

Definition 12 For a level-0 rule a,

$$\frac{}{a}\, a$$

is a *derivation* of a from $\{a\}$.

Now consider a level-$(\ell + 1)$ rule $(\Gamma_1 \rhd a_1), \ldots, (\Gamma_n \rhd a_n) \rhd b$. Suppose that for each i $(1 \leq i \leq n)$ a derivation

$$\begin{array}{c}\Sigma_i \cup \Gamma_i\\ \mathscr{D}_i\\ a_i\end{array}$$

of a_i from $\Sigma_i \cup \Gamma_i$ is given. Then

$$\dfrac{\begin{array}{ccc}\Sigma_1 & & \Sigma_n \\ \mathscr{D}_1 & & \mathscr{D}_n \\ a_1 & \dots & a_n\end{array}}{b}\ (\Gamma_1 \rhd a_1), \dots, (\Gamma_n \rhd a_n) \rhd b$$

is a *derivation* of b from $\Sigma_1 \cup \ldots \cup \Sigma_n \cup \{(\Gamma_1 \rhd a_1), \ldots, (\Gamma_n \rhd a_n) \rhd b\}$.

An atom b is *derivable* from Σ in a higher-level atomic system S, symbolically $\Sigma \vdash_S b$, if there is a derivation of b from $\Sigma \cup S$.

As an example, consider the atomic system $S = \{((b \rhd e) \rhd f), (((a \rhd b) \rhd c) \rhd e)\}$ and the set of assumptions $\Sigma = \{((a \rhd b) \rhd d), ((b, d) \rhd c)\}$. The following derivation shows that $\Sigma \vdash_S f$:

$$\dfrac{\dfrac{\dfrac{\dfrac{\dfrac{}{b}\,[b]^3 \qquad 1\,\dfrac{\dfrac{\dfrac{\dfrac{}{[a]^1}}{a}}{[a \rhd b]^2}}{b}}{d}}{b, d \rhd c}}{2\,\dfrac{c}{\langle((a \rhd b) \rhd c) \rhd e\rangle}}}{3\,\dfrac{e}{f}}\ \langle(b \rhd e) \rhd f\rangle$$

Angle brackets $\langle\,\rangle$ are used to indicate the rules of S, and square brackets [ ] with numerals indicate the discharge of assumptions.

5.2 Proof-Theoretic Validity

We now consider a notion of validity for intuitionistic propositional logic (see [16]), which will be based on the following clauses for the fragment $\{\rightarrow, \vee, \wedge\}$. Absurdity $\perp$ is taken as a distinguished atom. Extensions S' of atomic systems S are again understood in the set-theoretic sense: An atomic system S' is an *extension* of an atomic system S (written $S' \supseteq S$), if S' results from adding a (possibly empty) set of atomic rules to S.

Definition 13 *S-validity* ($\vDash_S$) and *validity* ($\vDash$) are defined as follows:

(S1) $\vDash_S a \;:\Longleftrightarrow\; \vdash_S a$,
(S2) $\vDash_S A \rightarrow B \;:\Longleftrightarrow\; A \vDash_S B$,
(S3) $\Gamma \vDash_S A \;:\Longleftrightarrow\; \forall S' \supseteq S : (\vDash_{S'} \Gamma \implies \vDash_{S'} A)$, where Γ is a set of formulas, and where $\vDash_{S'} \Gamma$ stands for $\{\vDash_{S'} A_i \mid A_i \in \Gamma\}$,
(S4) $\vDash_S A \vee B \;:\Longleftrightarrow\; \vDash_S A$ or $\vDash_S B$,
(S5) $\vDash_S A \wedge B \;:\Longleftrightarrow\; \vDash_S A$ and $\vDash_S B$,
(S6) $\Gamma \vDash A \;:\Longleftrightarrow\; \forall S : \Gamma \vDash_S A$.

Since only the logical constants of the fragment $\{\rightarrow, \vee, \wedge\}$ are considered, and $\perp$ is just an atom, one could also speak of *minimal validity* or *validity for minimal*

logic here. This notion is very similar to the 'minimal part' of Kreisel validity, given by clauses (K1)–(K4) and (K8) of Definition 6, when restricted to a propositional language and for words w identified with atoms a.

In analogy to substitution-Kreisel-validity, we define in addition *validity under substitution* as *validity for all substitution instances* (resulting from uniform substitutions of formulas for atoms). Thus validity under substitution is by definition closed under substitution.

Definition 14 *S-validity under substitution* ($\Vdash_S$) and *validity under substitution* ($\Vdash$) are defined as follows:

(i) $\Gamma \Vdash_S A \;:\Longleftrightarrow\;$ for each substitution instance Γ', A' of Γ, A: $\Gamma' \vDash_S A'$.
(ii) $\Gamma \Vdash A \;:\Longleftrightarrow\;$ for each substitution instance Γ', A' of Γ, A: $\Gamma' \vDash A'$.

These notions of validity are now extended for intuitionistic propositional logic:

Definition 15 *Intuitionistic S-validity* ($\vDash^i_S$) is defined as follows. Let ($\bot$) stand for the set of rules $\left\{\frac{\bot}{a} \,\middle|\, a \text{ atomic}\right\}$. Then $\Gamma \vDash^i_S A \;:\Longleftrightarrow\; \Gamma \vDash_{S\cup(\bot)} A$.

Correspondingly, $\Gamma \vDash^i A$, $\Gamma \Vdash^i_S A$ and $\Gamma \Vdash^i A$ are defined as $\Gamma \vDash_{(\bot)} A$, $\Gamma \Vdash_{S\cup(\bot)} A$ and $\Gamma \Vdash_{(\bot)} A$, respectively.

The treatment of absurdity $\bot$, and therefore of negation if understood as $\neg A := A \to \bot$, differs from the one given by clause (K5) of Kreisel validity and from the one given by clauses (G5) or (G5′). If $\bot$ were defined as a non-atomic constant by a semantical clause which says that there is no atomic system S such that $\vDash_S \bot$, then $\vDash_S \neg\neg a$ would hold for any atom a; this is the case, since $\nvDash_{S'} \neg a$ for any $S' \supseteq S$, as $\vDash_{S''} a$ for some $S'' \supseteq S'$.

We note the following properties of S-validity:

Lemma 3

(i) $\vDash_S$ *is a consequence relation, that is,*

(1) $A \vDash_S A$,
(2) $\Gamma \vDash_S A \implies \Gamma, \Delta \vDash_S A$,
(3) $(\Gamma \vDash_S A$ *and* $\Delta, A \vDash_S B) \implies \Gamma, \Delta \vDash_S B$.

(ii) $\vDash_S$ *is monotone with respect to S, that is,* $\Gamma \vDash_S A \implies \forall S' \supseteq S : \Gamma \vDash_{S'} A$.
(iii) $\Gamma \vDash_S A \to B \iff \Gamma, A \vDash_S B$.

For intuitionistic S-validity (i.e., for $\vDash_S$ *replaced with* $\vDash^i_S$*) these properties hold as well.*

Atomic rules can be represented by formulas and vice versa (for details see [16]). Let Σ^* stand for the set of formulas representing a finite set Σ of atomic rules. The following completeness and soundness result holds:

Lemma 4 $\Sigma^* \vDash_S a \iff \Sigma^* \vdash_S a$.

5.3 Failure of Strong Completeness

We now consider the system *NI* of natural deduction for intuitionistic propositional logic, for which one can show that it is not complete for validity.

Definition 16 Derivability of a formula A from a (possibly empty) set of assumptions Γ in *NI* is written $\Gamma \vdash A$.

Definition 17 (i) *Soundness* of *NI* means: $\Gamma \vdash A \implies \Gamma \vDash^i A$.
(ii) *Strong completeness* of *NI* means: $\Gamma \vDash^i A \implies \Gamma \vdash A$.
(iii) *Completeness* (simpliciter) of *NI* means: $\Gamma \Vdash^i A \implies \Gamma \vdash A$.

Soundness holds. Since derivability $\Gamma \vdash A$ in *NI* is closed under substitution, this implies $\Gamma \Vdash^i A$, that is, intuitionistic validity under substitution. The distinction between strong completeness and completeness (simpliciter) is useful, because one can show that validity is not closed under substitution; the given semantics validates a formula which is not derivable in *NI*. Thus strong completeness does not hold:

Theorem 5 *NI is not strongly complete. The set of valid formulas is not closed under substitution.*

Three proofs of this result are discussed in [16]. Here we only mention the counterexample (cf. also Goldfarb [8] and Sect. 4)

$$a \to (b \vee c) \vDash (a \to b) \vee (a \to c)$$

which is already a counterexample for strong completeness of minimal logic, and hence of *NI*. This counterexample is independent of the level of atomic systems. There are other counterexamples, for which this is not the case. For example, $\neg\neg a \vDash^i a$ holds for first-level atomic systems, but fails for atomic systems of levels higher than 1. Thus certain counterexamples in the realm of first-level atomic systems can be avoided by allowing for higher-level atomic systems. What the given counterexample therefore also shows is that strong completeness already fails for the (more standard) notion of validity based on first-level atomic systems.

5.4 Strong Completeness Results

Strong completeness holds for the fragment of disjunction-free formulas and for the fragment of arbitrary negative formulas $\neg A$ (see [16]):

Lemma 5 *Let Γ and A be disjunction-free. Then $\Gamma \vDash^i A \iff \Gamma \vdash A$.*

Lemma 6 *Let Γ and A be disjunction-free. Then $\Gamma \vDash A \iff \Gamma \vdash^m A$, where $\vdash^m$ denotes derivability in minimal logic. In other words, strong completeness holds for the $\{\to, \wedge\}$-fragment of minimal (and intuitionistic) logic (see Schroeder-Heister [33]).*

Lemma 7 *For any formula of the form $\neg A$ it holds that* $\models^i \neg A \iff \vdash \neg A$.

These results depend on higher-level atomic systems, for which Lemma 4 holds.

5.5 Failure of Completeness

Theorem 6 *Intuitionistic logic is not complete with respect to the semantics based on higher-level atomic systems.*

This has been proved in [16] by showing that the intuitionistically non-derivable Harrop or Kreisel–Putnam formula (see Harrop [9], Kreisel and Putnam [11]) is intuitionistically valid under substitution, that is, that

$$\Vdash^i (\neg A \to (B \vee C)) \to ((\neg A \to B) \vee (\neg A \to C))$$

holds. We emphasize that the given proof of this theorem depends on the fact that the considered semantics is based on higher-level atomic systems.

Since higher-level rules can be reduced to second-level rules by an appropriate coding (see Sandqvist [27]), it follows that intuitionistic logic is incomplete for S-validity based on second-level atomic systems. Whether intuitionistic logic is complete (simpliciter) for validity based on first-level atomic systems is an open problem.

Similarly to Gabbay's completeness conjecture for substitution-Kreisel-validity, the following conjecture can be made for intuitionistic validity under substitution:

Conjecture 3 *Intuitionistic propositional logic is complete (simpliciter) for intuitionistic validity based on first-level atomic systems. That is,* $\Gamma \Vdash^i A \implies \Gamma \vdash A$, *for first-level atomic systems only.*

5.6 Comparison with Kripke Semantics

Proof-theoretic validity shares some similarities with the notion of validity in Kripke semantics, which is sound and complete for intuitionistic logic (see Kripke [12]; cf. Moschovakis [14]). We mention that the semantical clauses for conjunction and disjunction have the same form in both cases, and that the clauses for implication are similar in that they depend on the idea of extensions. In Kripke semantics the clause for implication is

$$k \text{ forces } A \to B \;:\iff\; \forall k' \geq k : (k' \text{ forces } A \implies k' \text{ forces } B)$$

for nodes k, k' and partial orders $\geq$. The forcing relation for atoms a and nodes k is given by truth-value assignments $v(k, a)$, which obey the monotonicity requirement

that if $k' \geq k$ and $v(k, a) = \text{true}$, then $v(k', a) = \text{true}$. Thus k' is an extension of k in the sense that $\{a \mid k' \text{ forces } a\} \supseteq \{a \mid k \text{ forces } a\}$, just like $S' \supseteq S$ for atomic systems S, S' of level 0 in the case of proof-theoretic validity.

Besides these similarities, there are the following main differences to Kripke semantics. In proof-theoretic validity, the S-validity of atoms is given by their derivability in S, whereas in Kripke semantics the validity (resp. the forcing relation) for nodes k and atoms a is given by truth-value assignments $v(k, a)$.

In S-validity, atomic systems S are not only sets of atoms (which in Kripke semantics would be assigned to nodes k by v) but sets of atomic *rules*. This also means that $S' \supset S$ can be the case, although $\{a \mid \vdash_{S'} a\} = \{a \mid \vdash_S a\}$ (and consequently $\{a \mid \vDash_{S'} a\} = \{a \mid \vDash_S a\}$), simply because S' might contain inapplicable additional rules besides the ones in S, which therefore do not enlarge the set of atoms derivable in S'. For example, let S contain only the axiom a and let $S' = S \cup \left\{\frac{b}{c}\right\}$; then $S' \supset S$, while both in S' and S only a is derivable. A notion like weak validity (see Sect. 3), where

$$S' \text{ is an extension of } S \;:\Longleftrightarrow\; \forall a : (\vdash_S a \implies \vdash_{S'} a),$$

is in this respect closer to the notion of validity in Kripke semantics than to S-validity.

In Kripke semantics, a formula has to be forced by every node in every Kripke structure in order to be Kripke valid. Besides different sets of nodes k and different truth-value assignments $v(k, a)$, one therefore has to consider different partial orders $\geq$, whereas in proof-theoretic validity only one kind of structure is taken into account (cf. Goldfarb [8]; see also [16]), namely the one where the partial order is set inclusion $\supseteq$ for atomic systems S.

Furthermore, inconsistent extensions are possible in the case of S-validity, since absurdity $\perp$ could be added as an axiom to atomic systems S. This is not the case in Kripke semantics, where the forcing relation is consistent in the sense that a node k cannot force both A and $\neg A$ (cp., however, the modified Kripke models of Veldman [35]).

5.7 A Completeness Result for Intuitionistic Logic

A completeness result for intuitionistic propositional logic is available for the following notion of validity, which is given for second-level atomic systems S (see Sandqvist [27]; we adjust it to our notation):

Definition 18

(T1) $\vDash_S a \;:\Longleftrightarrow\; \vdash_S a$,
(T2) $\vDash_S A \to B \;:\Longleftrightarrow\; A \vDash_S B$,
(T3) $\Gamma \vDash_S A \;:\Longleftrightarrow\; \forall S' \supseteq S : (\vDash_{S'} \Gamma \implies \vDash_{S'} A)$, where Γ is a set of formulas, and where $\vDash_{S'} \Gamma$ stands for $\{\vDash_{S'} A_i \mid A_i \in \Gamma\}$,

(T4) $\vDash_S A \vee B \;:\Longleftrightarrow\; \forall S' \supseteq S$ and $\forall c : (A \vDash_{S'} c$ and $B \vDash_{S'} c \;\Longrightarrow\; \vDash_{S'} c)$,
(T5) $\vDash_S A \wedge B \;:\Longleftrightarrow\; \vDash_S A$ and $\vDash_S B$,
(T6) $\vDash_S \bot \;:\Longleftrightarrow\; \forall a : \vDash_S a$,
(T7) $\Gamma \vDash A \;:\Longleftrightarrow\; \forall S : \Gamma \vDash_S A$.

Compared to S-validity (see Definition 13) there are two differences (besides the restriction to second-level atomic systems S):

(i) Clause (T4) for disjunction replaces (S4). It resembles the natural deduction elimination rule for disjunction. Note that the definiens is restricted to extensions $S' \supseteq S$, and that propositional quantification is made use of in the universal quantification over all atoms c (not over all formulas; cf. Ferreira [4]).
(ii) Absurdity $\bot$ is not an atom but a logical constant, whose meaning is given by clause (T6). This clause is based on Dummett's introduction rule for $\bot$ (cf. Dummett [3, Chap. 13]).

Theorem 7 (Sandqvist [27]) *Intuitionistic propositional logic is sound and complete for this semantics, that is,* $\Gamma \vDash A \iff \Gamma \vdash A$.

6 Completeness Results for Classical Logic

So far, we have only discussed notions of proof-theoretic validity intended for intuitionistic logic or for certain fragments thereof. Now we will discuss a notion of proof-theoretic validity for classical logic.

Sandqvist [26] gives a semantics for the fragment $\{\rightarrow, \bot, \forall\}$ of the language of first-order logic. He considers basic sequents of the form $(\Gamma : a)$, which are relations between finite sets Γ of basic sentences and basic sentences a. Basic sentences are closed atomic formulas, that is, formulas containing neither logical constants nor free variables. Sets of basic sequents are called 'bases'. In our terminology, basic sequents are first-level rules, and bases are first-level atomic systems S. Sandqvist shows that minimal logic can be justified and that the law of double negation elimination is valid for the fragment $\{\rightarrow, \bot, \forall\}$. The other logical constants can then be defined, and a justification of classical logic is achieved without making use of the principle of bivalence. That classical logic is sound and complete for the given semantics is surprising, since this semantics is very similar to semantics proposed for intuitionistic logic. Discussions of these results can be found in Makinson [13] and in [2].

Sandqvist's semantics is the following (again, we use our notation):

Definition 19

(C1) For closed atoms a: $\vDash_S a \;:\Longleftrightarrow\;$ every set of closed atoms which is closed under S contains a.
(C2) For non-empty Γ: $\Gamma \vDash_S A \;:\Longleftrightarrow\; \vDash_S A$ for every $S' \supseteq S$ such that $\vDash_{S'} B$ for every $B \in \Gamma$.
(C3) $\vDash_S A \rightarrow B \;:\Longleftrightarrow\; A \vDash_S B$.

(C4) $\vDash_S \bot \;:\Longleftrightarrow\; \vDash_S a$ for every closed atom a.
(C5) $\vDash_S \forall x A(x) \;:\Longleftrightarrow\; \vDash_S A(x)[x/t]$ for every closed term t.
(C6) $\Gamma \vDash A \;:\Longleftrightarrow\; \forall S : \Gamma \vDash_S A$.
(C7) $\Gamma \vDash^* A \;:\Longleftrightarrow\; \Gamma\sigma \vDash A\sigma$ for all ground substitutions σ.

Note that the definiens in clause (C1) could be expressed equivalently as $\vdash_S a$. Another (equivalent) formulation has been given by Makinson [13], where $\underline{S}(\Delta)$ is written for the closure of a set Δ of closed atoms under the rules in S. That is, $\underline{S}(\Delta)$ is the intersection of all sets Λ of closed atoms such that $\Delta \subseteq \Lambda$, and if $\dfrac{a_1 \quad \dots \quad a_n}{b} \in S$ with $\{a_1, \dots, a_n\} \subseteq \Lambda$, then $b \in \Lambda$. Clauses (C1) and (C4) can then be written as follows:

(C1′) For closed atoms a: $\vDash_S a \;:\Longleftrightarrow\; a \in \underline{S}(\emptyset)$.
(C4′) $\vDash_S \bot \;:\Longleftrightarrow\; a \in \underline{S}(\emptyset)$ for every closed atom a.

We point out that $\bot$ is *not* an atom here. In clause (C5), the notation $A(x)[x/t]$ means that each occurrence of x in A is replaced by the term t. The relation $\Gamma \vDash^* A$ defined in clause (C7) deals with open formulas; a ground substitution is a substitution of variable-free terms for variables. The sets Γ of formulas are finite, but in Definition 19 infinite sets Γ could be allowed as well. The relation $\vDash_S$ is called 'valid inferability' by Sandqvist; by 'validity' we refer to the relation $\vDash$ defined in clause (C6).

The given semantics validates minimal logic (see Sandqvist [26, Lemma 3]). Furthermore, Sandqvist [26, Lemma 4] shows that the law of double negation elimination holds: $(A \to \bot) \to \bot \vDash^* A$. Since minimal logic plus double negation elimination amounts to classical logic, the following soundness and completeness result for classical first-order logic holds:

Theorem 8 (Sandqvist [25, 26]) $\Gamma \vDash A \iff \Gamma \vdash A$ *in classical first-order logic.*

The theorem is proved constructively by Sandqvist. An alternative proof is given by Makinson [13], who uses classical meta-reasoning.

Sandqvist [26] refers to the implication from right to left as soundness, whereas Makinson [13] takes the opposite perspective, in which the implication from right to left expresses that Sandqvist's semantics is complete with respect to the usual model-theoretic semantics of classical logic. The implication from left to right, that is, completeness in the sense that Sandqvist validity ($\Gamma \vDash A$) implies classical derivability, or equivalently classical validity, holds as well.

6.1 Other Logical Constants

Sandqvist's semantics contains clauses only for the logical constants of the fragment $\{\to, \bot, \forall\}$. A clause for conjunction $\wedge$ like (S5)

$$\vDash_S A \wedge B \;:\Longleftrightarrow\; \vDash_S A \text{ and } \vDash_S B$$

could be added without causing any problems with respect to completeness (cf. Makinson [13]). However, as noted by Sandqvist [26], if a clause for disjunction $\vee$ like (S4)

$$\vDash_S A \vee B \;:\Longleftrightarrow\; \vDash_S A \text{ or } \vDash_S B$$

were added, then Theorem 8 would no longer hold. For example, the law of double negation elimination $(A \to \bot) \to \bot \vDash A$ does then not hold for each substitution instance anymore; a counterexample is $A := B \vee (B \to \bot)$ (cf. [2]). In other words, validity fails to be closed under substitution, if disjunction is taken as primitive and understood according to the given semantical clause. This is also the case for the following stricter disjunction clauses (see Makinson [13]):

$$\vDash_S A \vee B \;:\Longleftrightarrow\; \forall S' \supseteq S : (\vDash_{S'} A \text{ or } \vDash_{S'} B),$$

and

$$\vDash_S A \vee B \;:\Longleftrightarrow\; \forall S' \supseteq S : \vDash_{S'} A \text{ or } \forall S' \supseteq S : \vDash_{S'} B.$$

Similar observations can be made for the existential quantifier.

Makinson also gives an alternative clause for disjunction (see [13, p. 149]), which does not affect completeness. However, this clause is modeled on the definition $A \vee B := (A \to \bot) \to B$, which represents a classical understanding of disjunction, whereas by clause (S4) disjunction is given its intuitionistic meaning.

6.2 Remarks

Theorem 8 still holds if atomic rules of S are allowed to have empty conclusions, and the closure $\underline{S}(\Delta)$ of a set Δ of closed atoms under the rules in S is understood as follows (see [13, p. 152]): $\underline{S}(\Delta)$ is the intersection of all sets Λ of closed atoms such that

(i) $\Delta \subseteq \Lambda$, and if $\dfrac{a_1 \quad \dots \quad a_n}{b} \in S$ with $\{a_1, \dots, a_n\} \subseteq \Lambda$, then $b \in \Lambda$, and

(ii) if $\dfrac{a_1 \quad \dots \quad a_n}{\emptyset} \in S$ with $\{a_1, \dots, a_n\} \subseteq \Lambda$, then $b \in \Lambda$ for every closed atom b (where again $\bot$ is not an atom).

This generalization introduces a kind of negation at the level of atomic rules. In logic programming terms, this is a generalization of definite Horn clauses to Horn clauses.

Theorem 8 fails, however, if second-level rules are allowed in S. For example, consider the atomic system S which contains only the second-level rule

$$\dfrac{\begin{array}{c}[a]\\ b\end{array}}{a}$$

Then $\vDash_S (a \to b) \to a$, but $\nvDash_S a$, since $\nvdash_S a$. Thus $\nvDash ((a \to b) \to a) \to a$, that is, Peirce's law is no longer valid, and soundness fails.

We already remarked that absurdity $\bot$ is not an atom here. Furthermore, it is essential that there are infinitely many atoms in the language; otherwise completeness would be lost, since for finite sets of n atoms the classically non-derivable formula $a_1 \to (\ldots \to (a_n \to \bot) \ldots)$ becomes valid (see Makinson [13]). Soundness would fail if instead of clause (C4) the clause

$$\text{There is no } S \text{ such that } \vDash_S \bot$$

were used (cf. [2, 13]). The use of a semantical clause for $\bot$ could also be avoided. Instead of showing the validity of the law of double negation, which depends both on clause (C3) for $\to$ and on clause (C4) for $\bot$, one can show the validity of Peirce's law, which does not depend on clause (C4) at all (cf. [2, 26]).

Sandqvist's result is remarkable, since it shows that the intuitionistically acceptable semantics given by Definition 19 allows for a justification of classical logic, as long as disjunction is understood classically.

The fact that the semantics is given for only a fragment of the language of first-order logic might be seen as a critical point. This leads to the question of whether such a semantics fulfills the requirements of proof-theoretic semantics for a justification of a logic. Makinson [13] argues that one might require to treat every logical constant used in informal mathematical discourse as a primitive in the formal language of the semantics and to give adequate semantical clauses for each of them. But, as he points out, such a requirement would be difficult to fulfill since it is too vague.

From the point of view of the formal systems used to represent logical reasoning in mathematical discourse one could argue that it is sufficient to have semantical clauses only for the standard logical constants present in the respective formal systems, such as the set $\{\to, \vee, \wedge, \bot, \forall, \exists\}$ of logical constants in natural deduction for intuitionistic or classical logic. In the case of classical logic the restriction to a semantics for a fragment like $\{\to, \bot, \forall\}$, which is sufficient to define all the standard logical constants, should then be acceptable for the purpose of giving a justification for the whole logic.

7 Conclusion

We saw that within proof-theoretic semantics several similar notions of validity have been proposed. For some of these notions completeness results are available for certain fragments of intuitionistic (propositional) logic or for full intuitionistic (propositional) logic. In other cases, such as validity based on higher-level atomic systems, completeness for minimal and intuitionistic logic does not hold. For yet another notion a completeness result holds for classical logic, provided that disjunction is understood classically.

The considered notions of validity have in common that they are not closed under substitution. As derivability in intuitionistic or classical logic is closed under substitution, it seems questionable to even consider these notions as candidates for completeness. Indeed, for intuitionistic logic the failure of completeness with respect to validity based on first- or higher-level atomic systems could be proved by showing the validity of instances of classical laws which are not valid as a schema. For a notion of validity based on atomic systems of level 0, that is, for sets of atoms alone, there are counterexamples of not even classically derivable valid formulas.

As a way out, strengthened notions of validity have been proposed, which are by definition closed under substitution. Thus a formula can now only be valid (in the strengthened sense), if each of its substitution instances, resulting from uniform substitutions of arbitrary formulas for atoms, is valid (in the sense of the underlying, non-strengthened notion of validity). Intuitionistic propositional logic is complete with respect to two of these strengthened notions considered here. In the case of Goldfarb's account, it is essential for completeness (Theorem 4) that only consistent extensions of atomic systems are taken into account. In the case of Sandqvist's completeness result for intuitionistic propositional logic and validity based on second-level atomic systems (Theorem 7) it is crucial that disjunction is explained by the given clause (T4), and not by a more standard clause like (S4).

An essential component of all the considered notions of validity is their dependency on atomic systems. In each notion the validity of atoms a with respect to an atomic system S is defined by derivability of a in S (or as membership in a set of atoms closed under the rules of S), and the validity of implications (or of logical consequences $\Gamma \vDash_S A$) with respect to atomic systems S is defined by making use of extensions S' of S. Using extensions guarantees that validity is monotone with respect to atomic systems S. Whether extensions of atomic systems should be an integral part of any proof-theoretic notion of validity cannot be discussed here; we just point out that, for example, Prawitz has given up to consider extensions of atomic systems from the mid-1970s on and now emphasizes that this is not an intrinsic part of his analysis [personal communication]. His main argument is that atomic systems should not be looked at as descriptions of one's knowledge but as rules defining the meaning of atomic propositions (cf. Prawitz [22, 23]), which would be changed by considering extensions (see [17] for a critical discussion).

With respect to completeness, the choice of the kind of atomic systems can be critical. For example, certain counterexamples to completeness of intuitionistic logic, namely examples of valid classically derivable formulas, can be prevented, if one allows for second-level instead of only first-level atomic systems. With regard to the completeness result for classical logic (Theorem 8) this means that the choice of first-level atomic systems is essential, since completeness does no longer hold for second-level atomic systems. Other results, such as strong completeness for certain fragments of intuitionistic logic, depend on the availability of arbitrary higher-level atomic systems.

For the philosophical endeavor of justifying a certain logic one might want to restrict oneself to first-level atomic systems in the first place, since higher-level systems already presuppose a feature of implication at the atomic level by allowing for

the discharge of atomic assumptions. This presupposition might be deemed too strong for any adequate justification. For a justification of intuitionistic logic one would therefore prefer a proof-theoretic semantics which is restricted to first-level atomic systems, possibly allowing for inconsistent extensions. The question of whether intuitionistic logic is complete for such a semantics is still open.

Acknowledgments This work was supported by the French-German ANR-DFG project "Hypothetical Reasoning—Its Proof-Theoretic Analysis" (HYPOTHESES), DFG grant Schr 275/16-2. It was written during a research stay at the IHPST (Paris), and I am very grateful for the hospitality I received there. I would like to thank Grigory Olkhovikov and Tor Sandqvist for discussions, and Peter Schroeder-Heister for discussions and comments. The completion of this paper was supported by the French-German ANR-DFG project "Beyond Logic: Hypothetical Reasoning in Philosophy of Science, Informatics, and Law", DFG grant Schr 275/17-1.

References

1. de Campos Sanz, W., Piecha, T.: Inversion by definitional reflection and the admissibility of logical rules. Rev. Symb. Log. **2**(3), 550–569 (2009)
2. de Campos Sanz, W., Piecha, T., Schroeder-Heister, P.: Constructive semantics, admissibility of rules and the validity of Peirce's law. Log. J. IGPL **22**(2), 297–308 (2014). First published online 6 August 2013
3. Dummett, M.: The Logical Basis of Metaphysics. Duckworth, London (1991)
4. Ferreira, F.: Comments on predicative logic. J. Philos. Log. **35**, 1–8 (2006)
5. Gabbay, D.M.: On Kreisel's notion of validity in Post systems. Studia Logica **35**(3), 285–295 (1976)
6. Gabbay, D.M.: Semantical Investigations in Heyting's Intuitionistic Logic. Reidel, Dordrecht (1981)
7. Gentzen, G.: Untersuchungen über das logische Schließen. Mathematische Zeitschrift, **39**, 176–210, 405–431 (1934/35). English translation in: Szabo, M.E. (ed.) The Collected Papers of Gerhard Gentzen, pp. 68–131. North-Holland, Amsterdam (1969)
8. Goldfarb, W.: On Dummett's "Proof-theoretic justifications of logical laws". In: Piecha, T., Schroeder-Heister, P. (eds.) Advances in Proof-Theoretic Semantics. Springer, Dordrecht (2016). This volume (Circulated manuscript, 1998)
9. Harrop, R.: Concerning formulas of the types in intuitionistic formal systems. J. Symb. Log. **25**, 27–32 (1960)
10. Kreisel, G.: Unpublished appendix to the paper "Set theoretic problems suggested by the notion of potential totality" in: Infinitistic Methods. Proceedings of the Symposium on Foundations of Mathematics, Warsaw, 2–9 September 1959. Pergamon, Oxford (1961) (reference taken from [5]; unpublished appendix not found)
11. Kreisel, G., Putnam, H.: Eine Unableitbarkeitsbeweismethode für den intuitionistischen Aussagenkalkül. Archiv für mathematische Logik und Grundlagenforschung **3**, 74–78 (1957)
12. Kripke, S.A.: Semantical analysis of intuitionistic logic I. In: Crossley, J., Dummett, M.A.E. (eds.) Formal Systems and Recursive Functions, pp. 92–130. North-Holland, Amsterdam (1965)
13. Makinson, D.: On an inferential semantics for classical logic. Log. J. IGPL **22**(1), 147–154 (2014)

14. Moschovakis, J.: Intuitionistic logic. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (2014). http://plato.stanford.edu/archives/fall2014/entries/logic-intuitionistic/
15. Olkhovikov, G.K., Schroeder-Heister, P.: Proof-theoretic harmony and the levels of rules: generalised non-flattening results. In: Moriconi, E., Tesconi, L. (eds.) Second Pisa Colloquium in Logic, Language and Epistemology, pp. 245–287. ETS, Pisa (2014)
16. Piecha, T., de Campos Sanz, W., Schroeder-Heister, P.: Failure of completeness in proof-theoretic semantics. J. Philos. Log. **44**(3), 321–335 (2014). First published online 1 August 2014
17. Piecha, T., Schroeder-Heister, P.: Atomic systems in proof-theoretic semantics: two approaches. In: Redmond, J., Nepomuceno Fernández, A., Pombo, O. (eds.) Epistemology, Knowledge and the Impact of Interaction. Springer, Dordrecht (2016)
18. Prawitz, D.: Ideas and results in proof theory. In: Fenstad, J.E. (ed.) Proceedings of the Second Scandinavian Logic Symposium, Studies in Logic and the Foundations of Mathematics, vol. 63, pp. 235–307. North-Holland, Amsterdam (1971)
19. Prawitz, D.: Towards a foundation of a general proof theory. In: Suppes, P., et al. (eds.) Logic, Methodology and Philosophy of Science IV, pp. 225–250. North-Holland, Amsterdam (1973)
20. Prawitz, D.: On the idea of a general proof theory. Synthese **27**, 63–77 (1974)
21. Prawitz, D.: Meaning approached via proofs. In: Kahle, R., Schroeder-Heister, P. (eds.) Proof-Theoretic Semantics. Synthese, vol. 148, pp. 507–524. Springer, Berlin (2006)
22. Prawitz, D.: An approach to general proof theory and a conjecture of a kind of completeness of intuitionistic logic revisited. In: Pereira, L.C., Haeusler, E.H., de Paiva, V. (eds.) Advances in Natural Deduction, Trends in Logic, vol. 39, pp. 269–279. Springer, Berlin (2014)
23. Prawitz, D.: On the relation between Heyting's and Gentzen's approaches to meaning. In: Piecha, T., Schroeder-Heister, P. (eds.) Advances in Proof-Theoretic Semantics. Springer, Dordrecht (2016). This volume
24. Read, S.: Proof-theoretic validity. In: Caret, C.R., Hjortland, O.T. (eds.) Foundations of Logical Consequence, Mind Association Occasional Series, pp. 136–158. Oxford University Press, Oxford (2015)
25. Sandqvist, T.: An inferentialist interpretation of classical logic. Ph.D. thesis, Department of Philosophy, Uppsala University (2005)
26. Sandqvist, T.: Classical logic without bivalence. Analysis **69**, 211–217 (2009)
27. Sandqvist, T.: Basis-extension semantics for intuitionistic sentential logic (2014). To appear
28. Schroeder-Heister, P.: The completeness of intuitionistic logic with respect to a validity concept based on an inversion principle. J. Philos. Log. **12**, 359–377 (1983)
29. Schroeder-Heister, P.: A natural extension of natural deduction. J. Symb. Log. **49**, 1284–1300 (1984)
30. Schroeder-Heister, P.: Proof-theoretic validity and the completeness of intuitionistic logic. In: Dorn, G., Weingartner, P. (eds.) Foundations of Logic and Linguistics: Problems and Their Solutions, pp. 43–87. Plenum Press, New York (1985)
31. Schroeder-Heister, P.: Validity concepts in proof-theoretic semantics. In: Kahle, R., Schroeder-Heister, P. (eds.) Proof-Theoretic Semantics. Synthese, vol. 148, pp. 525–571. Springer, Berlin (2006)
32. Schroeder-Heister, P.: The calculus of higher-level rules, propositional quantification, and the foundational approach to proof-theoretic harmony. In: Indrzejczak, A. (ed.) Gentzen's and Jaśkowski's Heritage. 80 Years of Natural Deduction and Sequent Calculi. Studia Logica, vol. 103, pp. 1185–1216. Springer, Berlin (2014)
33. Schroeder-Heister, P.: Examples of proof-theoretic validity (Supplement to [34]). In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (2012). http://plato.stanford.edu/entries/proof-theoretic-semantics/examples.html
34. Schroeder-Heister, P.: Proof-theoretic semantics. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy (2012). http://plato.stanford.edu/entries/proof-theoretic-semantics/
35. Veldman, W.: An intuitionistic completeness theorem for intuitionistic predicate logic. J. Symb. Log. **41**, 159–166 (1976)
36. Wansing, H.: The idea of a proof-theoretic semantics and the meaning of the logical operations. Studia Logica **64**, 3–20 (2000)

Open Problems in Proof-Theoretic Semantics

Peter Schroeder-Heister

Abstract I present three open problems the discussion and solution of which I consider relevant for the further development of proof-theoretic semantics: (1) The nature of hypotheses and the problem of the appropriate format of proofs, (2) the problem of a satisfactory notion of proof-theoretic harmony, and (3) the problem of extending methods of proof-theoretic semantics beyond logic.

Keywords Proof-theoretic semantics · Hypothesis · Natural deduction · Sequent calculus · Harmony · Identity of proofs · Definitional reflection

1 Introduction

Proof-theoretic semantics is the attempt to give semantical definitions in terms of proofs. Its main rival is truth-theoretic semantics, or, more generally, semantics that treats the denotational function of syntactic entities as primary. However, since the distinction between truth-theoretic and proof-theoretic approaches is not as clear cut as it appears at first glance, particularly if 'truth-theoretic' is understood in its model-theoretic setting (see Hodges [33], and Došen [10]). it may be preferable to redirect attention from the negative characterisation of proof-theoretic semantics to its positive delineation as the explication of meaning through proofs. Thus, we leave aside the question of whether alternative approaches can or do in fact deal with the phenomena that proof-theoretic semantics tries to explain. In proof-theoretic semantics, proofs are not understood simply as formal derivations, but as entities expressing arguments by means of which we can acquire knowledge. In this sense, proof-theoretic semantics is closely connected and strongly overlaps with what Prawitz has called *general proof theory*.

The task of this paper is not to provide a philosophical discussion of the value and purpose of proof-theoretic semantics. For that the reader may consult Schroeder-Heister [56, 61] and Wansing [72]. The discussion that follows presupposes some

P. Schroeder-Heister (✉)
Department of Computer Science, University of Tübingen, Tübingen, Germany
e-mail: psh@uni-tuebingen.de

T. Piecha and P. Schroeder-Heister (eds.), *Advances in Proof-Theoretic Semantics*,
Trends in Logic 43, DOI 10.1007/978-3-319-22686-6_16

acquaintance with basic issues of proof-theoretic semantics. Three problems are addressed, which I believe are crucial for the further development of the proof-theoretic approach. This selection is certainly personal, and many other problems might be added. However, it is my view that grappling with these three problems opens up further avenues of enquiry that are needed if proof-theoretic semantics is to mature as a discipline.

The first problem is the understanding of *hypotheses* and the *format of proofs*. It is deeply philosophical and deals with the fundamental concepts of reasoning, but has important technical implications when it comes to formalizing the notion of proof. The second problem is the proper understanding of proof-theoretic *harmony*. This is one of the key concepts within proof-theoretic semantics. Here we claim that an intensional notion of harmony should be developed. The third problem is the need to widen our perspective *from logical to extra-logical* issues. This problem proceeds from the insight that the traditional preoccupation of proof-theoretic semantics with logical constants is far too limited.

I work within a conventional proof-theoretic framework where natural deduction and sequent calculus are the fundamental formal models of reasoning. Using categorial logic, which can be viewed as abstract proof theory, many new perspectives on these three problems would become possible. This task lies beyond the scope of what can be achieved here. Nevertheless, I should mention that the proper recognition of categorial logic within proof-theoretic semantics is still a *desideratum*. For the topic of categorial proof theory the reader is referred to Došen's work, in particular to his programmatic statement of 1995 [6], his contribution to this volume [11] and the detailed expositions in two monographs [7, 12].

2 The Nature of Hypotheses and the Format of Proofs

The notion of proof from hypotheses—hypothetical proof—lies at the heart of proof-theoretic semantics. A hypothetical proof is what justifies a hypothetical judgement, which is formulated as an implication. However, it is not clear what is to be understood by a hypothetical proof. In fact, there are various competing conceptions, often not made explicit, which must be addressed in order to describe this crucial concept.

2.1 Open Proofs and the Placeholder View

The most widespread view in modern proof-theoretic semantics is what I have called the *primacy of the categorical over the hypothetical* [58, 60]. According to this view, there is a primitive notion of proof, which is that of an assumption-free proof of an assertion. Such proofs are called *closed proofs*. A closed proof proves outright, without referring to any assumptions, that which is being claimed. A proof from assumptions is then considered an open proof, that is, a proof which, using Frege's

term, may be described as 'unsaturated'. The open assumptions are marks of the places where the proof is unsaturated. An open proof can be closed by substituting closed proofs for the open assumptions, yielding a closed proof of the final assertion. In this sense, the open assumptions of an open proof are placeholders for closed proofs. Therefore, one can speak of the *placeholder view of assumptions*.

Prawitz [46], for example, speaks explicitly of open proofs as codifying open arguments. Such arguments are, so-to-speak, arguments with holes that can be filled with closed arguments, and similarly for open judgements and open grounds [50]. A formal counterpart of this conception is the Curry-Howard correspondence, in which open assumptions are represented by free term variables, corresponding to the function of variables to indicate open places. Thus, one is indeed justified in viewing this conception as extending to the realm of proofs Frege's idea of the unsaturatedness of concepts and functions. That hypotheses are merely placeholders entails that no specific speech act is associated with them. Hypotheses are not posed or claimed but play a subsidiary role in a superordinated claim of which the hypothesis marks an open place.

I have called this *placeholder view* of assumptions and the *transmission view* of hypothetical proofs a *dogma* of standard semantics. It should be considered a dogma, as it is widely accepted without proper discussion, despite alternative conceptions being readily available. It belongs to *standard semantics*, as it underlies not only the dominant conception of proof-theoretic semantics, but in some sense it also underlies classical truth-condition semantics. In the classical concept of consequence according to Bolzano and Tarski, the claim that B follows from A is justified by the fact that, in any model of A—in any world, in which A is true—, B is true as well. This means that hypothetical consequence is justified by reference to the transmission of the categorical concept of truth from the condition to the consequent. We are not referring here to functions or any sort of process that takes us from A to B, but just to the metalinguistic universal implication that, whenever A is true, B is true as well. However, as with the standard semantics of proofs, we retain the idea that the categorical concept precedes the hypothetical concept, and the latter is justified by reference to the former concept.

The concept of open proofs in proof-theoretic semantics employs not only the idea that they can be closed by substituting something into the open places, but also the idea that they can be closed outright by a specific operation of *assumption discharge*. This is what happens with the application of implication introduction as described by Jaśkowski [34] and Gentzen [25]. Here the open place disappears because what is originally expressed by the open assumption now becomes the condition of an implication. This can be described as a *two-layer system*. In addition to hypothetical judgements given by an open proof with the hypotheses as open places, we have hypothetical judgements in the form of implications, in which the hypothesis is a subsentence. A hypothetical judgement in the sense of an implication is justified by an open proof of the consequent from the condition of the implication. The idea of two layers of hypotheses is typical for assumption-based calculi and in particular for natural deduction, which is the main deductive model of proof-theoretic semantics.

According to the placeholder view of assumptions there are two operations that one can perform as far as assumptions are concerned: Introducing an assumption as an open place, and eliminating or closing it. The closing of an assumption can be achieved in either of two ways: by substituting another proof for it, or by discharge. In the substitution operation the proof substituted for the assumption need not necessarily be closed. However, when it is open then instead of the original open assumption the open assumptions of the substitute become open assumptions of the whole proof. Thus a full closure of an open assumption by means of substitution requires a closed proof as a substitute. To summarise, the three basic operations on assumptions are: assumption introduction, assumption substitution, and assumption discharge.

These three basic operations on assumptions are unspecific in the following sense: They do not depend on the internal form of the assumption, that is, they do not depend on its logical or non-logical composition. They take the assumption as it is. In this sense these operations are structural. The operation of assumption discharge, although pertaining to the structure of the proof, is normally used in the context of a logical inference, namely the introduction of implication. The crucial idea of implication introduction, as first described by Jaśkowski and Gentzen, involves that, in the context of a logical inference, the structure of the proof is changed. This structure-change is a non-local effect of the logical rule.

The unspecific character of the operations on assumptions means that, in the placeholder view, assumptions are not manipulated in any sense. The rules that govern the internal structure of propositions are always rules that concern the assertions, but not the assumptions made. Consequently, the placeholder view is assertion-centred as far as content is concerned. In an inference step we pass from assertions already made to another assertion. At such a step the structure of the proof, in particular which assumptions are open, may be changed, but not the internal form of assumptions. This may be described as the forward-directedness of proofs. When proving something, we may perform structural changes in the proof that lies 'behind' us, but without changing the content of what lies behind. It is important to note that we do not here criticise this view of proofs. We merely highlight what one must commit oneself to in affirming this view.

2.2 The No-Assumptions View

The most radical alternative to the placeholder view of assumptions is the claim that there are no assumptions at all. This view is much older than the placeholder view and was strongly advocated by Frege (see for example Frege [22]). Frege argues that the aim of deduction is to establish truth, and, in order to achieve that goal, deductions proceed from true assertions to true assertions. They start with assertions that are evident or for other reasons true. This view of deduction can be traced back to Bolzano's *Wissenschaftslehre* [1] and its notion of 'Abfolge'. This notion means the

relationship between true propositions A and B, which obtains if B holds *because* A holds.

However, in view of the fact that hypothetical claims abound in everyday life, science, legal reasoning etc., it is not very productive simply to deny the idea of hypotheses. Frege was of course aware that 'if …, then …' statements play a central role wherever we apply logic. His logical notation (the *Begriffsschrift*) uses implication as one of the primitive connectives. The fact that he can still oppose the idea of reasoning from assumptions is that he denies a two-layer concept of hypotheses. As we have the connective of implication at hand, there is no need for a second kind of hypothetical entity that consists of a hypothetical proof from assumptions. Instead of maintaining that B can be proved from the hypothesis A, we should just be able to prove A *implies* B non-hypothetically, which has the same effect. There is no need to consider a second structural layer at which hypotheses reside.

For Frege this means, of course, that implication is not justified by some sort of introduction rule, which was a much later invention of Jaśkowski and Gentzen. The laws of implication are justified by truth-theoretic considerations and codified by certain axioms. That A implies itself, is, for example, one of these axioms (in Frege's *Grundgesetze*, [21]), from which a proof can start, as it is true.

The philosophical burden here lies in the justification of the primitive axioms. If, like Frege, we have a truth-theoretic semantics at hand, this is no fundamental problem, as Frege demonstrates in detail by a truth-valuational procedure. It becomes a problem when the meaning of implication is to be explained in terms of proofs. This is actually where the need for the second structural layer arises. It was the ingenious idea of Jaśkowski and Gentzen to devise a two-layer method that reduces the meaning of implication to something categorically different. Even though this interpretation was not, or was only partly, intended as a meaning theory for implication (in Gentzen [25] as an explication of actual reasoning in mathematics, in Jaśkowski [34] as an explication of suppositional reasoning) it has become crucial in that respect in later proof-theoretic semantics. The single-layer alternative of starting from axioms in reasoning presupposes an external semantics that is not framed in terms of proofs.

Such an external semantics need not be a classical truth-condition semantics, or an intuitionistic Kripke-style semantics. It could, for example, be a constructive BHK-style semantics, perhaps along the lines of Goodman-Kreisel or a variant of realisability (see, for example, Dean and Kurokawa [4]). We would then have a justification of a formal system by means of a soundness proof. As soon as the axioms or rules of our system are sound with respect to this external semantics, they are justified. In proof-theoretic semantics, understood in the strict sense of the term, such an external semantics is not available. This means that a single-layer concept of implication based only on axioms and rules, but without assumptions, is not a viable option.[1]

[1] Digression on Frege: Even though formally Frege has only a single-layer system, there is a hidden two-layer system that lies in the background. Frege makes an additional distinction between upper member and lower members of (normally iterated) implications. This distinction is not a syntactical property of the implication itself, but something that we attach to it, and that we can attach to it

2.3 Bidirectionality

The transmission view of consequence is incorporated in natural deduction in that we have the operations of assumption introduction, assumption-closure by substitution and assumption-closure by discharge. There are various notations for it. The most common such notation in proof-theoretic semantics is Gentzen's [25] tree notation that was adopted and popularised by Prawitz [44]. Alternative notations are Jaśkowski's [34] box notation, which is the origin of Fitch's [18] later notation. A further notation is sequent-style natural deduction. In this notation proofs consist of judgements of the form $A_1, \ldots, A_n \vdash B$, where the A's represent the assumptions and B the conclusion of what is claimed. This can be framed either in tree form, in boxed form, or in a mixture of both. With sequent-style natural deduction the operation of assumption discharge is no longer non-local. In the case of implication introduction

$$\frac{A_1, \ldots, A_n, B \vdash C}{A_1, \ldots, A_n \vdash B \rightarrow C} \tag{1}$$

we pass from one sequent to another without changing the structure of the proof. In fact, in such a proof there are no sequents as assumptions, but every top sequent is an axiom, normally of the form $A \vdash A$ or $A_1, \ldots, A_n, A \vdash A$. However, this is not a no-assumptions system in the sense of Sect. 2.2: It combines assumption-freeness with a two-layer approach. The structural layer is the layer of sequents composed out of lists of formulas building up the left and right sides (antecedent and succedent) of a sequent, whereas the logical layer concerns the internal structure of formulas. By using introduction rules such as (1), the logical layer is characterised in terms of the structural layer. In this sense sequent-style natural deduction is a variant of 'standard' natural deduction.

However, a different picture emerges if we look at the sequent calculus LK or LJ that Gentzen devised. These are two-layer systems in which we can manipulate not only what is on the right hand side and what corresponds to assertions, but also what is on the left hand side and what corresponds to assumptions. In sequent-style natural deduction a sequent

$$A_1, \ldots, A_n \vdash B \tag{2}$$

just means that we have a proof of B with open assumptions $A_1, \ldots, A_n$:

$$\begin{array}{c} A_1, \ldots, A_n \\ \mathscr{D} \\ B \end{array} \tag{3}$$

(Footnote 1 continued)
in different ways. This distinction is analogous to that between assumptions and assertion, so that, when we prove an implication, we can at the same time regard this proof as a proof of the upper member of this implication from its lower members taken as assumptions. In the *Grundgesetze* Frege [21] even specifies rules of proofs in terms of this second-layer distinction. This means that he himself goes beyond his own idea of a single-layer system. See Schroeder-Heister [64].

In the symmetric sequent calculus, which is the original form of the sequent calculus, the $A_1, \ldots, A_n$ can still be interpreted as assumptions in a derivation of B, but no longer as open assumptions in the sense of the transmission view. Or at least they can be given an alternative interpretation that leads to a different concept of reasoning. In the sequent calculus we have introduction rules not only on the right hand side, but also on the left hand side, for example, in the case of conjunction:

$$(\wedge\mathrm{L})\ \frac{A, A_1, \ldots, A_n \vdash C}{A \wedge B, A_1, \ldots, A_n \vdash C}$$

This can be interpreted as a novel model of reasoning, which is different from assertion-centred forward reasoning in natural deduction. When reading a sequent (2) in the sense of (3), the step of ($\wedge$L) corresponds to:

$$\begin{array}{c} \underline{A \wedge B} \\ A \\ \mathscr{D} \\ C \end{array}$$

This is a step that continues a given derivation $\mathscr{D}$ of C from A upwards to a derivation of C from $A \wedge B$. Therefore, from a philosophical point of view, the sequent calculus presents a model of bidirectional reasoning, that is, of reasoning that, by means of right-introduction rules, extends a proof downwards, and, by means of left-introduction rules, extends a proof upwards.

This is, of course, a philosophical interpretation of the sequent calculus, reading it as describing a certain way of constructing a proof from hypotheses.[2] Since under this schema both assumptions and assertions are liable to application of rules, assumptions are no longer understood simply as placeholders for closed proofs. Both assumptions and assertions are now entities in their own right. Read in that way we have a novel picture of the nature of hypotheses. We can give this reading a format in natural-deduction style, namely by formulating the rule ($\wedge$L) as

$$\frac{\underline{} \qquad \begin{array}{c}[A] \\ C\end{array}}{C}$$

where the line above $A \wedge B$ expresses that $A \wedge B$ stands here as an assumption. We may call this system a natural-deduction sequent calculus, that is, a sequent calculus in natural-deduction style. It is a system in which major premisses of elimination rules occur only as assumptions ('stand proud' in Tennant's [70] terminology). The

[2]Gentzen [25] himself devised the sequent calculus as a technical device to prove his Hauptsatz after giving a philosophical motivation of the calculus of natural deduction. He wanted to give a calculus in 'logistic' style, by which he meant a calculus without assumptions that just moves from claim to claim and whose rules are local due to the assumption-freeness of the system. The term 'logistic' comes from the designation of modern symbolic as opposed to traditional logic in the 1920s (see Carnap, [3]).

intuitive idea of this step is that we can introduce an assumption in the course of a derivation. If a proof of C from A is given, we can, by introducing the assumption $A \wedge B$ and discharging the given assumption A, pass over to C. That $A \wedge B$ occurs only as an assumption and cannot be a conclusion of any other rule, demonstrates that we have a different model of reasoning, in which assumptions are not just placeholders for other proofs, but stand for themselves. The fact that, given a proof $\begin{array}{c}\mathscr{D}\\ A \wedge B\end{array}$ of $A \wedge B$ and a proof of the form

$$\dfrac{\dfrac{}{A \wedge B} \qquad \begin{array}{c}[A]\\ \mathscr{D}'\\ C\end{array}}{C}$$

we obtain a proof of C by combining these two proofs is no longer built into the system and its semantics, but something that must be proved in the form of a cut elimination theorem. According to this philosophical re-interpretation of the sequent calculus, assumptions and assertions now resemble handles at the top and at the bottom of a proof, respectively. It is no longer the case that one side is a placeholder whereas the other side represents proper propositions.

This interpretation also means that the sequent calculus is not just a meta-calculus for natural deduction, as Prawitz ([44], Appendix A) suggests. It is a meta-calculus in the sense that whenever there is a natural-deduction derivation of B from $A_1, \ldots, A_n$, there is a sequent-calculus derivation of the sequent $A_1, \ldots, A_n \vdash B$, and vice versa. However, this does not apply to proofs. There is no rule application in natural deduction that corresponds to an application of a left-introduction rule in the sequent calculus. This means that at the level of proof construction there is no one-one correspondence, but something genuinely original in the sequent calculus. This is evidenced by the fact that the translation between sequent calculus and natural deduction is non-trivial. Prawitz is certainly right that a sequent-calculus proof can be viewed as giving instructions about how to construct a corresponding natural-deduction derivation. However, we would like to emphasise that it can be interpreted to be more than such a metalinguistic tool, namely as representing a way of reasoning in its own right. We do not want to argue here in favour of either of these positions. However, we would like to emphasise that the philosophical significance of the bidirectional approach has not been properly explored (see also Schroeder-Heister, [59]).

2.4 Local and Global Proof-Theoretic Semantics

We have discussed the philosophical background of three conceptions of hypotheses and hypothetical proofs. Each of them has strong implications for the form of proof-theoretic semantics. According to the no-assumptions view (Sect. 2.2) with its single-layer conception there is no structural way of dealing with hypotheses. Therefore,

there is no proof-theoretic semantics of implication, at least not along the common line that an implication expresses that we have or can generate a hypothetical proof. We would need instead a semantics from outside.

A proof-theoretic semantics for the placeholder view of assumptions (Sect. 2.1), even though it is assertion-centred, is not necessarily verificationist in the sense that it considers introduction rules for logical operators to be constitutive of meaning. Nothing prevents us from considering elimination rules as primitive meaning-giving rules and justifying introduction rules from them (see Prawitz [49], Schroeder-Heister [67]). However, the placeholder-view forces one particular feature that might be seen as problematic from certain points of view, namely the global character of the semantics.

According to the placeholder-view of assumptions, an open derivation $\mathscr{D}$ of B from A

$$\begin{array}{c} A \\ \mathscr{D} \\ B \end{array}$$

would be considered valid if for every closed derivation of A

$$\begin{array}{c} \mathscr{D}' \\ A \end{array}$$

the derivation

$$\begin{array}{c} \mathscr{D}' \\ A \\ \mathscr{D} \\ B \end{array}$$

obtained by substituting the derivation $\mathscr{D}'$ of A for the open assumption A is valid. This makes sense only if proof-theoretic validity is defined for whole proofs rather than for single rules, since the entity in which the assumption A is an open place is a proof. A proof would not be considered valid because it is composed of valid rules, but conversely, a rule would be considered valid if it is a limiting case of a proof, namely a one-step proof. This is actually how the definitions of validity in the spirit of Prawitz's work proceed (Prawitz [48], Schroeder-Heister [56, 61]).

This global characteristic of validity has strong implications. We must expect now that a proof as a whole is well-behaved in a certain sense, for example, that it has certain features related to normalisability. In all definitions of validity we have as a fundamental property that a closed proof is valid iff it reduces to a valid closed proof, which means that validity is always considered modulo reduction. And reduction applies to the proof as a whole, which means that it is a global issue. As validity is global, there is no way for partial meaning in any sense. A proof can be valid, and it can be invalid. However, there is no possibility of the proof being only partially valid as reflected in the way the proof behaves.

This global proof semantics has its merits as long as one considers only cases such as the standard logical constants, where everything is well-founded and we can build valid sentences from the bottom up. However, it reaches its limits of applicability, if proof-theoretic semantics should cover situations where we do not have such a full specification of meaning. When dealing with iterated inductive definitions, we can, of course, require that definitions be well-behaved, as Martin-Löf [40] did in his theory. However, when it comes to partial inductive definitions, the situation is different (see Sect. 4).

Here it is much easier to say: We have locally valid rules, but the composition of such rules is not necessarily a globally valid derivation. In a rule-based approach we can make the composition of rules and its behaviour a problem, whereas on the transmission view the validity of composition is always enforced. Substitution becomes an explicit step, which can be problematised. It will be possible, in particular, to distinguish between the validity of rules and the effect the composition of rules has on a proof. In fact, it is not even mandatory to allow from the very beginning that each composition of (locally) valid rules renders a proof valid. We might impose further restriction on the composition of valid rules. This occurs especially when the composition of locally valid rules does not yield a proof the assumptions of which can be interpreted as placeholders, that is, for which the substitution property does not hold.

Therefore, the bidirectional model of proof allows for a local proof-theoretic semantics. Here we can talk simply of *rules* that extend a proof on the assertion or on the assumption side. There will be rules for each side, and one may discuss issues such as when these rules are in harmony or not. Whether one side is to be considered primary, and, if yes, which one, does not affect the model of reasoning as such. In any case a derivation would be called valid if it consists of the application of valid rules, which is exactly what local proof-theoretic semantics requires.

3 The Problem of Harmony

In the proof-theoretic semantics of logical constants, harmony is a, or perhaps the, crucial concept. If we work in a natural-deduction framework, harmony is a property that introduction and elimination rules for a logical constant are expected to satisfy with respect to each other in order to be appropriate. Harmony guarantees that we do not gain anything when applying an introduction rule followed by an elimination rule, but also, conversely, that from the result of applying elimination rules we can, by applying introduction rules, recover what we started with. The notion of harmony or 'consonance' was introduced by Dummett ([13], pp. 396–397).[3]

[3] At least in his more logic-oriented writings, Dummett tends to use 'harmony' as comprising only the 'no-gain' direction of introductions followed by eliminations, and not the 'recovery' direction of eliminations followed by introductions, which he calls 'stability'. See Dummett [14].

However, it is not absolutely clear how to define harmony. Various competing understandings are to be found in the literature. We identify a particular path that has not yet been explored, and we call this path 'intensional' or 'strong' harmony. The need to consider such a notion on the background of the discussion, initiated by Prawitz [45], on the identity of proofs, in particular in the context of Kosta Došen's work (see Došen [6, 8, 9], Došen and Petrić [12]), was raised by Luca Tranchini.[4] As the background to this issue we first present two conceptions of harmony, which are not reliant on the notion of identity of proofs.

3.1 Harmony Based on Generalised Rules

According to Gentzen "the introductions represent so-to-speak the 'definitions' of the corresponding signs" whereas the eliminations are "consequences" thereof, which should be demonstrated to be "unique functions of the introduction inferences on the basis of certain requirements" (Gentzen [25], p. 189). If we take this as our characterisation of harmony, we must specify a function $\mathscr{F}$ which generates elimination rules from given introduction rules. If elimination rules are generated according to this function, then introduction and elimination rules are in harmony with each other.

There have been various proposals to formulate elimination rules in a uniform way with respect to given introduction rules, in particular those by von Kutschera [36], Prawitz [47] and Schroeder-Heister [53]. At least implicitly, they all intend to capture the notion of harmony. Read [51, 52] has proposed to speak of 'general-elimination harmony'. Formulated as a principle, we could say: Given a set $c\mathscr{I}$ of introduction rules for a logical constant c, the set of elimination rules harmonious with $c\mathscr{I}$ is the set of rules generated by $\mathscr{F}$, namely $\mathscr{F}(c\mathscr{I})$. In other words, $c\mathscr{I}$ and $\mathscr{F}(c\mathscr{I})$ are by definition in harmony with each other. If alternative elimination rules $c\mathscr{E}$ are given for c, one would say that $c\mathscr{E}$ is in harmony with $c\mathscr{I}$, if $c\mathscr{E}$ is equivalent to $\mathscr{F}(c\mathscr{I})$ in the presence of $c\mathscr{I}$. This means that, in the system based on $c\mathscr{I}$ and $c\mathscr{E}$, we can derive the rules contained in $\mathscr{F}(c\mathscr{I})$, and in the system based on $c\mathscr{I}$ and $\mathscr{F}(c\mathscr{I})$, we can derive the rules contained in $c\mathscr{E}$.

Consequently the generalised elimination rules $\mathscr{F}(c\mathscr{I})$ are *canonical harmonious elimination rules*[5] given introduction rules $c\mathscr{I}$. The approaches mentioned above develop arguments that justify this distinguishing characteristic, for example by referring to an inversion principle. The canonical elimination rule ensures that

[4] This topic will be further pursued by Tranchini and the author.

[5] Of course, this usage of the term 'canonical' is different from its usage in connection with meaning-giving introduction rules, for example, for derivations using an introduction rule in the last step.

everything that can be obtained from the premisses of each introduction rule can be obtained from their conclusion. For example, if the introduction rules for φ have the form

$$(\varphi\ \mathrm{I})\ \frac{\Delta_1}{\varphi} \quad \dots \quad \frac{\Delta_m}{\varphi}$$

the canonical elimination rule takes the form

$$(\varphi\ \mathrm{E})_{can}\ \frac{\varphi \quad \begin{array}{c}[\Delta_1]\\ r\end{array} \quad \dots \quad \begin{array}{c}[\Delta_m]\\ r\end{array}}{r}$$

The exact specification of what the Δ_i can mean, and what it means to use the Δ_i as dischargeable assumptions, depends on the framework used (see Schroeder-Heister [63, 65]).[6]

While the standard approaches use introduction rules as their starting point, it is possible in principle, and in fact not difficult, to develop a corresponding approach based on elimination rules. Given a set of elimination rules $c\mathscr{E}$ of a connective c, we would define a function $\mathscr{G}$ that associates with $c\mathscr{E}$ a set of introduction rules $\mathscr{G}(c\mathscr{E})$ as the set of introduction rules harmonious to $c\mathscr{I}$. While the rules in $\mathscr{G}(c\mathscr{E})$ are the *canonical harmonious introduction rules*, any other set $c\mathscr{I}$ of introduction rules for c would be in harmony with $c\mathscr{E}$ if $c\mathscr{I}$ is equivalent to $\mathscr{G}(c\mathscr{E})$ in the presence of $c\mathscr{E}$. This means that, in the system based on $c\mathscr{E}$ and $c\mathscr{I}$, we can derive the rules contained in $\mathscr{G}(c\mathscr{E})$, and in the system based on $c\mathscr{E}$ and $\mathscr{G}(c\mathscr{E})$, we can derive the rules in $c\mathscr{I}$. For example, if the elimination rules have the form

$$(\varphi\ \mathrm{E})\ \frac{\varphi \quad \Delta_1}{q_1} \quad \dots \quad \frac{\varphi \quad \Delta_m}{q_m}$$

then the canonical introduction rule takes the form

$$(\varphi\ \mathrm{I})_{can}\ \frac{\begin{array}{c}[\Delta_1]\\ q_1\end{array} \quad \dots \quad \begin{array}{c}[\Delta_m]\\ q_m\end{array}}{\varphi}$$

Here the conclusions of the elimination rules become premisses of the canonical introduction rules. Again, the exact specification of Δ_i depends on the framework used. For example, if, for the four-place connective $\wedge\!\rightarrow$, the set $\wedge\!\rightarrow\mathscr{E}$ consists of the three elimination rules

$$(\wedge\!\rightarrow\mathrm{E})\ \frac{\wedge\!\rightarrow(p_1, p_2, p_3, p_4)}{p_1} \qquad \frac{\wedge\!\rightarrow(p_1, p_2, p_3, p_4)}{p_2}$$
$$\frac{\wedge\!\rightarrow(p_1, p_2, p_3, p_4) \qquad p_3}{p_4}$$

[6]In this section, we do not distinguish between schematic letters for formulas and propositional variables, as we are also considering propositional quantification. Therefore, in a rule schema such as $(\varphi\ \mathrm{E})_{can}$, the propositional variable r is used as a schematic letter.

we would define $\mathscr{G}(\wedge\rightarrow \mathscr{E})$ as consisting of the single introduction rule

$$(\wedge\rightarrow \mathrm{I})_{can}\ \frac{p_1 \qquad p_2 \qquad \begin{matrix}[p_3]\\ p_4\end{matrix}}{\wedge\rightarrow (p_1, p_2, p_3, p_4)}$$

In Schroeder-Heister [63] functions $\mathscr{F}$ and $\mathscr{G}$ are defined in detail.

3.2 Harmony Based on Equivalence

Approaches based on generalised eliminations or generalised introductions maintain that these generalised rules have a distinguished status, so that harmony can be defined with respect to them. An alternative way would be to explain what it means that given introductions $c\mathscr{I}$ and given eliminations $c\mathscr{E}$ are in harmony with each other, independent of any syntactical function that generates $c\mathscr{E}$ from $c\mathscr{I}$ or vice versa. This way of proceeding has the advantage that rule sets $c\mathscr{I}$ and $c\mathscr{E}$ can be said to be in harmony without starting either from the introductions or from the eliminations as primary meaning-giving rules. That for certain syntactical functions $\mathscr{F}$ and $\mathscr{G}$ the rule sets $c\mathscr{I}$ and $\mathscr{F}(c\mathscr{I})$, or $c\mathscr{E}$ and $\mathscr{G}(c\mathscr{E})$, are in harmony, is then a special *result* and not the *definiens* of harmony. The canonical functions generating harmonious rules operate on sets of introduction and elimination rules for which harmony is already defined independently. This symmetry of the notion of harmony follows naturally from an intuitive understanding of the concept.

Such an approach is described for propositional logic in Schroeder-Heister [66]. Its idea is to translate the meaning of a connective c according to given introduction rules $c\mathscr{I}$ into a formula c^I of second-order intuitionistic propositional logic IPC2, and its meaning according to given elimination rules $c\mathscr{E}$ into an IPC2-formula c^E. Introductions and eliminations are then said to be in harmony with each other, if c^I and c^E are equivalent (in IPC2). The introduction and elimination meanings c^I and c^E can be read off the proposed introduction and elimination rules. For example, consider the connective && with the introduction and elimination rules

$$\frac{p_1 \qquad \begin{matrix}[p_1]\\ p_2\end{matrix}}{p_1 \,\&\&\, p_2} \qquad\qquad \frac{p_1\&\&p_2 \qquad \begin{matrix}[p_1]\\ r_1\end{matrix} \qquad \begin{matrix}[p_2]\\ r_2\end{matrix} \qquad \begin{matrix}[r_1, r_2]\\ r\end{matrix}}{r}$$

Its introduction meaning is $p_1 \wedge (p_1 \rightarrow p_2)$, and its elimination meaning is $\forall r_1 r_2 r(((p_1 \rightarrow r_1) \wedge (p_2 \rightarrow r_2) \wedge ((r_1 \wedge r_2) \rightarrow r)) \rightarrow r)$. As these formulas are equivalent in IPC2, the introduction and elimination rules for && are in harmony with each other. Further examples are discussed in Schroeder-Heister [66].

The translation into IPC2 presupposes, of course, that the connectives inherent in IPC2 are already taken for granted. Therefore, this approach works properly only for generalised connectives different from the standard ones. As it reduces semantical

content to what can be expressed by formulas of IPC2, it was called a 'reductive' rather than 'foundational' approach. As described in Schroeder-Heister [63] this can be carried over to a framework that employs higher-level rules, making the reference to IPC2 redundant. However, as the handling of quantified rules in this framework corresponds to what can be carried out in IPC2 for implications, this is not a presupposition-free approach either. The viability of both approaches hinges on the notion of equivalence, that is, the idea that meanings expressed by equivalent propositions (or rules in the foundational approach), one representing the content of introduction-premisses and the other one representing the content of elimination-conclusions, is sufficient to describe harmony.

3.3 The Need for an Intensional Notion of Harmony

Even though the notion of harmony based on the equivalence of c^I and c^E in IPC2 or in the calculus of quantified higher-level rules is highly plausible, a stronger notion can be considered.[7] Let us illustrate this by an example: Suppose we have the set $\wedge\mathscr{I}$ consisting of the standard conjunction introduction

$$\frac{p_1 \quad p_2}{p_1 \wedge p_2}$$

and two alternative sets of elimination rules: $\wedge\mathscr{E}$ consisting of the standard projection rules

$$\frac{p_1 \wedge p_2}{p_1} \qquad \frac{p_1 \wedge p_2}{p_2}$$

and $\wedge\mathscr{E}'$ consisting of the alternative rules

$$\frac{p_1 \wedge p_2}{p_1} \qquad \frac{p_1 \wedge p_2 \quad p_1}{p_2}$$

It is obvious that $\wedge\mathscr{E}$ and $\wedge\mathscr{E}'$ are equivalent to each other, and also equivalent to the rule

$$\frac{p_1 \wedge p_2 \qquad \begin{matrix}[p_1, p_2] \\ q\end{matrix}}{q}$$

which is the canonical generalised elimination rule for $\wedge$. However, do $\wedge\mathscr{E}$ and $\wedge\mathscr{E}'$ mean the same in every possible sense? According to $\wedge\mathscr{E}$, conjunction just expresses pairing, that is, a proof of $p_1 \wedge p_2$ is a pair $\langle \Pi_1, \Pi_2 \rangle$ of proofs, one for p_1 and one for p_2. According to $\wedge\mathscr{E}'$, conjunction expresses something different. A proof of $p_1 \wedge p_2$ is now a pair that consists of a proof of p_1, and a proof of p_2 which is conditional on p_1. Using a functional interpretation of conditional proofs,

[7] The content of this subsection uses ideas presented by Kosta Došen in personal discussion.

this second component can be read as a procedure f that transforms a proof of p_1 into a proof of p_2 so that, according to $\wedge\mathscr{E}'$, conjunction expresses the pair $\langle \Pi_1, f \rangle$. Now $\langle \Pi_1, \Pi_2 \rangle$ and $\langle \Pi_1, f \rangle$ are different. From $\langle \Pi_1, f \rangle$ we can certainly construct the pair $\langle \Pi_1, f(\Pi_1) \rangle$, which is of the desired kind. From the pair $\langle \Pi_1, \Pi_2 \rangle$ we can certainly construct a pair $\langle \Pi_1, f' \rangle$, where f' is the constant function that maps any proof of p_1 to Π_2. However, if we combine these two constructions, we do not obtain what we started with, since we started with an arbitrary function and we end up with a constant function. This is an intuition that is made precise by the consideration that $p_1 \wedge p_2$, where conjunction here has the standard rules $\wedge\mathscr{I}$ and $\wedge\mathscr{E}$, is equivalent to $p_1 \wedge (p_1 \rightarrow p_2)$, but is not isomorphic to it (see Došen [6, 9]). Correspondingly, only $\wedge\mathscr{E}$, but not $\wedge\mathscr{E}'$ is in harmony with $\wedge\mathscr{I}$.

Unlike the notion of equivalence, which only requires a notion of proof in a system, the notion of isomorphism requires a notion of identity of proofs. This is normally achieved by a notion of reduction between proofs, such that proofs that are linked by a chain of reductions are considered identical.[8] In intuitionistic natural deduction these are the reductions reducing maximum formulas (in the case of implication this corresponds to β-reduction), as well as the contractions of an elimination immediately followed by an introduction (in the case of implication this corresponds to η-reduction) and the permutative reductions in the case of disjunction and existential quantification. Using these reductions, moving from $p_1 \wedge p_2$ to $p_1 \wedge (p_1 \rightarrow p_2)$ and back to $p_1 \wedge p_2$ reduces to the identity proof $p_1 \wedge p_2$ (i.e., the formula $p_1 \wedge p_2$ conceived as a proof from itself), whereas conversely, moving from $p_1 \wedge (p_1 \rightarrow p_2)$ to $p_1 \wedge p_2$ and back to $p_1 \wedge (p_1 \rightarrow p_2)$ does not reduce to the identity proof $p_1 \wedge (p_1 \rightarrow p_2)$. In this sense $\wedge\mathscr{E}$ and $\wedge\mathscr{E}'$ cannot be identified.

More precisely, given a formal system together with a notion of identity of proofs, two formulas ψ_1 and ψ_2 are called isomorphic if there are proofs of ψ_2 from ψ_1 and of ψ_1 from ψ_2, such that each of the combination of these proofs (yielding a proof of ψ_1 from ψ_1 and ψ_2 from ψ_2) reduces to the trivial identity proof ψ_1 or ψ_2, respectively. As this notion, which is best made fully precise in categorial terminology, requires not only a notion of proof but also a notion of identity between proofs, it is an intensional notion, distinguishing between possibly different ways of proving something. The introduction of this notion into the debate on harmony calls for a more finegrained analysis. We may now distinguish between purely extensional harmony, which is just based on equivalence and which may be explicated in the ways described in the previous two subsections, and intensional harmony, which requires additional means on the proof-theoretic side based on the way harmonious proof conditions can be transformed into each other.

However, even though the notion of an isomorphism has a clear meaning in a formal system given a notion of identity of proofs, it is not so clear how to use it to define a notion of intensional harmony. The notion of intensional harmony will also be called *strong harmony* in contradistinction to extensional harmony which is also called *weak harmony*.

[8] We consider only a notion of identity that is based on reduction and normalisation. For further options, see Došen [8].

3.4 Towards a Definition of Strong Harmony

For simplicity, take the notion of reductive harmony mentioned in Sect. 3.2. Given introduction rules $c\mathscr{I}$ and elimination rules $c\mathscr{E}$ of an operator c, it associates with c the introduction meaning c^I and the elimination meaning c^E, and identifies extensional harmony with the equivalence of c^I and c^E in IPC2. It then appears to be natural to define intensional harmony as the availability of an *isomorphism* between c^I and c^E in IPC2. However, this definition turns out to be unsuccessful, as the following observation shows.

What we would like to achieve, in any case, is that the canonical eliminations for given introductions are in strong harmony with the introductions, and similarly that the canonical introductions for given eliminations are in strong harmony with the eliminations. The second case is trivial, as the premisses of the canonical introduction are exactly the conclusions of the eliminations. For example, if for the connective $\leftrightarrow$ the elimination rules

$$(\leftrightarrow \mathrm{E})_{can}\ \frac{p_1 \leftrightarrow p_2 \qquad p_1}{p_2} \qquad\qquad \frac{p_1 \leftrightarrow p_2 \qquad p_2}{p_1}$$

are assumed to be given, then its canonical introduction rule has the form

$$(\leftrightarrow \mathrm{I})_{can}\ \frac{\begin{array}{c}[p_1]\\ p_2\end{array} \qquad \begin{array}{c}[p_2]\\ p_1\end{array}}{p_1 \leftrightarrow p_2}$$

Both the elimination meaning $\leftrightarrow^I$ and the introduction meaning $\leftrightarrow^E$ have the form $(p_1 \rightarrow p_2) \wedge (p_2 \rightarrow p_1)$, so that the identity proof identifies them. However, in the first case, where the canonical eliminations are given by the general elimination rules, the situation is more problematic.

Consider disjunction with the rules

$$\frac{p_1}{p_1 \vee p_2} \quad \frac{p_2}{p_1 \vee p_2} \qquad\qquad \frac{p_1 \vee p_2 \qquad \begin{array}{c}[p_1]\\ q\end{array} \qquad \begin{array}{c}[p_2]\\ q\end{array}}{q}$$

Here the introduction meaning of $p_1 \vee p_2$ is $p_1 \vee p_2$, viewed as a formula of IPC2, and its elimination meaning is $\forall q(((p_1 \rightarrow q) \wedge (p_2 \rightarrow q)) \rightarrow q)$. However, though $p_1 \vee p_2$ and $\forall q(((p_1 \rightarrow q) \wedge (p_2 \rightarrow q)) \rightarrow q)$ are equivalent in IPC2, they are not isomorphic. There are proofs

$$\begin{array}{c} p_1 \vee p_2 \\ \mathscr{D}_1 \\ \forall q(((p_1 \rightarrow q) \wedge (p_2 \rightarrow q)) \rightarrow q) \end{array} \qquad\qquad \begin{array}{c} \forall q(((p_1 \rightarrow q) \wedge (p_2 \rightarrow q)) \rightarrow q) \\ \mathscr{D}_2 \\ p_1 \vee p_2 \end{array}$$

of $\forall q(((p_1 \rightarrow q) \wedge (p_2 \rightarrow q)) \rightarrow q)$ from $p_1 \vee p_2$ and of $p_1 \vee p_2$ from $\forall q(((p_1 \rightarrow q) \wedge (p_2 \rightarrow q)) \rightarrow q)$, so that the composition $\mathscr{D}_2 \circ \mathscr{D}_1$ yields

the identity proof $p_1 \vee p_2$, but there are no such proofs so that $\mathscr{D}_1 \circ \mathscr{D}_2$ yields the identity proof $\forall q(((p_1 \to q) \wedge (p_2 \to q)) \to q)$. One might object that, due to the definability of connectives in IPC2, $p_1 \vee p_2$ should be understood as $\forall q(((p_1 \to q) \wedge (p_2 \to q)) \to q)$, so that the isomorphism between $p_1 \vee p_2$ and $\forall q(((p_1 \to q) \wedge (p_2 \to q)) \to q)$ becomes trivial (and similarly, if conjunction is also eliminated due to its definability in IPC2). To accommodate this objection, we consider the example of the trivial connective $+$ with the introduction and elimination rules

$$\frac{p}{+p} \qquad \frac{+p \qquad \begin{array}{c}[p]\\ q\end{array}}{q}$$

Here the elimination rule is the canonical one according to the general-elimination schema. In order to demonstrate strong harmony, we would have to establish the isomorphism of p and $\forall q((p \to q) \to q)$ in IPC2, but this fails. This failure may be related to the fact that for the second-order translations of propositional formulas, we do not have η-conversions in IPC2 (see Girard et al. [27], p. 85[9]). This shows that for the definition of strong harmony the definition of introduction and elimination meaning by translation into IPC2 is perhaps not the best device. We consider the lack of an appropriate definition of strong harmony a major open problem, and we provide two tentative solutions (with the emphasis on 'tentative').

First proposal: Complementation by canonical rules. In order to avoid the problems of second-order logic, we can stay in intuitionistic propositional logic as follows. Suppose for a constant c certain introduction rules $c\mathscr{I}$ and certain elimination rules $c\mathscr{E}$ are proposed, and we ask: When are $c\mathscr{I}$ and $c\mathscr{E}$ in harmony with each other? Suppose $\overline{c\mathscr{I}}$ is the canonical elimination rule for the introduction rules $c\mathscr{I}$, and $\overline{c\mathscr{E}}$ is the canonical introduction rule for the elimination rules $c\mathscr{E}$. We also call $\overline{c\mathscr{I}}$ the *canonical complement* of $c\mathscr{I}$, and $\overline{c\mathscr{E}}$ the *canonical complement* of $c\mathscr{E}$. We define two new connectives c_1 and c_2. Connective c_1 has $c\mathscr{I}$ as its introduction rules and its complement $\overline{c\mathscr{I}}$ as its elimination rule. Conversely, connective c_2 has $c\mathscr{E}$ as its elimination rules and its complement $\overline{c\mathscr{E}}$ as its introduction rules. In other words, for one connective we take the given introduction rules as complemented by the canonical elimination rule, and for the other connective we take the given elimination rules as complemented by the canonical introduction rule. Furthermore, we associate with c_1 and c_2 reduction procedures in the usual way, based on the pairs $c\mathscr{I}/\overline{c\mathscr{I}}$ and $c\mathscr{E}/\overline{c\mathscr{E}}$ as primitive rules. Then we say that $c\mathscr{I}$ and $c\mathscr{E}$ are in *strong harmony*, if c_1 is isomorphic to c_2, that is, if there are proofs from c_1 to c_2 and back, such that the composition of these proofs is identical to the identity proof c_1 or c_2, depending on which side one starts with.[10] In this way, by splitting up c into two connectives, we avoid the explicit translation into IPC2.[11]

[9] This was pointed out to me by Kosta Došen.

[10] For better readability, we omit possible arguments of c_1 and c_2.

[11] This procedure also works for weak harmony as a device to avoid the translation into second-order logic.

Second proposal: Change to the notion of canonical elimination. As mentioned above, we do not encounter any problem in IPC2, if we translate the introduction meaning of disjunction by its disjunction-free second-order translation, as isomorphism is trivial in this case. In fact, whenever we have more than one introduction rule for some c then the disjunction-free second-order translation is identical to the second-order translation of the elimination meaning for the canonical (indirect) elimination. We have a problem in the case of the connective $+$, which has the introduction meaning p and the elimination meaning $\forall q((p \to q) \to q)$. However, for $+$ an alternative elimination rule is derivable, namely the rule

$$\frac{+p}{p}$$

In fact, this sort of elimination rule is available for all connectives with only a single introduction rule. We call it the 'direct' as opposed to the 'indirect' elimination rule. For example, the connective $\&\supset$ with the introduction rule

$$\frac{p_1 \qquad \begin{array}{c}[p_1]\\ p_2\end{array}}{p_1 \,\&\supset\, p_2}$$

has as its direct elimination rules

$$\frac{p_1 \,\&\supset\, p_2}{p_1} \qquad \frac{p_1 \,\&\supset\, p_2 \quad p_1}{p_2}$$

If we require that the canonical elimination rules always be direct where possible, that is, whenever there is not more than one introduction rule, and indirect only if there are multiple introduction rules, then the problem of reduction to IPC2 seems to disappear. In the direct case of a single introduction rule, the elimination meaning is trivially identical to the introduction meaning. In the indirect case, they now become trivially identical again. This is because disjunction, which is used to express the introduction meaning for multiple introduction rules, is translated into disjunction-free second-order logic in a way that makes its introduction meaning identical to its elimination meaning.

This second proposal would require the revision of basic tenets of proof-theoretic semantics, because ever since the work of von Kutschera [36], Prawitz [47] and Schroeder-Heister [53] on general constants, and since the work on general elimination rules, especially for implication, by Tennant [69, 70], López-Escobar [37] and von Plato [43],[12] the idea of the indirect elimination rules as the basic form of elimination rules for *all* constants has been considered a great achievement. That said, the abandonment of projection-based conjunction and modus-ponens-based implication has received some criticism (Dyckhoff [15], Schroeder-Heister [66], Sect. 15.8). In fact, even the first proposal above might require this priority of the direct elimination

[12]For a discussion see Schroeder-Heister [65].

rules. If we consider conjunction with $\wedge\mathscr{I}$ and $\wedge\mathscr{E}$ given by the standard rules

$$\frac{p_1 \quad p_2}{p_1 \wedge p_2} \qquad \frac{p_1 \wedge p_2}{p_1} \quad \frac{p_1 \wedge p_2}{p_2}$$

then $\overline{\wedge\mathscr{I}}$ is the generalised $\wedge$-elimination rule

$$\frac{p_1 \wedge p_2 \qquad \begin{array}{c}[p_1, p_2]\\ q\end{array}}{q}$$

whereas $\overline{\wedge\mathscr{E}}$ is identical with $\wedge\mathscr{I}$. This means that strong harmony would require that projection-based conjunction and conjunction with general elimination are isomorphic, but no such isomorphism obtains.[13]

4 Proof-Theoretic Semantics Beyond Logic

Proof-theoretic semantics has been occupied almost exclusively with logical reasoning, and, in particular, with the meaning of logical constants. Even though the way we can acquire knowledge logically is extremely interesting, this is not and should not form the central pre-occupation of proof-theoretic semantics. The methods used in proof-theoretic semantics extend beyond logic, often so that their application in logic is nothing but a special case of these more general methods.

What is most interesting is the handling of reasoning with information that is incorporated into sentences, which, from the viewpoint of logic, are called 'atomic'. A special way of providing such information, as long as we are not yet talking about empirical knowledge, is by definitions. By defining terms, we introduce claims into our reasoning system that hold in virtue of the definition. In mathematics the most prominent example is inductive definitions. Now definitional reasoning itself obeys certain principles that we find otherwise in proof-theoretic semantics. As an inductive definition can be viewed as a set of rules the heads of which contain the *definiendum* (for example, an atomic formula containing a predicate to be defined), it is only natural to consider inductive clauses as kinds of introduction rules, suggesting a straightforward extension of principles of proof-theoretic semantics to the atomic case. A particular challenge here comes from logic programming, where we consider inductive definitions of a certain kind, called 'definite-clause programs', and use them not only for descriptive, but also for computational purposes. In the context of dealing with negation, we even have the idea of inverting clauses in a certain sense. Principles such as the 'completion' of logic programs or the 'closed-world assumption' (which logic programming borrowed from Artificial Intelligence research), are

[13]This last observation is due to Luca Tranchini.

strongly related to principles generating elimination rules from introduction rules and, thus, to the idea of harmony between these rules.

4.1 Definitional Reflection

In what follows, we sketch the idea of *definitional reflection*, which employs the idea of clausal definitions as a powerful paradigm to extend proof-theoretic semantics beyond the specific realm of logic. It is related to earlier approaches developed by Lorenzen [38, 39] who based logic (and also arithmetic and analysis) on a general theory of admissible rules using a sophisticated inversion principle (he coined the term 'inversion principle' and was the first to formulate it in a precise way). It is also related to Martin-Löf's [40] idea of iterated inductive definitions, which gives introduction and elimination rules for inductively defined atomic sentences. Moreover, it is inspired by ideas in logic programming, where programs can be read as inductive definitions and where, in the attempt to provide a satisfactory interpretation of negation, ideas that correspond to the inversion of rules have been considered (see Denecker et al. [5], Hallnäs and Schroeder-Heister [31]). We take definitional reflection as a specific example of how proof-theoretic semantics can be extended beyond logic, and we claim that such an extension is quite useful. Other extensions beyond logic are briefly mentioned at the end of this section.

A particular advantage that distinguishes definitional reflection from the approaches of Lorenzen and Martin-Löf and makes it more similar to what has been done in logic programming is the idea that the meaning assignment by means of a clausal or inductive definition can be partial, which means in particular that definitions need not be well-founded. In logic programming this has been common from the very beginning. For example, clauses such as

$$p \Leftarrow \neg p$$

which defines p by its own negation, or related circular clauses have been standard examples for decades in the discussion of normal logic programs and the treatment of negation (see, e.g. Gelfond and Lifschitz [24], Gelder et al. [23]). Within mainstream proof-theoretic semantics, such circular definitions have only recently garnered attention, in particular within the discussion of paradoxes, mostly without awareness of logic programming semantics and developments there. The idea of definitional reflection can be used to incorporate smoothly partial meaning and non-wellfounded definitions. We consider definitional reflection as an example of how to move beyond logic and, with it, beyond the totality and well-foundedness assumptions of the proof-theoretic semantics of logic.

As definitional reflection is a local approach not based on the placeholder view of assumptions, we formulate it in a sequent-style framework. A *definition* is a list of clauses. A clause has the form

$$a \Leftarrow B$$

where the head a is an atomic formula ('atom'). In the simplest case, the body B is a list of atoms $b_1, \ldots, b_m$, in which case a definition looks like a definite logic program. We often consider an extended case where B may also contain structural implication[14] '$\Rightarrow$', and sometimes even structural universal implication, which essentially is handled by restricting substitution. Given a definition $\mathbb{D}$, the list of clauses with a head starting with the predicate P is called the *definition of* P. In the propositional case where atoms are just propositional letters, we speak of the *definition of* a having the form

$$\mathbb{D}_a \begin{cases} a \Leftarrow B_1 \\ \quad \vdots \\ a \Leftarrow B_n \end{cases}$$

However, it should be clear that the definition of P or of a is normally just a particular part of a definition $\mathbb{D}$, which contains clauses for other expressions as well. It should also be clear that this definition $\mathbb{D}$ cannot always be split up into separate definitions of its predicates or propositional letters. So 'definition of a' or 'of P' is a mode of speech. What is always meant is the list of clauses for a predicate or propositional letter within a definition $\mathbb{D}$.

Syntactically, a clause resembles an introduction rule. However, in the theory of definitional reflection we separate the definition, which is incorporated in the set of clauses, from the inference rules, which put it into practice. So, instead of different introduction rules which define different expressions, we have a general schema that applies to a given definition. Separating the specific definition from the inference schema using arbitrary definitions gives us wider flexibility. We need not consider introduction rules to be basic and other rules to be derived from them. Instead we can speak of certain inference principles that determine the inferential meaning of a clausal definition and which are of equal stance. There is a pair of inference principles that put a definition into action, which are in harmony with each other, without one of them being preferential. As we are working in a sequent-style framework, we have inferential principles for introducing the defined constant on the right and on the left of the turnstile, that is, in the assertion and in the assumption positions. For simplicity we consider the case of a propositional definition $\mathbb{D}$, which has no predicates, functions, individual variables or constants, and in which the bodies of clauses are just lists of propositional letters. Suppose $\mathbb{D}_a$ (as above) is the definition

[14] We speak of 'structural' implication to distinguish it from the implicational sentence connective which may form part of a defined atom. Some remarks on this issue are made in Sect. 4.2.

of a (within $\mathbb{D}$), and the B_i have the form '$b_{i1}, \ldots, b_{ik_i}$', as in propositional logic programming. Then the right-introduction rules for a are

$$(\vdash a)\ \frac{\Gamma \vdash b_{i1} \qquad \ldots \qquad \Gamma \vdash b_{ik_i}}{\Gamma \vdash a}, \quad \text{in short} \quad \frac{\Gamma \vdash B_i}{\Gamma \vdash a} \quad (1 \leq i \leq n),$$

and the left-introduction rule for a is

$$(a \vdash)\ \frac{\Gamma, B_1 \vdash C \qquad \ldots \qquad \Gamma, B_n \vdash C}{\Gamma, a \vdash C}$$

If we talk generically about these rules, that is, without mentioning a specific a, but just the definition $\mathbb{D}$, we also write $(\vdash \mathbb{D})$ and $(\mathbb{D} \vdash)$. The right introduction rule expresses reasoning 'along' the clauses. It is also called *definitional closure*, by which is meant 'closure under the definition'. The intuitive meaning of the left introduction rule is the following: Everything that follows from every possible *definiens* of a, follows from a itself. This rule is called the *principle of definitional reflection*, as it reflects upon the definition as a whole. If $B_1, \ldots, B_n$ exhaust *all possible conditions* to generate a according to the given definition, and if each of these conditions entails the very same conclusion, then a itself entails this conclusion.

This principle, which gives the whole approach its name, extracts deductive *consequences* of a from a definition in which only the defining *conditions* of a are given. If the clausal definition $\mathbb{D}$ is viewed as an inductive definition, definitional reflection can be viewed as being based on the extremal clause of $\mathbb{D}$: Nothing else beyond the clauses given in $\mathbb{D}$ defines a. To give a very simple example, consider the following definition:

$$\begin{cases} \text{child\_of\_tom} \Leftarrow \text{anna} \\ \text{child\_of\_tom} \Leftarrow \text{robert} \end{cases}$$

Then one instance of the principle of definitional reflection with respect to this definition is

$$\frac{\text{anna} \vdash \text{tall} \quad \text{robert} \vdash \text{tall}}{\text{child\_of\_tom} \vdash \text{tall}}$$

Therefore, if we know anna $\vdash$ tall and robert $\vdash$ tall, we can infer child_of_tom $\vdash$ tall.

Since definitional reflection depends on the definition as a whole, taking *all definientia* of a into account, it is non-monotonic with respect to $\mathbb{D}$. If $\mathbb{D}$ is extended with an additional clause

$$a \Leftarrow B_{n+1}$$

for a, then previous applications of the $(\mathbb{D} \vdash)$ rule may no longer remain valid. In the present example, if we add the clause

$$\text{child\_of\_tom} \Leftarrow \text{john}$$

we can no longer infer child_of_tom $\vdash$ tall, except when we also know john $\vdash$ tall. Note that due to the definitional reading of clauses, which gives rise to inversion, the sign '$\Leftarrow$' expresses more than just implication, in contradistinction to structural implication '$\Rightarrow$' that may occur in the body of a clause. To do justice to this fact, one might instead use ':-' as in PROLOG, or ':=' to express that we are dealing with some sort of definitional equality.

In standard logic programming one has, on the declarative side, only what corresponds to definitional closure. Definitional reflection leads to powerful extensions of logic programming (due to computation procedures based on this principle) that lie beyond the scope of the present discussion.

4.2 Logic, Paradoxes, Partial Definitions

Introduction rules (clauses) for logically compound formulas are not distinguished in principle from introduction rules (clauses) for atoms. The introduction rules for conjunction and disjunction would, for example, be handled by means of clauses for a truth predicate with conjunction and disjunction as term-forming operators:

$$\mathbb{D}_{log} \begin{cases} T(p \wedge q) \Leftarrow T(p), T(q) \\ T(p \vee q) \Leftarrow T(p) \\ T(p \vee q) \Leftarrow T(q) \end{cases}$$

In order to define implication, we need a rule arrow in the body, which, for the whole clause, corresponds to using a higher-level rule:

$$T(p \rightarrow q) \Leftarrow (T(p) \Rightarrow T(q))$$

This definition requires some sort of 'background logic'. By that we mean the structural logic governing the comma and the rule arrow $\Rightarrow$, which determine how the bodies of clauses are handled. In standard logic we have just the comma, which is handled implicitly. In extended versions of logic programming we would have the (iterated) rule arrow, that is, structural implication and associated principles governing it, and perhaps even structural disjunction (this is present in disjunctive logic programming, but not needed for the applications considered here).

It is obvious that ($\vdash \mathbb{D}_{log}$) gives us the right-introduction rules for conjunction and disjunction or, more precisely, those for $T(A)$, where A is a conjunction or disjunction. Definitional reflection ($\mathbb{D}_{log} \vdash$) gives us the left-introduction rules. The clause for $T(p \rightarrow q)$ gives us the rules for implication, where the precise formulation of these rules depends on the exact formulation of the background logic governing $\Rightarrow$.

Definitional reflection in general provides a much wider perspective on inversion principles than deductive logic alone. Using the definitional rule

$$t \in \{x : a\} \Leftarrow a[t/x]$$

we obtain a principle of naive comprehension, which does not lead to a useless theory in which everything is derivable, even if we allow a to be the formula $x \notin x$ and t the term $\{x : x \notin x\}$. A definition of the form

$$p \Leftarrow \neg p$$

yields a paraconsistent system in which both $\vdash p$ and $\vdash \neg p$ are derivable, without every other formula being derivable. Formally, this means that the rule of cut

$$\frac{\Gamma \vdash A \quad A, \Delta \vdash B}{\Gamma, \Delta \vdash B}$$

is not always admissible. For special cases cut can be obtained, for example, if the definition is stratified, which essentially means that it is well-founded. So the well-behaviour of a definition in the case of logic, where we do have cut elimination, is due to the fact that it obeys certain principles, which in the general case cannot be expected to hold. This connects the proof theory of clausal definitions with theories of paradoxes, which conceive paradoxes as based on locally correct reasoning (Prawitz [44] (Appendix B), Tennant [68], Schroeder-Heister [62], Tranchini [71]).

For the situation that obtains here, Hallnäs [28] proposed the terms 'total' vs. 'partial' in analogy with the terminology used in recursive function theory. That a computable (i.e., partial recursive) function is total is not something required by definition, but is a matter of (mathematical) fact, actually an undecidable matter. Similarly, that a clausal definition yields a system that admits the elimination of cuts is a result that may or may not hold true, but nothing that should enter the requirements for something to be admitted as a definition. If it holds, the definition is called 'total', otherwise it is properly 'partial'.

4.3 Variables and Substitution

The idea of proof-theoretic semantics beyond logic invites consideration of powerful inversion principles that extend the simple form of definitional reflection considered above. Although we mentioned clauses that contained variables, we formally defined definitional reflection only for propositional definitions of the form $\mathbb{D}_a$, where it says that everything that can be inferred from each *definiens* can be obtained from the *definiendum* of a definition. However, this is insufficient for the more general case which is the standard case in logic programming. We show this by means of an example. Suppose we have the following definition in which the atoms have a predicate-argument-structure:

$$\left\{\begin{array}{rcl} \text{child\_of\_tom(anna)} & \Leftarrow & \text{daughter\_of\_tom(anna)} \\ \text{child\_of\_tom(robert)} & \Leftarrow & \text{son\_of\_tom(robert)} \\ \text{tall(anna)} & \Leftarrow & \text{daughter\_of\_tom(anna)} \\ \text{tall(robert)} & \Leftarrow & \text{son\_of\_tom(robert)} \end{array}\right.$$

Given our propositional rule of definitional reflection, we could just infer propositional results such as child_of_tom(anna) $\vdash$ tall(anna) or child_of_tom(robert) $\vdash$ tall (robert). However, what we would like to infer is the principle

$$\text{child\_of\_tom}(x) \vdash \text{tall}(x)$$

with free variable x, since anna and robert are the only objects for which the predicate child_of_tom is defined, and since for them the desired principle holds.

In an even more general case, we have clauses that contain variables the instances of which match instances of the claim we want to obtain by definitional reflection. Consider the definition

$$\mathbb{D}_1 \left\{\begin{array}{rcl} p(y, a) & \Leftarrow & q(y) \\ p(x, f(a)) & \Leftarrow & q(f(x)) \end{array}\right.$$

According to the principle of *general definitional reflection*, we obtain, with respect to $\mathbb{D}_1$,

$$p(a, z) \vdash q(z)$$

The intuitive argument is as follows: Suppose $p(a, z)$. Any, and in fact the only, substitution instance of $p(a, z)$ that can be obtained by the first clause is generated by substituting a for y in the first clause, and a for z in $p(a, z)$. We denote this substitution by $[a/y, a/z]$ and call it σ_1. Any, and in fact the only, substitution instance of $p(a, z)$ that can be obtained by the second clause is generated by substituting a for x in the second clause, and $f(a)$ for z in $p(a, z)$. We denote this substitution by $[a/x, f(a)/z]$ and call it σ_2. When σ_1 is applied to the body of the first clause, $q(a)$ is obtained, which is also obtained when σ_1 is applied to $q(z)$. When σ_2 is applied to the body of the second clause, $q(f(a))$ is obtained, which is also obtained when σ_2 is applied to $q(z)$. Therefore, we can conclude $q(z)$.

In the propositional case we could describe definitional reflection by saying that every C that is a consequence of each defining condition is a consequence of the *definiendum* a. We cannot now even identify a single formula as a *definiendum*, as any formula which is a substitution instance of a head of a definitional clause is considered to be defined. Therefore we should now say: Suppose a formula a is given. If for each substitution instance $a\sigma$ that can be obtained as $b\sigma$ from the head of a clause $b \Leftarrow C$, we have that $C\sigma$ implies $A\sigma$ for some A, then A can be inferred from a.

Formally, this leads to a principle of definitional reflection according to which, for the introduction of an atom a on the left side of the turnstile, the most general unifiers $mgu(a, b)$ of a with the heads of all definitional clauses are considered:

$$(\mathbb{D}\vdash)_\omega \; \frac{\{\Gamma\sigma, C\sigma \vdash A\sigma : \sigma = mgu(a, b) \text{ for some clause } b \Leftarrow C \text{ in } \mathbb{D}\}}{\Gamma, a \vdash A}$$

with the proviso: The variables free in $\Gamma, a \vdash A$ must be different from those in the $b \Leftarrow C$ above the line. This means that we always assume that variables in clauses are standardised apart. We call this principle the ω-version of definitional reflection, as it is in a certain way related to the ω-rule in arithmetic, a point that we cannot elaborate on here (see Schroeder-Heister [55]).

This powerful principle is typical of applications outside logic. When we consider logical definitions such as $\mathbb{D}_{log}$, we see that each formula $T(a)$, where a is a compound logical formula, determines exactly a single head b in one or two clauses (two in the case of disjunction), such that $T(a)$ is a substitution instance of b. This means that (1) there is just matching and no unification between b and $T(a)$, and (2) there is just a single substitution for the main logical connective in $T(a)$ involved due to the strict separation between the clauses for the different logical connectives, implying that there is no overlap between substitution instances of clauses.

The power of the ω-version of definitional reflection is demonstrated, for example, by the fact that the rules of free equality can be obtained from the definition consisting of the single clause

$$\mathbb{D}_= \{ \; x = x \Leftarrow$$

For example, the transitivity of equality is derived by a single inference step as follows:

$$(\mathbb{D}_= \vdash)_\omega \; \frac{x_2 = x_3 \vdash x_2 = x_3}{x_1 = x_2, x_2 = x_3 \vdash x_1 = x_3}$$

Here we use that the substitution $[x_2/x_1, x_2/x]$ is an *mgu* of $x_1 = x_2$ with the head $x = x$ of the clause $x = x \Leftarrow$. In a similar way, all freeness axioms of Clark's equational theory [35] can be derived.[15]

4.4 Outlook: Applications and Extensions of Definitional Reflection

We have not discussed computational issues here. Clausal definitions give rise to computational procedures as investigated and implemented in logic programming,

[15]For further discussion see Girard [26], Schroeder-Heister [32], Eriksson [16, 17], Schroeder-Heister [54]. Of all inversion principles mentioned in the literature, only Lorenzen's original one [39] comes close to the power of definitional reflection (though substantial differences remain, see Schroeder-Heister [57]).

and definitional reflection adds a strong component to such computation (see Hallnäs and Schroeder-Heister [31], Eriksson [17]). This computational aspect is important to proof-theoretic semantics. We should not only be able to give a definition of a semantically correct proof, where 'semantically' is understood in the sense of proof-theoretic semantics, we should also be interested in ways to construct such proofs that proceed according to such principles. Programming languages, theorem provers and proof editors based on inversion principles make important contributions to this task. Devising principles of proof construction that can be used for proof search is itself an issue of proof-theoretic semantics that is a *desideratum* in the philosophically dominated community. Theories that go beyond logic are of particular interest here, as theorems outside pure logic are what we normally strive for in reasoning.

If we want to deal with more advanced mathematical theories, stronger closure and reflection principles are needed. At an elementary level, clauses $a \Leftarrow B$ in a definition can be used to describe function computation in the form

$$f(x_1, \ldots, x_k) \Leftarrow (x_1, \ldots, x_k)$$

which is supposed to express that from the arguments $x_1, \ldots, x_k$ the value $f(x_1, \ldots, x_k)$ is obtained, so that by means of definitional reflection $f(x_1, \ldots, x_k)$ can be computed. More generally, one might describe functionals F by means of (infinitary) clauses the bodies of which describe the evaluation of functions f which are arguments of F (for some hints see Hallnäs [29, 30]). An instructive example is the analysis of abstract syntax (see McDowell and Miller [41]).

There are several other approaches that deal with the atomic level proof-theoretically, that is, with issues beyond logic in the narrower sense. These approaches include Negri and von Plato's [42] proof analysis, Brotherston and Simpson's [2] infinite derivations, or even derivations concerning subatomic expressions (see Więckowski [73]), and corresponding linguistic applications, as discussed by Francez and Dyckhoff [19] and Francez et al. [20]. Proof-theoretic semantics beyond logic is a broad field with great potential, the surface of which, thus far, has barely been scratched.

Acknowledgments I am grateful to Kosta Došen, Thomas Piecha, Dag Prawitz, Luca Tranchini and John William Devine for helpful comments and suggestions. The completion of this paper was supported by the French-German ANR-DFG project 'Beyond Logic: Hypothetical Reasoning in Philosophy of Science, Informatics, and Law' (DFG Schr 275/17-1).

References

1. Bolzano, B.: Wissenschaftslehre. Versuch einer ausführlichen und größtentheils neuen Darstellung der Logik mit steter Rücksicht auf deren bisherige Bearbeiter, vol. I–IV. Seidel, Sulzbach (1837)
2. Brotherston, J., Simpson, A.: Complete sequent calculi for induction and infinite descent. In: Proceedings of the 22nd Annual IEEE Symposium on Logic in Computer Science (LICS), pp. 51–62. IEEE Press, Los Alamitos (2007)
3. Carnap, R.: Abriß der Logistik. Mit besonderer Berücksichtigung der Relationstheorie und ihrer Anwendungen. Springer, Wien (1929)
4. Dean, W., Kurokawa, H.: Kreisel's Theory of Constructions, the Kreisel-Goodman paradox, and the second clause. In: Piecha, T., Schroeder-Heister, P. (eds.) Advances in Proof-Theoretic Semantics. Springer, Dordrecht (2016) (This volume)
5. Denecker, M., Bruynooghe, M., Marek, V.: Logic programming revisited: logic programs as inductive definitions. ACM Trans. Comput. Log. **2**, 623–654 (2001)
6. Došen, K.: Logical consequence: a turn in style. In: Dalla Chiara, M.L., Doets, K., Mundici, D., van Benthem, J. (eds.) Logic and Scientific Methods: Volume One of the Tenth International Congress of Logic, Methodology and Philosophy of Science, Florence, August 1995, pp. 289–311. Kluwer, Dordrecht (1997)
7. Došen, K.: Cut Elimination in Categories. Springer, Berlin (2000)
8. Došen, K.: Identity of proofs based on normalization and generality. Bull. Symb. Log. **9**, 477–503 (2003)
9. Došen, K.: Models of deduction. Synthese **148**, 639–657. Special issue: Kahle, R., Schroeder-Heister, P. (eds.) Proof-Theoretic Semantics (2006)
10. Došen, K.: Comments on an opinion. In: Piecha, T., Schroeder-Heister, P. (eds.) Advances in Proof-Theoretic Semantics. Springer, Dordrecht (2016) (This volume)
11. Došen, K.: On the paths of categories. In: Piecha, T., Schroeder-Heister, P. (eds.) Advances in Proof-Theoretic Semantics. Springer, Dordrecht (2016) (This volume)
12. Došen, K., Petrić, Z.: Proof-Theoretical Coherence. College Publications, London (2004)
13. Dummett, M.: Frege: Philosophy of Language. Duckworth, London (1973)
14. Dummett, M.: The Logical Basis of Metaphysics. Duckworth, London (1991)
15. Dyckhoff, R.: Some remarks on proof-theoretic semantics. In: Piecha, T., Schroeder-Heister, P. (eds.) Advances in Proof-Theoretic Semantics. Springer, Dordrecht (2016) (This volume)
16. Eriksson, L.-H.: A finitary version of the calculus of partial inductive definitions. In: Eriksson, L.-H., Hallnäs, L., Schroeder-Heister, P. (eds.) Extensions of Logic Programming. Second International Workshop, ELP '91, Stockholm, January 1991, Proceedings. Lecture Notes in Computer Science, vol. 596, pp. 89–134. Springer, Berlin (1992)
17. Eriksson, L.-H.: Finitary partial inductive definitions and general logic. Ph.D. Thesis. Royal Institute of Technology, Stockholm (1993)
18. Fitch, F.B.: Symbolic Logic: An Introduction. Ronald Press, New York (1952)
19. Francez, N., Dyckhoff, R.: Proof-theoretic semantics for a natural language fragment. Linguist. Philos. **33**, 447–477 (2010)
20. Francez, N., Dyckhoff, R., Ben-Avi, G.: Proof-theoretic semantics for subsentential phrases. Studia Logica **94**, 381–401 (2010)
21. Frege, G.: Grundgesetze der Arithmetik. Begriffsschriftlich abgeleitet, vol. I. Hermann Pohle, Jena (1893)
22. Frege, G.: Logische Untersuchungen. Dritter Teil: Gedankengefüge. Beiträge zur Philosophie des deutschen Idealismus **3**, 36–51 (1923)
23. Gelder, A.V., Ross, K.A., Schlipf, J.S.: The well-founded semantics for general logic programs. J. Assoc. Comput. Mach. **38**, 620–650 (1991)
24. Gelfond, M., Lifschitz, V.: The stable model semantics for logic programming. In: Proceedings of the 5th International Conference and Symposium on Logic Programming, pp. 1070–1080. IEEE, New York (1988)

25. Gentzen, G.: Untersuchungen über das logische Schließen. Mathematische Zeitschrift **39**, 176–210, 405–431 (1934/35). English translation in: Szabo, M. E. (ed.) The Collected Papers of Gerhard Gentzen, North Holland, Amsterdam 1969, pp. 68–131
26. Girard, J.-Y.: A fixpoint theorem in linear logic. Linear Logic Mailing List (linear@cs.stanford.edu) 6 February 1992
27. Girard, J.-Y., Lafont, Y., Taylor, P.: Proofs and Types. Cambridge University Press, Cambridge (1989)
28. Hallnäs, L.: Partial inductive definitions. Theor. Comput. Sci. **87**, 115–142 (1991)
29. Hallnäs, L.: On the proof-theoretic foundation of general definition theory. Synthese **148**, 589–602 (2006)
30. Hallnäs, L.: On the proof-theoretic foundations of set theory. In: Piecha, T., Schroeder-Heister, P. (eds.) Advances in Proof-Theoretic Semantics. Springer, Dordrecht (2016) (This volume)
31. Hallnäs, L., Schroeder-Heister, P.: A proof-theoretic approach to logic programming: I. Clauses as rules. II. Programs as definitions. J. Log. Comput. **1**, 261–283, 635–660 (1990/91)
32. Hallnäs, L., Schroeder-Heister, P.: Girard's fixpoint theorem. Linear Logic Mailing List (linear@cs.stanford.edu) 19 February 1992
33. Hodges, W.: A strongly differing opinion on proof-theoretic semantics? In: Piecha, T., Schroeder-Heister, P. (eds.) Advances in Proof-Theoretic Semantics. Springer, Dordrecht (2016) (This volume)
34. Jaśkowski, S.: On the rules of suppositions in formal logic. Stud. Log. **1**, 5–32 (1934). Reprinted in: McCall, S. (ed.), Polish Logic 1920–1939, Oxford 1967, pp. 232–258
35. Kunen, K.: Negation in logic programming. J. Log. Program. **4**, 289–308 (1987)
36. von Kutschera, F.: Die Vollständigkeit des Operatorensystems $\{\neg, \wedge, \vee, \supset\}$ für die intuitionistische Aussagenlogik im Rahmen der Gentzensemantik. Archiv für mathematische Logik und Grundlagenforschung **11**, 3–16 (1968)
37. López-Escobar, E.G.K.: Standardizing the N systems of Gentzen. In: Caicedo, X., Montenegro, C.H. (eds.), Models, Algebras, and Proofs, pp. 411–434. Dekker, New York (1999)
38. Lorenzen, P.: Konstruktive Begründung der Mathematik. Mathematische Zeitschrift **53**, 162–202 (1950)
39. Lorenzen, P.: Einführung in die operative Logik und Mathematik. Springer, Berlin (1955). 2nd edn. 1969
40. Martin-Löf, P.: Hauptsatz for the intuitionistic theory of iterated inductive definitions. In: Fenstad, J.E. (ed.) Proceedings of the Second Scandinavian Logic Symposium, pp. 179–216. North-Holland, Amsterdam (1971)
41. McDowell, R., Miller, D.: A logic for reasoning with higher-order abstract syntax. In: Logic in Computer Science (LICS 1997, Warsaw), pp. 434–445. IEEE Computer Society (1997)
42. Negri, S., von Plato, J.: Proof Analysis: A Contribution to Hilbert's Last Problem. Cambridge University Press, Cambridge (2011)
43. von Plato, J.: Natural deduction with general elimination rules. Arch. Math. Log. **40**, 541–567 (2001)
44. Prawitz, D.: Natural Deduction: A Proof-Theoretical Study. Almqvist & Wiksell, Stockholm (1965). Reprinted Dover Publications, Mineola NY 2006
45. Prawitz, D.: Ideas and results in proof theory. In: Fenstad, J.E. (ed.) Proceedings of the Second Scandinavian Logic Symposium (Oslo 1970), pp. 235–308. North-Holland, Amsterdam (1971)
46. Prawitz, D.: On the idea of a general proof theory. Synthese **27**, 63–77 (1974)
47. Prawitz, D.: Proofs and the meaning and completeness of the logical constants. In: Hintikka, J., Niiniluoto, I., Saarinen, E. (eds.) Essays on Mathematical and Philosophical Logic: Proceedings of the Fourth Scandinavian Logic Symposium and the First Soviet-Finnish Logic Conference, Jyväskylä, Finland, June 29—July 6, 1976, pp. 25–40. Kluwer, Dordrecht (1979). Revised German translation 'Beweise und die Bedeutung und Vollständigkeit der logischen Konstanten', Conceptus **16**, 31–44 (1982)
48. Prawitz, D.: Remarks on some approaches to the concept of logical consequence. Synthese **62**, 152–171 (1985)

49. Prawitz, D.: Pragmatist and verificationist theories of meaning. In: Auxier, R.E., Hahn, L.E. (eds.) The Philosophy of Michael Dummett, pp. 455–481. Open Court, Chicago (2007)
50. Prawitz, D.: Inference and knowledge. In: Peliš, M. (ed.) The Logica Yearbook 2008, pp. 175–192. College Publications, London (2009)
51. Read, S.: General-elimination harmony and the meaning of the logical constants. J. Philos. Log. **39**, 557–576 (2010)
52. Read, S.: General-elimination harmony and higher-level rules. In: Wansing, H. (ed.) Dag Prawitz on Proofs and Meaning, pp. 293–312. Springer, Cham (2015)
53. Schroeder-Heister, P.: A natural extension of natural deduction. J. Symb. Log. **49**, 1284–1300 (1984)
54. Schroeder-Heister, P.: Cut elimination in logics with definitional reflection. In: Pearce, D., Wansing, H., (eds) Nonclassical Logics and Information Processing, International Workshop, Berlin, November 1990, Proceedings (Lecture Notes in Computer Science vol. 619) Springer, Berlin, pp. 146–171 (1992)
55. Schroeder-Heister, P.: Rules of definitional reflection. In: Proceedings of the 8th Annual IEEE Symposium on Logic in Computer Science (Montreal 1993), pp. 222–232. IEEE Press, Los Alamitos (1993)
56. Schroeder-Heister, P.: Validity concepts in proof-theoretic semantics. Synthese **148**, 525–571. Special issue: Kahle R., Schroeder-Heister, P. (eds.) Proof-Theoretic Semantics (2006)
57. Schroeder-Heister, P.: Generalized definitional reflection and the inversion principle. Logica Universalis **1**, 355–376 (2007)
58. Schroeder-Heister, P.: Proof-theoretic versus model-theoretic consequence. In: Peliš, M. (ed.) The Logica Yearbook 2007, pp. 187–200. Filosofia, Prague (2008)
59. Schroeder-Heister, P.: Sequent calculi and bidirectional natural deduction: on the proper basis of proof-theoretic semantics. In: Peliš, M. (ed.) The Logica Yearbook 2008, pp. 237–251. College Publications, London (2009)
60. Schroeder-Heister, P.: The categorical and the hypothetical: A critique of some fundamental assumptions of standard semantics. Synthese, **187**, 925–942. Special issue: Lindström, S., Palmgren E., Westerståhl, D. (eds.) The Philosophy of Logical Consequence and Inference (2012)
61. Schroeder-Heister, P.: Proof-theoretic semantics. In: Zalta, E.N. (ed.) Stanford Encyclopedia of Philosophy, Stanford (2012) http://plato.stanford.edu/entries/proof-theoretic-semantics
62. Schroeder-Heister, P.: Proof-theoretic semantics, self-contradiction, and the format of deductive reasoning. Topoi **31**, 77–85 (2012)
63. Schroeder-Heister, P.: The calculus of higher-level rules, propositional quantifiers, and the foundational approach to proof-theoretic harmony. Studia Logica, **102**, 1185–1216. Special issue: Indrzejczak, A. (ed.) Gentzen's and Jaśkowski's Heritage: 80 Years of Natural Deduction and Sequent Calculi (2014)
64. Schroeder-Heister, P.: Frege's sequent calculus. In: Indrzejczak, A., Kaczmarek, J., Zawidzki, M. (eds.) Trends in Logic XIII: Gentzen's and Jaśkowski's Heritage—80 Years of Natural Deduction and Sequent Calculi, pp. 233–245. Łódź University Press, Łódź (2014)
65. Schroeder-Heister, P.: Generalized elimination inferences, higher-level rules, and the implications-as-rules interpretation of the sequent calculus. In: Pereira, L.C., Haeusler, E.H., de Paiva, V. (eds.) Advances in Natural Deduction: A Celebration of Dag Prawitz's Work, pp. 1–29. Springer, Heidelberg (2014)
66. Schroeder-Heister, P.: Harmony in proof-theoretic semantics: A reductive analysis. In: Wansing, H. (ed.) Dag Prawitz on Proofs and Meaning, pp. 329–358. Springer, Cham (2015)
67. Schroeder-Heister, P.: Proof-theoretic validity based on elimination rules. In: Haeusler, E.H., de Campos Sanz, W., Lopes, B. (eds.) Why is this a proof? Festschrift for Luiz Carlos Pereira, pp. 159–176. College Publications, London (2015)
68. Tennant, N.: Proof and paradox. Dialectica **36**, 265–296 (1982)
69. Tennant, N.: Autologic. Edinburgh University Press, Edinburgh (1992)
70. Tennant, N.: Ultimate normal forms for parallelized natural deductions. Log. J. IGPL **10**, 299–337 (2002)

71. Tranchini, L.: Proof-theoretic semantics, paradoxes, and the distinction between sense and denotation. J. Log. Comput. (2015) Published online June 2014
72. Wansing, H.: The idea of a proof-theoretic semantics. Studia Logica **64**, 3–20 (2000)
73. Więckowski, B.: Rules for subatomic derivation. Rev. Symb. Log. **4**, 219–236 (2011)