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 potential 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 $(\mathrm{P}_{\rightarrow})$, $(\mathrm{P}_{\neg})$, and $(\mathrm{P}_{\forall})$ 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 $(\mathrm{P}_{\neg})$.

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 ($\mathrm{P}_{\rightarrow}$), ($\mathrm{P}_{\neg}$), and ($\mathrm{P}_{\forall}$) 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.) ($P_{\to}$) 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 ($P_{\to}$), ($P_{\neg}$), and ($P_{\forall}$) 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 ($P_{\to}$):

($P^2_{\to}$) 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 ($P_{\to}$) 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 ($\mathrm{P}_{\rightarrow}$) 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 ($\mathrm{P}_{\neg}$) and ($\mathrm{P}_{\forall}$).

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 ($\mathrm{P}^2_{\rightarrow}$). 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 ($\mathrm{P}^2_{\rightarrow}$) 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 ($\mathrm{P}^2_{\rightarrow}$) 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 $\texttt{Proof}_{\mathsf{T}}(x, y)$ 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:

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

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:

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

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]:

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

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 πst, Kreisel's clauses can be understood as defining $\Pi(A, s)$ in terms of πxy 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 $\vdash_{\mathscr{T}^+} \pi sc \equiv \top$.

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. $P(\ulcorner A\urcorner) \rightarrow A$.

(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) $\mathsf{T}^+ \vdash P(\ulcorner D\urcorner) \rightarrow D$ 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* $Y =_{\text{df}} \lambda f.(\lambda x.f(xx))(\lambda x.f(xx))$. 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 $h(y,x) =_{\text{df}} \lambda y.\lambda x.(\pi yx \supset_1 \bot)$. 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]:

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

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 $R(A, p_1)$, $R(p_2, R(A, p_2))$, $R(p_3, R(p_2, R(A, p_2)))$, ... 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 ($\mathrm{P}_{\to}$), ($\mathrm{P}_{\neg}$), and ($\mathrm{P}_{\forall}$), 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 $\to$, $\neg$ 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 ($\mathrm{P}_{\to}$), ($\mathrm{P}_{\neg}$), and ($\mathrm{P}_{\forall}$) 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 ($\mathrm{K}_{\rightarrow}$), ($\mathrm{K}_{\neg}$), or ($\mathrm{K}_{\forall}$).

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 ($\mathrm{P}^2_{\rightarrow}$), ($\mathrm{P}^2_{\neg}$), and ($\mathrm{P}^2_{\forall}$) 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 $[\Pi(A \rightarrow B, y)]$ 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 $\Pi(A \rightarrow B, y)$—i.e. the proof condition for the implication $A \rightarrow B$—in the same manner as Kreisel's second clause variant ($\mathrm{K}_{\rightarrow}$). 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 ($\mathrm{P}_{\rightarrow}$). 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 ($\mathrm{P}^2_{\rightarrow}$), ($\mathrm{P}^2_{\neg}$), 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 →-Intro and →-Elim take the forms

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

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 →, ¬, and ∀.

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 $\mathtt{n} + \mathtt{1} = D\mathtt{n}\mathtt{0}$. 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* $B =_{\text{df}} L_0$ 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 ($\mathrm{P}^2_\rightarrow$) 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 $A \rightarrow B$. 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 ($\mathrm{P}_\rightarrow$) 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 $\lambda\beta^\rightarrow$ (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 $\texttt{Proof}_{\mathsf{T}}$ (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$_{\mathsf{T}}$) $\texttt{Proof}_{\mathsf{T}}(\texttt{n}, \ulcorner \phi \urcorner) \rightarrow \phi$

[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. $R(A, x) \rightarrow A$ and $\mathtt{Proof}_{\mathsf{T}}(x, \ulcorner \phi \urcorner) \rightarrow \phi$ (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) $\exists x R(A, x) \rightarrow A$

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

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 $\exists y(\mathtt{Proof}_{\mathsf{T}}(y, \ulcorner \exists x \mathtt{Proof}_{\mathsf{T}}(x, \ulcorner \phi \urcorner) \rightarrow \phi \urcorner)$ 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 $\texttt{Proof}_{\mathsf{T}}(x, y)$ (i.e. if $\mathsf{T} \vdash \phi$, then $\mathsf{T} \vdash \exists x \texttt{Proof}_{\mathsf{T}}(x, \ulcorner \phi \urcorner)$) 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 $\mathcal{T}$, 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 □ 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 $\mathcal{T}^{\omega}$ [17, p. 118] is

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

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 $\mathcal{T}$, 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) $t : (A \rightarrow B) \rightarrow (s : A \rightarrow t \cdot s : B)$
(LP2) $t : A \rightarrow A$
(LP3) $t : A \rightarrow !t : t : A$

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 $A \rightarrow B$ and A by *modus ponens* is handled by LP1 as follows: if we assume (as induction hypotheses) that $u(\overrightarrow{x}) : A \rightarrow B$ 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. $\vdash_{\mathcal{T}} \top \equiv \top$), 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 $\pi \bot x \vdash_{\mathcal{T}} \bot \equiv \top$).

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 $(\mathrm{P}_\forall)$, 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 →, ¬ and ∀ 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 $\vdash_{\mathsf{QLP}} (r(y)\forall y) : (x : \bot \to \bot)$—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)