Notre Dame Journal of Formal Logic
Volume 55, Number 2, 2014

Montague's Paradox, Informal Provability, and Explicit Modal Logic

Walter Dean

Abstract The goal of this paper is to explore the significance of Montague's paradox—that is, any arithmetical theory $T \supseteq Q$ over a language containing a predicate $P(x)$ satisfying (T) $P(\ulcorner\varphi\urcorner) \rightarrow \varphi$ and (Nec) $T \vdash \varphi \ \therefore \ T \vdash P(\ulcorner\varphi\urcorner)$ is inconsistent—as a limitative result pertaining to the notions of formal, informal, and constructive provability, in their respective historical contexts. To this end, the paradox is reconstructed in a quantified extension $\mathcal{QLP}$ (the quantified logic of proofs) of Artemov's *logic of proofs* ($\mathcal{LP}$). $\mathcal{QLP}$ contains both *explicit modalities* $t : \varphi$ ("t is a proof of φ") and also *proof quantifiers* $(\exists x)x : \varphi$ ("there exists a proof of φ"). In this system, the basis for the rule Nec is decomposed into a number of distinct principles governing how various modes of reasoning about proofs and provability can be internalized within the system itself. A conceptually motivated resolution to the paradox is proposed in the form of an argument for rejecting the unrestricted rule Nec on the basis of its subsumption of an intuitively invalid principle pertaining to the interaction of proof quantifiers and the proof-theorem relation expressed by explicit modalities.

1 On the Origins of Montague's Paradox

In this paper I will take "Montague's paradox" to correspond to the following result.

Proposition 1.1 ([43, Theorem 3]) *Let T_1 be a theory in the language $\mathcal{L}_P = \{0, s, +, \times, P\}$ extending Q (Robinson arithmetic) such that $P(x)$ is a unary predicate satisfying the axiom scheme*

(T) $\quad P(\ulcorner\varphi\urcorner) \rightarrow \varphi$

and the rule

(Nec) $\quad$ If $T_1 \vdash \varphi$, then $T_1 \vdash P(\ulcorner\varphi\urcorner)$.

Received February 1, 2012; accepted September 29, 2012
2010 Mathematics Subject Classification: Primary 03F03; Secondary 03F45, 03F50
Keywords: Montague's paradox, Knower paradox, constructive proof, informal proof, provability logic, explicit modal logic, justification logic
 10.1215/00294527-2420636

Then T_1 is inconsistent.

Proof Since T_1 extends Q, by Gödel's diagonal lemma there exists a sentence δ such that

(1) $\quad T_1 \vdash \delta \leftrightarrow \neg P(\ulcorner \delta \urcorner)$.

We may now reason in T_1 as follows:

(i) $T_1 \vdash \delta \leftrightarrow \neg P(\ulcorner \delta \urcorner)$	(1)
(ii) $T_1 \vdash P(\ulcorner \delta \urcorner) \rightarrow \delta$	T
(iii) $T_1 \vdash \neg P(\ulcorner \delta \urcorner)$	i, ii
(iv) $T_1 \vdash \delta$	i, iii
(v) $T_1 \vdash P(\ulcorner \delta \urcorner)$	Nec, iv
(vi) $T_1 \vdash \bot$	iii, v

□

Proposition 1.1 corresponds to Theorem 3 of Richard Montague's 1963 paper [43] titled "Syntactical treatments of modality, with corollaries on reflexion principles and finite axiomatizability." As suggested by this title, Montague is concerned with the interpretation

(2) $\quad P(\ulcorner F \urcorner)$ if and only if F is logically necessary.

T and Nec are presumably both plausible principles about logical necessity, respectively expressing that if F is logically necessary, then it is true, and if F is provable from logically necessary principles, then the fact that F is logically necessary is provable from logically necessary principles. Montague hence presents Proposition 1.1 (as well as several other inconsistency results based on similar principles) as calling into doubt our ability to consistently regard necessity as a predicate of sentences (i.e., "syntactically"). On this basis he famously remarked: "[I]f necessity is to be treated syntactically, . . . then virtually all of modal logic. . . must be sacrificed" [43, p. 294].

Since the appearance of [43], the appropriateness of this reaction has repeatedly been challenged. One reason for this is that despite Montague's proposal that $P(x)$ should be interpreted according to (2), the principles T and Nec are also compatible with interpreting this predicate as expressing several other notions which have traditionally been treated as sentential or propositional operators—for example, truth, knowledge, and provability.[1] The question thus naturally arises as to which of our "naive" notions Proposition 1.1 should be taken to reflect upon most directly.

The main thesis of this paper will be that the significance of Montague's paradox is most readily appreciated if we understand $P(x)$ as expressing some form of mathematical provability. Such an interpretation provides the historical context for two anticipations of Proposition 1.1 (respectively by Myhill [44] and Kreisel [33]). But in addition to this, I will suggest that by interpreting $P(x)$ as expressing provability, we not only gain some insight into what distinguishes the significance of Proposition 1.1 from formally similar inconsistency results (such as Tarski's formalization of the liar paradox), but also open new avenues for resolving the underlying conceptual tension which the result might be taken to highlight on this interpretation.

1.1 Provability, necessitation, and internalization In attempting to interpret $P(x)$ in the manner just suggested, we must confront the fact that it is far from clear that there is a univocal sense of provability in mathematics. For instance, it is possible to identify at least three different senses which a statement of the form "φ is provable"

might be assigned in different settings: (1) φ is *formally provable* (i.e., derivable in some relevant axiomatic system such as PA or ZF); (2) φ is *informally provable* (i.e., demonstrable not in any particular formal system but by some recognizably correct mathematical argument); (3) φ is *constructively provable* (i.e., demonstrable by methods of proof recognized within intuitionistic or constructive mathematics).

These senses of provability are each associated with a different nexus of historical and technical developments. But within all of the resulting traditions, attempts have been made to investigate the relevant notion of provability axiomatically which are similar in spirit to Montague's exploration of logical necessity. For instance, *provability logic* (in the sense of Smoryński [49] or Boolos [7]) can be understood as an attempt to use modal logic to investigate the properties of formal provability by considering a mapping $(\cdot)^*$ between the languages of modal logic and formal arithmetic which interprets statements of the form $\Box F$ as $\mathrm{Prov}_{PA}(\ulcorner F^* \urcorner)$. Prior to this, however, Gödel [20] observed that we can reason schematically about provability by treating the occurrence of "provable" in "φ is provable" as a modal operator and adopting appropriate axioms. As I will explore further below, this proposal gave rise to independent traditions in the axiomatic investigation of informal and constructive provability (see, e.g., Myhill [44], Halldén [27], Reinhardt [47], and Leitgeb [39] on informal provability, and Gödel [23], Kreisel [33], [34], Sundholm [50], Beeson [4], and Artemov [3] on constructive provability).

I will suggest below that not only do variants of Montague's paradox arise for each of the interpretations just considered, but also that a common resolution is available in each case. As an initial step in this direction, it will be useful to distill an interpretation of the predicate $P(x)$ which is grounded as closely as possible in the principles Nec and T on which the derivation of the inconsistency embodied by Proposition 1.1 depends.

Nec is an example of what I will refer to as an *internalization principle*—that is, a rule or axiom by which part or all of the reasoning which can be conducted in an $\mathcal{L}_P$-theory can be derived under the scope of $P(x)$ (or in the case of a modal theory, under the scope of its operator $\Box$). Nec is perhaps the strongest and simplest such principle, as it allows for the internalization of arbitrary theorems of T, including those whose proofs rely on "nonlogical" axioms and rules (among which we might place both the arithmetical axioms of Q and consequences of the principles T or Nec themselves).

This principle will be recognized as a predicate analogue to the necessitation rule common to all normal modal logics $\mathcal{L}$—that is,

(Nec) If $\vdash_{\mathcal{L}} F$, then $\vdash_{\mathcal{L}} \Box F$.

In the context of first-order theories such as T_1 we might also consider restricting Nec to some subclass of principles (e.g., the "logical" axioms of the system) and then adopt principles which have the effect of ensuring that $P(x)$ is closed under derivable consequence. Several of the other results presented in [43] attest to the fact that other combinations of such principles also lead to inconsistency in conjunction with the "reflection" axiom T, of which the following is characteristic.

Proposition 1.2 *Let $T_2 \supseteq Q$ be an $\mathcal{L}_P$-theory satisfying* T *and the following three principles:*

(K^-) if $T_2 \vdash P(\ulcorner \varphi \to \psi \urcorner)$ and $T_2 \vdash P(\ulcorner \varphi \urcorner)$, then $T_2 \vdash P(\ulcorner \psi \urcorner)$;

(U) $P(\ulcorner P(\ulcorner\varphi\urcorner) \to \varphi\urcorner)$;

(Log) $T_2 \vdash P(\ulcorner\varphi\urcorner)$ if φ is an axiom of first-order logic with identity.[2]

Then T_2 is inconsistent.

From a formal perspective, results like Propositions 1.1 and 1.2 are often compared with Tarski's theorem (see [52]) on the undefinability of truth—that is, that any $\mathcal{L}_P$-theory extending Q and containing all instances of the so-called *T-schema*

(TS) $P(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$

is inconsistent. In particular, as long as appropriate internalization principles are also assumed, Propositions 1.1 and 1.2 show that over Q, only the left-to-right direction of the T-schema is needed in order for the resulting theory to be inconsistent.

Several other classical results can be understood as strengthening the impression that what is characteristic about Montague's paradox is its reliance on internalization principles. For instance, whereas the principle T asserts that any sentence satisfying $P(x)$ is *true*, the following principle merely asserts that the set of sentences with this property is *consistent*:

(D) $P(\ulcorner\neg\varphi\urcorner) \to \neg P(\ulcorner\varphi\urcorner)$.

But now note that the schema

(K) $P(\ulcorner\varphi \to \psi\urcorner) \to (P(\ulcorner\varphi\urcorner) \to P(\ulcorner\psi\urcorner))$,

(4) $P(\ulcorner\varphi\urcorner) \to P(\ulcorner P(\ulcorner\varphi\urcorner)\urcorner)$

can both naturally be understood as principles by which reasoning conducted within an $\mathcal{L}_P$-theory T can be reproduced internally under the scope of $P(x)$.[3] But even though D can be understood as a further weakening of TS, Friedman and Sheard [19] showed that together with Nec, these principles are again sufficient to lead to inconsistency—that is, we have the following.

Proposition 1.3 *Let $T_3 \supseteq Q$ be an $\mathcal{L}_P$-theory satisfying* D, K, 4, *and* Nec. *Then T_3 is inconsistent.*

Friedman and Sheard's paper is generally understood in the context of the development of axiomatic theories of truth (see, e.g., [41], [38], [26]). In light of Tarski's theorem, however, they also issue the following caveat about the interpretation of a predicate $T(x)$ which might be added to an arithmetical base theory in an effort to capture some portion of our intuitions about truth:

> One may reasonably take the attitude that what we are really exploring here is the axiomatic properties of concepts which are somewhere between 'provability' (which is well understood but somehow insufficient) and full 'truth' (which is mysterious and perhaps inherently unstable). Possible interpretations of $T(x)$ might be 'intuitively provable' or 'knowably true.' [19, p. 3]

This passage will set the tone for the rest of this paper in the sense that two of the more refined theses for which I will argue are as follows: (1) the conceptual significance of results like Montague's paradox which show the inconsistency of combining weakened forms of TS together with internalization principles like Nec may be most directly appreciated when we interpret $P(x)$ as expressing one of the aforementioned forms of provability; (2) this fact is highlighted by the sort of justification which is

available for the adoption of internalization principles. From this it follows that if we seek a conceptually motivated resolution to Montague's paradox, a reasonable strategy is to consider the sort of basis which exists for adopting Nec relative to various potential interpretations of $P(x)$.[4]

Such a justification seems near at hand if we understand $P(x)$ as expressing a form of provability which at least subsumes the resources of a system S which we have adopted in order to formalize reasoning about this notion itself. For in this case, we may reason informally as follows.

(3) (i) Suppose that $S \vdash \varphi$.
(ii) There thus exists some S-derivation $\mathfrak{D}$ of φ—that is, a finite sequence (or tree) of statements defined according to the relevant definition of "$\vdash$" with hypotheses in S and conclusion φ.
(iii) To say that such a $\mathfrak{D}$ *exists* is to say that φ is *provable* in S. Since we intended $P(x)$ to subsume the methods of proof available in S, we are thus justified in concluding that $P(\text{“}\varphi\text{”})$ (where “φ” denotes some method of naming φ using the resources available in S).[5]

The evaluation of such an argument must, of course, depend on the precise interpretation we wish to assign to $P(x)$. But suppose that we adopt what is arguably the most conservative interpretation available—that is, $P(\text{“}\varphi\text{”})$ if and only if φ is *logically demonstrable* (i.e., either self-evidently a logical truth or provable from such truths by logically valid means).[6] If φ is logically demonstrable it is presumably also *logically necessary*, and hence not only (ipso facto) *true*, but also *knowable* and *believable* (in virtue of possessing a proof from logically valid principles which justify it).[7] This suggests that as long as we are willing to concede that the axioms of S are themselves logically demonstrable and that its rules preserve this property, then Nec ought to be understood as a valid rule of inference on any of the interpretations of $P(x)$ relative to which Montague's paradox has traditionally been discussed.[8]

Another notable feature of this argument is that it exposes a sense in which the justification of Nec relies on adopting an *existential* interpretation of $P(x)$—that is, one on which $P(\ulcorner\varphi\urcorner)$ expresses that *there exists* a proof or other sort of demonstration which serves to justify φ. Although such quantification is not expressible in the object language of theories like T_1 which treat $P(x)$ as a primitive predicate, this observation is already helpful in seeing why results like Proposition 1.1 might be understood as demonstrating paradoxical features of the concepts of provability or knowledge in addition to truth. For note that relative to these interpretations, we do not expect the corresponding right-to-left direction of TS to hold—that is, if φ is true, it need not (at least ipso facto) be provable or knowable. However, if we make the stronger assumption that φ is provable in S, then it is presumably also knowable on the basis of possessing an S-derivation.[9]

If $P(x)$ is treated as a defined predicate, the heuristic argument (3) may, of course, be replaced by a mathematical one about the metatheory of S. In the paradigmatic case, for instance, we may assume that S extends Q and that $P(x)$ is taken to represent formal provability in S via an explicit arithmetization of syntax. If S is also recursively axiomatizable, it will be possible to define a Δ_1^0 proof predicate $\text{Proof}_S(x, y)$ whose structure mirrors the inductive definition of $\vdash_S$ in the familiar manner of Gödel [21]. This allows us to express the existence of an S-proof of φ

via a Σ^0_1-sentence of the form $\exists x \mathrm{Proof}_S(x, \ulcorner \varphi \urcorner) =_{df} \mathrm{Prov}_S(\ulcorner \varphi \urcorner)$ using a numerical quantifier. If we now assume that $S \vdash \varphi$, then $\mathrm{Prov}_S(\ulcorner \varphi \urcorner)$ will thus be a true statement about the natural numbers. Since theories extending Q are Σ^0_1-complete (i.e., they prove all true Σ^0_1-sentences), it hence follows that $S \vdash \mathrm{Prov}_S(\ulcorner \varphi \urcorner)$. It thus follows that if we think of $P(x)$ as expressing derivability in a particular recursively axiomatizable formal system, the validity of Nec may be understood as a matter of mathematical necessity.

1.2 Constructive and informal provability The foregoing observations about arithmetical provability predicates were codified by Hilbert and Bernays's isolation of the schema (see [30])

(4) If $S \vdash \varphi$, then $S \vdash \mathrm{Prov}_S(\ulcorner \varphi \urcorner)$

as one of the conditions which a definition of $\mathrm{Prov}_S(x)$ must satisfy in order to be able to carry out the proof of Gödel's second incompleteness theorem for S. These considerations aside, however, one of Montague's goals in [43] was to demonstrate the extent to which inconsistency results like Proposition 1.1 depend only on the propositional (or more accurately, sentential) properties which are assumed to hold of $P(x)$, and not the details of any particular combinatorial analysis of provability or arithmetization thereof.[10]

It will be recalled, however, that Gödel's goal in [20] was not to undertake a general study of the properties of particular arithmetical proof predicates, but rather to use modal logic to provide a formalization of Heyting's [28], [29] *proof interpretation* (or, as it has come to be known, the *Brouwer–Heyting–Kolmogorov* (BHK) *interpretation*) of the intuitionistic propositional connectives. This interpretation seeks to associate with each formula F of intuitionistic propositional calculus ($\mathcal{INT}$) a so-called *proof condition* which can be roughly understood as giving an account of its constructive meaning by specifying what sort of object ought to be counted as a proof of F.

As formulated by Troelstra and van Dalen [55], for instance, the BHK clauses for conjunction and implication are as follows.

(BHK$_\wedge$) A proof of $F \wedge G$ is given by presenting a proof of F and a proof of G.

(BHK$_\rightarrow$) A proof of $F \rightarrow G$ is a construction which permits us to transform any proof of F into a proof of G.

It is often stressed (e.g., [50], [56]) that clauses of the sort just exemplified should be understood as explicating, rather than formally analyzing, the meaning of the intuitionistic connectives. However, the BHK interpretation has been the inspiration for a great deal of technical work which attempts to provide a formal characterization of the notions "proof," "construction," and "transformation" as they appear in clauses like BHK$_\wedge$ and BHK$_\rightarrow$.

Gödel [20] provided one of the earliest such proposals by suggesting that the clauses of the Heyting interpretation can be formalized in modal logic by using an operator $\Box$ with the intended interpretation:

(5) $\Box F$ if and only if F is constructively provable.

In order to substantiate this reading, he specified an embedding $(\cdot)^g$ of intuitionistic propositional calculus ($\mathcal{INT}$) into the language of propositional modal logic for which the following may be shown.

Proposition 1.4 *For all formulas F, $\mathcal{INT} \vdash F$ if and only if $\mathcal{S}4 \vdash (F)^g$.*[11]

This result has been said (see, e.g., [51]) to fall short of providing a completely satisfactory account of the BHK interpretation, as it fails to render explicit the existential quantification over constructive proofs which is presumably implicit in (5).[12] However, Gödel [20] (see also [23]) also indicates that it is possible to interpret the operator $\Box$ according to the following:

(6) $\Box F$ if and only if F is informally provable.

It might initially be thought that the notion of informal provability is sufficiently amorphous to not be amenable to logical treatment. In the wake of the incompleteness theorems, however, a number of authors have proposed that there is a sense of provability which underlies our ability to see, for example, that the Gödel sentence or the consistency statement for PA are true despite the fact that they are not formally derivable in PA (assuming its consistency). It is this "absolute" notion of provability—that is, demonstrability not in a particular axiomatic system but by any correct mathematical means—which has provided the inspiration for most axiomatic treatments of informal provability.

Gödel [20] is also typically cited as the origin of the current consensus that $\mathcal{S}4$ is the correct logic of informal provability understood as a propositional operator. Such a case may be made by arguing that the axioms of $\mathcal{S}4$ reflect conceptual truths about the notion of informal provability—for example, the T axiom can be taken to reflect the fact that what is provable by correct mathematical means is true, the K axiom to reflect that the class of informally provable statements is closed under modus ponens—and that intuitive counterexamples can be found to the validity of the axioms of stronger modal systems (such as $\mathcal{S}4.2$ or $\mathcal{S}4.3$) when $\Box$ is interpreted relative to (6) (see, e.g., [27], [8], [39]).

But since $\mathcal{S}4$ also contains the rule Nec (which we have seen is a modal analogue of Nec), it is natural to ask after the status of results like Proposition 1.1 on this interpretation. At least on the face of things, however, propositional modal systems do not provide the combinatorial apparatus necessary for generating self-referential statements about provability on which results of this type depend. In fact, it is easy to see that $\mathcal{S}4$ is incompatible with the existence of a certain class of self-referential statements—a point to which I will return below.

1.3 Myhill, Kreisel, and the anticipation of Montague's paradox Attempts to interpret a predicate $P(x)$ as expressing informal or constructive provability provide the context of two instances in which Proposition 1.1 appears to have been discovered independently of Montague's paper. The first such anticipation was by Myhill [44], who argued that while absolute provability is a legitimate notion, its properties should be studied axiomatically rather than through an arithmetization of syntax. Although he also endorses $\mathcal{S}4$ as the correct logic of informal provability (treated as a propositional operator), Myhill showed that a predicate version of this system—that is, one which includes the principles K, T, 4, and an unrestricted form of Nec—is inconsistent via a derivation which is essentially identical to that of Proposition 1.1.

The other anticipation of Montague's theorem which I wish to discuss originates with Kreisel's [33], [34] attempt to provide a formalization of the BHK interpretation of intuitionistic predicate calculus ($\mathcal{IQC}$). This represents one of several attempts to extend Gödel's modal interpretation of $\mathcal{INT}$ in a manner which more directly

accounts for the apparent quantification over proofs which is implicit in clauses like $\mathrm{BHK}_{\rightarrow}$. The form of Kreisel's system—which has come to be known as the *theory of constructions* ($\mathcal{C}$)—is thus quite different from the arithmetical and modal systems we have been considering.[13]

Kreisel proposed to take as basic the relation which an individual constructive proof bears to a statement which it demonstrates. $\mathcal{C}$ hence contains a class of structured expressions $t, u, v, \ldots$ (which I will refer to as *proof terms*) intended to denote constructive proofs, and a binary function symbol Π which expresses the characteristic function of this relation. Relative to a slight simplification of its syntax, $\mathcal{C}$ can be understood to contain terms of the form $\Pi(t; F)$ such that the equation $\Pi(t; F) = 0$ is intended to express that t is a constructive proof of F.[14]

The version of $\mathcal{C}$ considered in [33] took the form of an untyped equational theory with terms of the form $\Pi(t; F)$ as well as terms denoting operations on proofs (e.g., pairing, projection, application, and function abstraction), and constants 0 and 1 to respectively denote the values true and false. Kreisel's central result about the theory of constructions was as follows.

Proposition 1.5 *If $\mathcal{IQC} \vdash F$, then there exists a proof term t such that $\mathcal{C} \vdash \Pi(t; F) = 0$.*

To the extent that derivability in $\mathcal{IQC}$ can be taken to characterize intuitionistic validity, this result can be seen as at least partially making good on the intuitionistic credo that truth is to be equated with constructive provability.[15]

Goodman [25], however, argued that if a system like $\mathcal{C}$ is to provide an adequate foundation for intuitionistic logic, then its formulation should be "type- and logic-free" (p. 101). To this end he proposed replacing $\mathcal{C}$ with a system formulated in pure combinatorial logic, which he showed could be used to interpret a type-free version of $\mathcal{C}$ in which arbitrary lambda-abstraction on proof terms is allowed. He also showed that a variant $\mathcal{C}^*$ of this theory which contains a version of what I will refer to as an *explicit reflection principle* (i.e., $\Pi(t; F) = 0 \rightarrow F$) is inconsistent.

This result—which has come to be known as the *Kreisel–Goodman paradox*[16]—has conventionally been attributed to various features of the BHK interpretation which are particular to the intuitionistic interpretation of the logical connectives—for example, the impredicativity of the proof condition for the conditional, or Kreisel's assumption that the relation $\Pi(t; F) = 0$ must be decidable (see [57], [4], and note 14). What is of more immediate significance, however, is that Goodman's derivation of the inconsistency is again essentially identical to the proof of Montague's paradox given above, with the exception of the following two differences.[17]

First, since $\mathcal{C}^*$ does not contain arithmetical terms or axioms, it is not immediately obvious how it may be used to demonstrate the existence of self-referential statements about provability. What Goodman shows, however, is that if the relation $\Pi(t; F)$ is taken as primitive, then it is possible to mirror the construction of the so-called *fixed-point* (or "paradoxical") *combinator* of untyped lambda calculus (see [10]) to construct a statement D which is provably equivalent to the statement that it is not provable.[18] Once D has been obtained, the reasoning of the first four steps of Proposition 1.1 can be mimicked in $\mathcal{C}^*$ to yield that D is provable in $\mathcal{C}^*$.

The second respect in which Goodman's derivation of an inconsistency in $\mathcal{C}^*$ differs from that of Proposition 1.1 is that $\mathcal{C}^*$ does not possess a single rule analogous

to Nec which allows it to internalize its own theorems in a single step. By individually internalizing the relevant applications of axioms and rules of $\mathcal{C}^*$, however, it is possible to construct a term t such that $\mathcal{C}^* \vdash \Pi(t; D) = 0$. This can then be shown to lead to a contradiction similar to the clash between lines iii and v in the derivation of Proposition 1.1.

Goodman took this inconsistency to show that it is not legitimate to consider the totality of constructive proofs as constituting a domain over which we may meaningfully quantify (a feature which is required in order for D to be well-formed). He therefore proposed to respond to the inconsistency in $\mathcal{C}^*$ by arguing that we should conceive of constructive proofs as being stratified into levels, such that a given formula may only quantify over proofs of a fixed level.[19]

1.4 Explicit modal logic Prior to Kreisel's development of the theory of constructions, Gödel [23] had described in an unpublished lecture a system which attempts to directly axiomatize the relation which constructive proofs bear to the statements which they demonstrate. Essentially the same system was independently proposed by Artemov [2], [3] in the form of the *Logic of Proofs* ($\mathcal{LP}$). Rather than treating the assertion that proof t demonstrates F as a relation, the systems of both Gödel and Artemov employ a form of labeled modal operator. In the syntax of $\mathcal{LP}$, such expressions are known as *explicit modalities* $t{:}F$ and are assigned the intended interpretation

(7) $t{:}F$ if and only if t denotes a proof of F.

As we will see below, the axioms and rules of $\mathcal{LP}$ provide "explicit" analogues to those of $\mathcal{S}4$ wherein instances of $\Box$ are replaced by various forms of structured proof terms.

A reasonable case can thus be made that whereas $\mathcal{S}4$ is the correct logic of informal provability, $\mathcal{LP}$ represents at least a sound means of reasoning about the proof-theorem relation expressed by (7). More generally, explicit modal logics like $\mathcal{LP}$ can be understood as occupying a sort of middle ground between modal and arithmetical systems for reasoning about provability. For instance, the use of explicit modalities provides a means of formulating various combinatorial properties about proofs which seem implicit in our acceptance of the $\mathcal{S}4$ axioms under the interpretations (5) or (6), but which are arguably independent of what sorts of objects we ultimately take informal or constructive proofs to be.[20]

One way in which this is manifest is in the fact that although $\mathcal{LP}$ also does not contain a single rule which plays the role of Nec, it satisfies the following property analogous to Proposition 1.5.

Proposition 1.6 (Constructive necessitation) *For all formulas F, if $\mathcal{LP} \vdash F$, then $\mathcal{LP} \vdash t : F$ for some proof term t.*

Proposition 1.6 also provides a means of formalizing steps i–ii in the argument (3) given for the rule Nec above—that is, if F is provable in $\mathcal{LP}$, then we can construct a proof term t which mirrors the derivation standing behind this fact such that $t{:}F$ is provable in $\mathcal{LP}$ itself.

Note, however, that in order to formalize the statement that F is *provable* appearing at step iii of this argument requires that we introduce quantifiers over proofs (or, as I will call them, *proof quantifiers*) into the object language of the system in

question. Such quantifiers are eschewed in $\mathcal{LP}$ in favor of variables $x, y, z, \ldots$ over proofs (or, as I will call them, *proof variables*). Fitting [17] proposed a means of introducing proof quantifiers into the language of explicit modal logic to yield a system known as the *Quantified Logic of Proofs* ($\mathcal{QLP}$). In the language of $\mathcal{QLP}$, it is possible to formulate statements of the form $(\exists x)x : F$ with the intended interpretation

(8) $(\exists x)x : F$ if and only if there exists a proof of F.

Since it may be shown that $\mathcal{QLP}$ satisfies a quantified analogue to Proposition 1.6 (i.e., if $\mathcal{QLP} \vdash F$, then $\mathcal{QLP} \vdash (\exists x)x : F$), this in turn suggests that it is also possible to mimic step iii of the argument (3) within this system itself.

At the same time, however, the introduction of proof quantifiers allows for the formulation of an object language statement expressing that a sentence D is equivalent to its own unprovability—that is, $D \leftrightarrow \neg(\exists x)x : D$. On this basis one might think that a version of the Kreisel–Goodman paradox would reemerge. But rather than entailing the existence of self-referential statements like D, it may be shown that no statement of this form is derivable in $\mathcal{QLP}$ in a manner reminiscent of $\mathcal{S}4$. As we will also see below, the proof of this fact can be taken to mirror the derivation of Montague's paradox, wherein an application of constructive necessitation is used to achieve the role played by Nec in the original derivation.

It is this final observation which provides the context for the most specific point I will attempt to demonstrate below—that is, that the assumption of Nec in the derivation of Montague's paradox disguises a number of distinct principles about provability which ought to be regarded as individual assumptions on which the paradox rests. Once this is acknowledged, a particular principle (which I will refer to as *justified universal generalization* (JUG)) pertaining to how reasoning by universal generalization about proofs should be internalized stands out as suspect.

In Section 3, I will argue that a reasonable case can be made that this principle should be rejected relative to *each* of the provability interpretations mentioned above—that is, formal, informal, and constructive—albeit for somewhat different reasons. Abandoning JUG thus provides a principled basis for rejecting an unrestricted form of Nec, and thereby also a conceptually motivated resolution to Montague's paradox when $P(x)$ is interpreted as expressing one of these forms of provability. In order to see why this is so, however, we must first see in detail how the paradox can be reconstructed in $\mathcal{QLP}$ and also how this system interacts with self-reference via an interpretation into formal arithmetic. This will be the topic of Section 2.

2 The Explicit Reconstruction of Montague's Paradox

The considerations adduced in the previous section suggest that a system of explicit modal logic such as $\mathcal{QLP}$ is an appropriate medium in which to reconstruct Montague's paradox if we wish to better understand the role of the rule Nec in its derivation. The version of this system which will be presented here is a variant of the formulation given by Fitting [17] first presented in Dean [11]. (See Artemov [2], [3] for additional discussion of explicit modal logic in general.)

2.1 On $\mathcal{QLP}_0$ and $\mathcal{QLP}$ The language of $\mathcal{QLP}$ is similar to that of propositional modal logic. However, rather than a single modal operator $\Box$, $\mathcal{QLP}$ possesses an

infinite family of *explicit modalities* of the form $t{:}F$. The terms t appearing in such statements may themselves be structured expressions as specified by the grammar

$$t := x_i \mid a_i(x) \mid !t \mid t_1 \cdot t_2 \mid t_1 + t_2 \mid \langle t(x)\forall x\rangle,$$

where $x_1, x_2, \ldots$ are known as *proof variables* (which I will often abbreviate $x, y, z, \ldots$) and $a_1(x), a_2(x), \ldots$ as *primitive proof terms*, and where $!$, $\cdot$, $+$, and $\langle\cdot\forall\cdot\rangle$ denote *proof operations* respectively called *proof checker* (unary), *application* (binary), *sum* (binary), and *uniform verifier* (binary). The class of formulas of $\mathcal{QLP}$ may now be defined as follows.

Definition 2.1 If $P_0, P_1, \ldots$ are propositional letters, then the class of formulas of $\mathcal{QLP}$ is specified by the grammar

$$F := \bot \mid P_i \mid F \wedge G \mid F \vee G \mid F \to G \mid \neg F \mid t : F \mid (\forall x)F(x) \mid (\exists x)F(x).$$

The free variables of proof terms and formulas are respectively defined in the same manner as those of terms and formulas in first-order logic, with the exception that in terms of the form $\langle t(x)\forall x\rangle$, x is considered bound. As usual, a sentence is taken to be a formula with no free variables.

A Hilbert-style proof system for $\mathcal{QLP}$ can now be specified.

Definition 2.2 The axioms of $\mathcal{LP}$ are as follows:

(LP1) All tautologies of classical propositional logic,
(LP2) $t : (F \to G) \to (s : F \to t \cdot s : G)$,
(LP3) $t : F \to F$,
(LP4) $t : F \to !t : t : F$,
(LP5) $t : F \to t + s : F$ and $s : F \to t + s : F$,
(QLP1) $(\forall x)F(x) \to F(t)$, for any proof term t that is free for x in $F(x)$,
(QLP2) $F(t) \to (\exists x)F(x)$, for any proof term t that is free for x in $F(x)$,
(QLP3) $(\forall x)(F \to G(x)) \to (F \to \forall x G(x))$, where $x \notin \mathrm{FV}(F)$,
(QLP4) $(\forall x)(F(x) \to G) \to ((\exists x)F(x) \to G)$, where $x \notin \mathrm{FV}(G)$.

Axioms LP1–LP5 are the original axioms of $\mathcal{LP}$ presented in [3] and will be respectively recognized as versions of the *S4* axioms K, T, and 4 wherein instances of the operator $\Box$ have been replaced with explicit modalities. Axiom LP3 is a version of what I referred to above as an explicit reflection principle—that is, a codification of the fact that if we accept that a particular proof t demonstrates F, then we ought also to accept that F is true. Axioms LP2, LP4, and LP5 can be taken to record functional dependencies between proofs which are arguably implicit in the interpretation of $\Box$ when read in accordance with Gödel's [20] use of *S4* to formalize the BHK interpretation.[21] Axioms QLP1–QLP4 codify the fact that once we have elected to regard proofs as objects over which we may quantify, the quantifier axioms of classical first-order logic should continue to hold.

In order to state the rules of $\mathcal{QLP}$, we must first provide several auxiliary definitions which generalize the notion of a *constant specification* for $\mathcal{LP}$. Such a specification is a mapping $\mathfrak{CS}$ which to each proof constant a assigns a set $\mathfrak{CS}(a)$ of instances of axioms LP1–LP4 such that for each formula $F \in \mathfrak{CS}(a)$, a is understood as an unstructured proof of F (essentially recording the fact that, as an axiom, F requires no further justification). However, in $\mathcal{QLP}$, axioms may contain free variables which may ultimately be bound by proof quantifiers. As such, the proof constant associated with an axiom ought to reflect which free variables it contains.

Following Fitting [17], I will adopt the expression *primitive term specification* to refer to this generalization of the original definition of a constant specification. More precisely, a primitive term specification is a mapping $\mathfrak{P}$ which assigns to each primitive proof term a a set of formulas $\mathfrak{P}(a)$ such that if $F \in \mathfrak{P}(a)$, then $\mathrm{FV}(a) = \mathrm{FV}(F)$. Such a specification is said to *meet the free variable condition* if whenever $F(x_1, \ldots, x_n) \in \mathfrak{P}(a(x_1, \ldots, x_n))$ and $y_1, \ldots, y_n$ are variables which do not occur in $F(x_1, \ldots, x_n)$, then $F(y_1, \ldots, y_n) \in \mathfrak{P}(a(y_1, \ldots, y_n))$. The mapping $\mathfrak{P}$ is said to be *axiomatically appropriate* if for all (and only) instances F of the axioms listed above there exists a primitive proof term a such that $F \in \mathfrak{P}(a)$. For the rest of this section, I will assume that $\mathfrak{P}$ is a fixed primitive term specification satisfying these requirements.

Definition 2.3 The rules of $\mathcal{QLP}_0$ consist of modus ponens together with the following.

(UPG) If $\Gamma \vdash F(x)$, then $\Gamma \vdash (\forall x)F(x)$ if $x \notin \mathrm{FV}(\Gamma)$.
(AxNec) If F is an axiom of $\mathcal{QLP}$ and $F \in \mathfrak{P}(a)$, then $\vdash a : F$.

The names UPG and AxNec are short for *universal proof generalization* and *axiom necessitation*. With these rules in place, we can now define $\mathcal{QLP}_0$ to be the system consisting of axioms LP1–LP5, QLP1–QLP4, and the rules UPG and AxNec. I will write $\Gamma(\mathfrak{P}) \vdash_{\mathcal{QLP}_0} F$ to denote that F is derivable in $\mathcal{QLP}_0$ from assumptions in Γ such that $F \in \mathfrak{P}(a)$ for all instances $a : F$ of AxNec employed in the corresponding derivation. (I will suppress mention of $\mathfrak{P}$ when it has been fixed as above.)

AxNec may be considered as a special case of Nec in the case where the statement internalized is an axiom of the system. We would also like internalization to extend to the rest of the theorems of $\mathcal{QLP}_0$ in a manner which generalizes Proposition 1.6. The general situation which is faced here may be illustrated by considering the following example:

(i)	$(x{:}F \wedge G) \rightarrow G$	LP1
(ii)	$a(x) : ((x{:}F \wedge G) \rightarrow G)$	AxNec, i
(iii)	$a(x) : ((x{:}F \wedge G) \rightarrow G) \rightarrow (y : (x{:}F \wedge G) \rightarrow a(x) \cdot y : G)$	LP2
(iv)	$y : (x{:}F \wedge G) \rightarrow a(x) \cdot y : G$	ii, iii
(v)	$(\forall y)(y : (x{:}F \wedge G) \rightarrow a(x) \cdot y : G)$	UPG

Since steps i–iv use only the axioms and rules of the base system $\mathcal{LP}$, we can now invoke Proposition 1.6 to obtain a term t such that $t : (y : (x{:}F \wedge G) \rightarrow a(x) \cdot y : G)$ is provable in $\mathcal{LP}$ (and hence in $\mathcal{QLP}_0$).[22] However, we are as yet unable to internalize the inference from iv to v which is mediated by UPG.

A variety of factors discussed in [17] and [11] suggest that the internalization of reasoning by universal generalization about proofs is most appropriately handled by adopting the rule which I referred to earlier as *justified universal generalization*:

(JUG) $\vec{s} : \vec{\Gamma} \vdash t(x) : F(x) \quad \therefore \quad \vec{s} : \vec{\Gamma} \vdash \langle t(x) \forall x \rangle : (\forall x)F(x)$, where $x \notin \mathrm{FV}(\vec{s} : \vec{\Gamma})$

where $\vec{s} : \vec{\Gamma}$ denotes a sequence of premises of the form $s_1 : G_1, \ldots, s_n : G_n$.[23]

The adoption of this rule may be motivated by a comparison with the traditional form of the universal generalization rule for first-order logic as well as a derived form of the necessitation rule for *S4* which allows for hypotheses (see [54])—that is,

(UG) $\Gamma \vdash F(x) \quad \therefore \quad \Gamma \vdash (\forall x)F(x)$ if $x \notin \mathrm{FV}(\Gamma)$,

(S4Nec) $\Box\Gamma \vdash F \;\;\therefore\;\; \Box\Gamma \vdash \Box F$ where all $G \in \Box\Gamma$ are of the form $G \equiv \Box H$.

The condition on Γ in UG can be viewed as formalizing the fact that we are justified in concluding that $(\forall x)F(x)$ from a proof of $F(x)$ as long as this proof has not relied on any premises about the properties of the object denoted by the variable x. Similarly, the condition on Γ in S4Nec can be viewed as formalizing the fact that we are justified in concluding that F is necessary from a proof of F from premises if these premises are themselves necessary. JUG combines these conditions by requiring that a proof justifying that a statement of the form $(\forall x)F(x)$ can be derived by UPG in $\mathcal{QLP}$ must reflect both that we are able to prove $F(x)$ *uniformly* (i.e., that there is a term $t(x)$ such that $t(s)$ functions as a proof of $F(s)$ for all proof terms s), and also that the proof of $F(x)$ follows from premises which do not involve x and which are also themselves explicitly provable.[24]

The system $\mathcal{QLP}$ is defined to be $\mathcal{QLP}_0$ with the addition of the rule JUG. We define $\Gamma \vdash_{\mathcal{QLP}} F$ similarly (I will suppress the subscript when clear from context). $\mathcal{QLP}$ may be shown to satisfy the deduction theorem—that is, if $\Gamma, F \vdash G$, then $\Gamma \vdash F \rightarrow G$. Using JUG we may now internalize the inference by UPG at step iv in the previous derivation.[25] More generally, it is now possible to demonstrate the following.

Theorem 2.4 (Internalization) *If $\vec{s} : \vec{\Gamma}, \vec{y} : \Delta \vdash F$, then there exists a proof term $t(\vec{s}, \vec{y})$ such that $\vec{s} : \vec{\Gamma}, \vec{y} : \vec{\Delta} \vdash_{\mathcal{QLP}} t(\vec{s}, \vec{y}) : F$.*

As an immediate corollary we have that if $\mathcal{QLP} \vdash F$, then by Theorem 2.4 and QLP2, $\mathcal{QLP} \vdash (\exists x)x : F$. As we will see below, a particular instance of this fact plays a significant role in reconstructing the derivation of Montague's paradox in $\mathcal{QLP}$. Before embarking on this, however, it will also be useful to record two additional facts about the connection between $\mathcal{LP}$, $\mathcal{QLP}$, and $\mathcal{S}4$ which were respectively obtained in [3] and [17].

Define a *realization* r of a propositional modal formula F to be an assignment of proof terms to all occurrences of $\Box$ in F. We write F^r for the image of F under the realization r. A realization is said to be *normal* if all occurrences of $\Box$ in F appearing in negative positions are replaced with proof variables.

Theorem 2.5 (Realization) *If $\mathcal{S}4 \vdash F$, then $\mathcal{QLP} \vdash F^r$ for some normal realization r.*

This result can be taken to further substantiate the view that the constructive provability interpretation of $\mathcal{S}4$ provides a means of interpreting the operator $\Box$ as expressing a form of implicit quantification over proofs. Note in particular, that it immediately follows from the realization theorem and Proposition 1.4 that

(9) $\mathcal{INT} \vdash F \iff \mathcal{LP} \vdash (F^g)^r$ for some normal realization r.

Since a normal realization replaces instances of $\Box$ in negative positions with proof variables, a realization of the image of a formula of $\mathcal{INT}$ under the Gödel embedding $(\cdot)^g$ can thus be seen as analyzing the proof conditions of a complex statement in a manner which makes explicit the functional dependencies between the proofs of its constituents in the manner envisioned by the BHK interpretation.

Finally, note that since realizations assign proof terms to instances of $\Box$, Theorem 2.5 can be understood as quantifying over proofs in the metatheory of $\mathcal{QLP}$—that is, the theorem promises that if F is provable in $\mathcal{S}4$, then there is *some* way of

replacing the occurrences of $\Box$ in F with proof terms so as to yield a theorem of $\mathcal{QLP}$. In order to see how this quantification can be brought inside the object language of $\mathcal{QLP}$, define the mapping $(\cdot)^{\exists}$ between the language of $\mathcal{S}4$ and $\mathcal{QLP}$ as follows: $(P_i)^{\exists} = P_i$ for P_i a propositional letter; $(\cdot)^{\exists}$ commutes with propositional connectives; $(\Box F)^{\exists} = (\exists x)x : F^{\exists}$. The following may then be shown.

Theorem 2.6 *We have that $\mathcal{S}4 \vdash F$ if and only if $\mathcal{QLP} \vdash F^{\exists}$.*

2.2 Reconstruction The goal of this section is to reconstruct in as precise a manner as possible the derivation of Montague's paradox in $\mathcal{QLP}$. However, the term "reconstruction" must be taken with at least one grain of salt. For as we have just seen, there is a close connection between derivability in $\mathcal{QLP}$ and $\mathcal{S}4$. And as I observed in Section 1, there is a sense in which $\mathcal{S}4$ is incompatible with the existence of statements which, like (1), can be understood as asserting that a certain sentence is equivalent to its own unprovability.

In order to make this observation precise, consider the following two sentences of propositional modal logic:

(10) (a) $D \leftrightarrow \neg\Box D$
(b) $\Box(D \leftrightarrow \neg\Box D)$

If we interpret $\Box$ as expressing provability, (10.a) can be taken to assert that D is true if and only if D is unprovable, whereas (10.b) asserts that this former fact is provable.

It may easily be shown that both (10.a) and its negation are consistent with $\mathcal{S}4$ (e.g., by constructing appropriate Kripke models). On the other hand, (10.b) is *refutable* in $\mathcal{S}4$ for all formulas D as is evident from the following derivation:[26]

(11)	(i)	$\Box(D \leftrightarrow \neg\Box D) \vdash D \leftrightarrow \neg\Box D$	T, modus ponens
	(ii)	$\Box(D \leftrightarrow \neg\Box D) \vdash \Box D \rightarrow D$	T
	(iii)	$\Box(D \leftrightarrow \neg\Box D) \vdash \neg\Box D$	i, ii
	(iv)	$\Box(D \leftrightarrow \neg\Box D) \vdash D$	i, iii
	(v)	$\Box(D \leftrightarrow \neg\Box D) \vdash \Box D$	S4Nec, iv
	(vi)	$\Box(D \leftrightarrow \neg\Box D) \vdash \bot$	iii, v
	(vii)	$\vdash \neg\Box(D \leftrightarrow \neg\Box D)$	vi, deduction theorem

Despite the fact that it is conducted in modal logic rather than arithmetic, the logical structure of this derivation is again similar to that of Proposition 1.1. One difference, however, is that in (11) the self-referential statement (10.b) is assumed as a hypothesis, rather than being derived on its own via the diagonal lemma.[27] This suggests that rather than standing mute on the existence of statements which assert their own unprovability, $\mathcal{S}4$ is in fact incompatible with the *provability* of such statements (for if $\mathcal{S}4 \vdash F \leftrightarrow \neg\Box F$ for any formula F, then $\mathcal{S}4 \vdash \Box(F \leftrightarrow \neg\Box F)$ via Nec and would hence be inconsistent in virtue of the foregoing derivation).[28]

From an instrumental perspective, however, this suggests that due to the close relationship between $\mathcal{QLP}$ and $\mathcal{S}4$, the appropriate way to reconstruct the reasoning of Montague's paradox is to start from the assumption of a sentence expressing the provability of a statement which expresses its own unprovability. For instance, in parallel to (10.a,b) we have

(12) (a) $D \leftrightarrow \neg(\exists x)x : D$
(b) $(\exists y)y : (D \leftrightarrow \neg(\exists x)x : D)$

As might be expected, (12.a) is consistent with $\mathcal{QLP}$, whereas (12.b) is refutable. The latter fact can be taken to follow directly from Theorem 2.6. However, in order to understand more precisely the principles on which this fact depends, it will be useful to reconstruct its derivation in full detail. To this end, it is useful to adopt not (12.b), but rather $y : (D \leftrightarrow \neg(\exists x)x : D)$ (where y is a free proof variable which will ultimately be bound by UPG) as a reductio assumption. We may now reason in $\mathcal{QLP}$ as follows:

(13)

(i)	$y : (D \leftrightarrow \neg(\exists x)x : D) \vdash D \leftrightarrow \neg(\exists x)x : D$	LP3, modus ponens
(ii)	$y : (D \leftrightarrow \neg(\exists x)x : D) \vdash (\exists x)x : D \rightarrow D$	derivable in $\mathcal{QLP}$
(iii)	$y : (D \leftrightarrow \neg(\exists x)x : D) \vdash \neg(\exists x)x : D$	i, ii
(iv)	$y : (D \leftrightarrow \neg(\exists x)x : D) \vdash D$	i, iii
(v)	$y : (D \leftrightarrow \neg(\exists x)x : D) \vdash t(y) : D$	for some $t(y)$ via Theorem 2.4
(v′)	$y : (D \leftrightarrow \neg(\exists x)x : D) \vdash (\exists x)x : D$	v, QLP2
(vi)	$y : (D \leftrightarrow \neg(\exists x)x : D) \vdash \bot$	iii, v′
(vii)	$\vdash \neg y : (D \leftrightarrow \neg(\exists x)x : D)$	deduction theorem
(viii)	$\vdash (\forall y)\neg y : [D \leftrightarrow \neg(\exists x)x : D]$	UPG
(ix)	$\vdash \neg(\exists y)y : [D \leftrightarrow \neg(\exists x)x : D]$	

With the exception of the quantifier manipulations in steps vii–ix, this derivation again shares the structure of the proof of Proposition 1.1. As presented, however, the argument is enthymemic as neither step ii nor step v correspond to $\mathcal{QLP}$ axioms, nor are they derivable from the preceding steps by $\mathcal{QLP}$ rules. The first of these gaps may be filled in as follows:

(14)

(i)	$\vdash x : F \rightarrow F$	LP3
(ii)	$\vdash (\forall x)(x : F \rightarrow F)$	UPG
(iii)	$\vdash (\forall x)(x : F \rightarrow F) \rightarrow ((\exists x)x : F \rightarrow F)$	QLP4
(iv)	$\vdash (\exists x)x : F \rightarrow F$	ii, iii

With respect to the second gap, recall that step v in Montague's paradox and derivation (11) respectively correspond to applications of Nec and S4Nec. Although $\mathcal{QLP}$ does not have a single rule which allows for internalization in this manner, it does satisfy Theorem 2.4. In virtue of this, we know that it is possible to construct a term $t(y)$ as displayed in line v of (13) by internalizing the reasoning of steps i–iv. Steps i, iii, and iv involve only propositional reasoning and are thus straightforward to internalize. In order to internalize step ii, however, we must show that the derivation of $(\exists x)x : F \rightarrow F$ involving the application of the UPG rule can also be internalized.[29]

This may be accomplished using the rule JUG in the following manner:

(15)

(i)	$\vdash x : F \rightarrow F$	LP3
(ii)	$\vdash r(x) : (x : F \rightarrow F)$	AxNec
(iii)	$\vdash \langle r(x)\forall x\rangle : (\forall x)(x : F \rightarrow F)$	JUG, ii
(iv)	$\vdash q : (\forall x)(x : F \rightarrow F) \rightarrow ((\exists x)x : F \rightarrow F)$	AxNec
(v)	$\vdash q \cdot \langle r(x)\forall x\rangle : ((\exists x)x : F \rightarrow F)$	LP2, iii, iv

It is now routine to establish that the term $t(y)$ in derivation (13) may be taken to be of the form $(a_1 \cdot y) \cdot ((b \cdot (q \cdot \langle r(x)\forall x\rangle)) \cdot (a_2 \cdot y))$, where a_1, a_2 and b are primitive proof terms for the tautologies which underlie the derivation of step iii from steps i and ii.[30]

The axioms and rules of $\mathcal{QLP}$ required for the subderivations (14) and (15) correspond to the additional principles required to justify the application of the rule Nec

in the reasoning of Montague's paradox which I alluded to at the end of Section 1.[31] The remainder of this paper will be aimed at assessing whether these principles are justified relative to our desire to interpret explicit modalities as expressing facts about the various forms of provability discussed above. Consider, for example, the following principles in the language of $\mathcal{QLP}$, instances of which are involved in the derivation (13):

(T_q) $\quad (\exists x)x : F \to F,$

(U_q) $\quad (\exists y)y : ((\exists x)x : F \to F).$

It will be observed that T_q and U_q are both related to the reflection principle T. For instance, if we interpret the quantifiers of $\mathcal{QLP}$ as ranging over informal proofs, T_q reports that if there exists a proof of F, then F must be true, while U_q expresses that there is an informal proof of this fact. The former statement can most readily be compared to the modal reflection axiom T (i.e., $\Box F \to F$), which in turn can be taken to express an apparent conceptual truth about both informal and constructive provability. This formula is still valid if $\Box$ is interpreted as expressing formal provability—that is, all instances of $\mathrm{Prov}_{PA}(\ulcorner F^* \urcorner) \to F^*$ are true in the standard model (where F^* is an arithmetical realization of propositional modal logic in the sense of [7]).

These two facts come apart in the case of the modal analogue of U_q—that is,

(U) $\quad \Box(\Box F \to F).$

For while U is derivable in $\mathcal{S}4$, it has false realizations when $\Box$ is interpreted as $\mathrm{Prov}_{PA}(x)$ (e.g., when F is mapped to $0 = 1$ as noted in Section 1). For this reason, it seems prudent to at least reserve judgment as to whether U should be accepted as an evident property of informal or constructive provability. One of the primary questions we will be concerned with below is whether U_q should ultimately be accepted on the informal and constructive proof interpretations of explicit modalities.

2.3 Arithmetical semantics for $\mathcal{QLP}$ There are two goals one might hope to achieve in providing a semantical interpretation of an explicit modal logic like $\mathcal{LP}$ or $\mathcal{QLP}$. First, one might hope to provide a mathematical interpretation of explicit modalities which provides an analysis of the nature of the objects which proof terms are intended to denote or of the justificatory relationship these objects bear to statements in the language of the systems. Second, one might hope to establish various metatheoretic results about one of these systems—for example, the consistency or inconsistency of certain sets of sentences, the conservativeness of a subsystem with respect to a certain class of formulas, and so on—without claiming that the semantics itself provided an adequate analysis of the relation expressed by $t{:}F$.

The semantics formulated in this section is primarily intended to serve the latter purpose. In particular, we will see that it is possible to provide an arithmetical interpretation of the language of $\mathcal{QLP}$ which is similar in spirit to the definition of an *arithmetical realization* in provability logic (see [7]) and even more similar to the definition of an *arithmetical interpretation* of the language of $\mathcal{LP}$ (see [3]).

Such interpretations can be understood as means of mapping sentences from the relevant modal or explicit modal language into arithmetical sentences such that statements involving the operators $\Box$ and : are mapped to statements about formal provability with respect to an arithmetical theory such as PA. In particular, an arithmetical realization $(\cdot)^*$ of propositional modal logic maps sentences of the form $\Box F$ to sentences of the form $\mathrm{Prov}_{PA}(\ulcorner F^* \urcorner)$ in the language $\mathcal{L}_a$ of first-order arithmetic. Similarly, an arithmetical interpretation $(\cdot)^+$ of the language of $\mathcal{LP}$ maps sentences of the form $t{:}F$ to those of the form $\mathrm{Proof}_{PA}(\ulcorner t^+ \urcorner, \ulcorner F^+ \urcorner)$. Artemov [3] established that relative to an appropriate formulation of the latter mapping, the images of the theorems of $\mathcal{LP}$ are not only arithmetically sound (i.e., true in the standard model), but they are also individually provable in PA itself.

Although we are about to see that the situation is more complex in the case of $\mathcal{QLP}$, it will be useful to base the definition of a $\mathcal{QLP}$-interpretation $(\cdot)^\circ$ as closely as possible on the definition of $(\cdot)^+$ given in [3]. To this end, let $\mathrm{Prf}(x, y)$ be a proof predicate for PA which satisfies the following conditions:

(16) (a) $\mathrm{Prf}(x, y)$ is provably Δ_1^0 in PA,
(b) for every arithmetical formula φ, $PA \vdash \varphi \iff$ for some $n \in \mathbb{N}$, $N \models \mathrm{Prf}(\overline{n}, \ulcorner \varphi \urcorner)$,
(c) if we let $T(n) = \{k \mid N \models \mathrm{Prf}(\overline{n}, \overline{k})\}$, then $\mathrm{Prf}(x, y)$ satisfies
(i) for all n, m there is a k such that $T(n) \cup T(m) \subseteq T(k)$,
(ii) $T(n)$ is finite for all n.[32]

It will also be useful to assume that the arithmetical language which we are working in contains terms $\boldsymbol{f_0}, \boldsymbol{f}_1, \ldots$ which represent all recursive functions. I will accordingly assume that PA is formulated so as to contain the defining axioms for all these functions (see [49] for details of how this can be handled). It follows from the fact that $\mathrm{Prf}(x, y)$ is provably Δ_1^0 that the set $\boldsymbol{P} = \{n \mid N \models \exists y \mathrm{Prf}(\overline{n}, y)\}$ of Gödel numbers of proofs is recursive. There is hence a recursive function $p : \mathbb{N} \to \mathbb{N}$ which enumerates P injectively. Let $\boldsymbol{p}$ be a name for an $\mathcal{L}_a$-representation of this function. For readability, we revert to the official names for proof variables $x_0, x_1, \ldots$ and assume that the official names for arithmetical variables in $\mathcal{L}_a$ are $y_0, y_1, \ldots$. We then set $(x_i)^\circ = \boldsymbol{p}(y_i)$. In the case where x_i occurs free in a $\mathcal{QLP}$ formula $\varphi(x_i)$, this will mean that the y_i will occur free in $(\varphi(x_i))^\circ$ inside an arithmetical term of the form $\boldsymbol{p}(y_i)$.

Suppose as before that $\mathfrak{P}$ is some fixed axiomatically appropriate primitive term specification. For each instance of a $\mathcal{QLP}$ axiom F, there is thus some primitive proof term a such that $F \in \mathfrak{P}(a)$ and $\mathrm{FV}(a) = \mathrm{FV}(F)$. We wish statements of the form $a(\vec{x}) : F(\vec{x})$ with $F(\vec{x}) \in \mathfrak{P}(a(\vec{x}))$ to be mapped to arithmetical open sentences which are true for all numerical substitution instances of their free variables. This means that the image of a primitive proof term must itself be a function mapping numbers into functions on Gödel numbers of proofs. For instance, if $F(\vec{x})$ is an axiom of $\mathcal{QLP}$ and we have defined $(\cdot)^\circ$ so that $F(\vec{x})^\circ$ is an open sentence of $\mathcal{L}_a$ provable in PA, then the interpretation of $a(\vec{x})$ should be a function $a_{F(\vec{x})}(\vec{x}^\circ)$ which for every assignment of $\mathcal{QLP}$-terms to $\vec{x}$, returns the code of an arithmetical proof of the image of the corresponding substitution instance of $F(\vec{x})$.

We must also provide a means of interpreting the proof operations $\cdot$, $+$, !, and $\langle \cdot \forall \cdot \rangle$. In the case of the first three operators, this can be accomplished in a manner similar to Artemov's original arithmetical interpretation of $\mathcal{LP}$ via the use of

three functions $m(y_1, y_2)$, $s(y_1, y_2)$, and $c(y)$. In particular, $m(y_1, y_2)$ returns the least Gödel number of a proof containing all instances of sentences G such that $\ulcorner F \to G \urcorner \in T(y_1)$ and $\ulcorner F \urcorner \in T(y_2)$, $s(y_1, y_2)$ returns the least Gödel number of a proof containing $T(y_1)$ and $T(y_2)$, and $c(y)$ returns the least Gödel number of a proof z such that $\mathrm{Prf}(z, \ulcorner F \urcorner) \in T(y)$ for all $\ulcorner F \urcorner \in T(y_1)$. Finally, in order to interpret the universal verifier symbol $\langle \cdot \forall \cdot \rangle$, we define a function $g(y_1, y_2)$ which, if y_1 is the Gödel number of a proof and y_2 is the Gödel number of a variable y, returns the least Gödel number of a proof z such that $\mathrm{Prf}(z, \ulcorner (\forall y) F \urcorner) \in T(z)$ for all $F \in T(y_1)$ such that $y \in \mathrm{FV}(F)$.

Following [3], it is routine to show that these functions are primitive recursive and may hence be represented by arithmetical terms $\boldsymbol{m}, \boldsymbol{s}, \boldsymbol{c}$, and $\boldsymbol{g}$. Finally, in the manner of [7], let $su(x, y, z)$ be the primitive recursive function such that for all i, j, k, $su(i, j, k)$ is the Gödel number of the result of substituting the numeral $\bar{i}$ for the jth free variable in formula with Gödel number k.

Definition 2.7 An arithmetical interpretation for $\mathcal{QLP}$ is a mapping $(\cdot)^\circ$ on both proof terms and formulas defined relative to the definitions of $\mathrm{Prf}(y_1, y_2)$, $m(y_1, y_2)$, $p(y_1, y_2)$, $c(y)$, and $g(y_1, y_2)$ just given. The value of t° is defined inductively as follows:

- $x_i^\circ = \boldsymbol{p}(y_i)$;
- if $F(x_1, \ldots, x_n) \in \mathfrak{P}(a(x_1, \ldots, x_n))$, then
$$\big(a(x_1, \ldots, x_n)\big)^\circ = \boldsymbol{a}_{F(x_1,\ldots,x_n)}(x_1^\circ, \ldots, x_n^\circ);$$
- $(t \cdot s)^\circ = \boldsymbol{m}(t^\circ, s^\circ)$, $(t + s)^\circ = \boldsymbol{s}(t^\circ, s^\circ)$, $(!t)^\circ = \boldsymbol{c}(t^\circ)$, $(\langle t \forall x_i \rangle)^\circ = \boldsymbol{g}(\ulcorner t^\circ \urcorner, \ulcorner y_i \urcorner)$.[33]

The value of F° is defined inductively as follows:

- $(P_i)^\circ$ is an arbitrary closed sentence of $\mathcal{L}_a$ and $\bot^\circ = \bot$ (where the latter is understood to denote some fixed arithmetical sentence which is refutable in PA—e.g., $0 = 1$);
- $(\cdot)^\circ$ commutes with propositional connectives—for example, $(F \to G)^\circ = F^\circ \to G^\circ$;
- $((\forall x_i) F)^\circ = (\forall y_i) F^\circ$, $((\exists x_i) F)^\circ = (\exists y_i) F^\circ$;
- $t : F(x_{k_1}, \ldots, x_{k_m}) = \mathrm{Prf}(t^\circ, \boldsymbol{su}(x_{k_m}^\circ, k_m, \ldots, \boldsymbol{su}(x_2^\circ, k_2, \boldsymbol{su}(x_1^\circ, k_1, \ulcorner F(x_{k_1}^\circ, \ldots, x_{k_m}^\circ) \urcorner)) \ldots))$.

It will be useful to observe several differences between the way proof variables are handled with respect to arithmetical interpretations for $\mathcal{LP}$ and for $\mathcal{QLP}$. Note that both $\mathcal{LP}$ and $\mathcal{QLP}$ possess axioms containing free proof variables—for example, $x : P \to P$. Recall, however, that an arithmetical interpretation $(\cdot)^+$ for $\mathcal{LP}$ associates each proof variable x with a fixed natural number, meaning that the image of such an axiom under $(\cdot)^*$ will always be a closed formula in the language of arithmetic. On the other hand, under an arithmetical interpretation $(\cdot)^\circ$ for $\mathcal{QLP}$, this sentence will get mapped to an open formula of the form

$$\mathrm{Prf}(\boldsymbol{p}(y), \ulcorner F^\circ \urcorner) \to F^\circ. \tag{17}$$

It will be useful to define arithmetical soundness for the language of $\mathcal{QLP}$ so as to accommodate this. The most straightforward way of doing so is to treat free arithmetical variables as universally bound. In particular, I will say that a $\mathcal{QLP}$ sentence $F(x_{k_1}, \ldots, x_{k_m})$ with free variables displayed is *arithmetically sound with respect*

to $(\cdot)^\circ$ if $N \models \forall y_1, \ldots, \forall y_n (F)^\circ(y_{k_1}, \ldots, y_{k_m})$. If F° is arithmetically sound with respect to all interpretations $(\cdot)^\circ$ satisfying Definition 2.7, I will simply say that F° is *arithmetically sound*.

On the route to demonstrating the arithmetical soundness of $\mathcal{QLP}_0$, we must first convince ourselves that if $\mathfrak{P}$ is an axiomatically appropriate primitive term specification, then there exist $\mathfrak{P}$-arithmetically sound interpretations. This essentially involves showing that if the formula F with free variables $x_{k_1}, \ldots, x_{k_m}$ is an axiom of $\mathcal{QLP}_0$, and $F \in \mathfrak{P}(a)$, then there exists an m-ary primitive recursive function a_F such that for all $n_1, \ldots, n_m \in \mathbb{N}$, $a_F(n_1, \ldots, n_m)$ returns a proof of the interpretation of F when its free variables are mapped to $p(n_1), \ldots, p(n_m)$. If we can show that such functions exist for all primitive proof terms a and formulas F such that $F \in \mathfrak{P}(a)$, then we will have obtained an arithmetically sound interpretation. If $\mathfrak{P}$ is a primitive term specification, we similarly say that an interpretation $(\cdot)^\circ$ is $\mathfrak{P}$-*arithmetically sound* if $a{:}F$ is arithmetically sound for all a, F such that $F \in \mathfrak{P}(a)$.

In order to make the case that arithmetically sound interpretations exist, I will consider an example. Consider the axiom LP3 and recall that its arithmetical interpretation is given by (17). Although this is an open arithmetical sentence, it is easy to see that, for all $n \in \mathbb{N}$,

$$PA \vdash \mathrm{Prf}(\boldsymbol{p}(\overline{n}), \ulcorner F^\circ \urcorner) \rightarrow F^\circ. \tag{18}$$

For note that we have either (i) $N \models \mathrm{Prf}(\boldsymbol{p}(\overline{n}), \ulcorner F^\circ \urcorner)$ or (ii) $N \not\models \mathrm{Prf}(\boldsymbol{p}(\overline{n}), \ulcorner F^\circ \urcorner)$. In the first case, $\mathrm{Prf}(\boldsymbol{p}(\overline{n}), \ulcorner F^\circ \urcorner)$ is a true Δ^0_1-sentence and hence $p(n)$ is indeed the Gödel number of a proof of F° in PA. Hence by the Σ^0_1-completeness of PA, $PA \vdash F^\circ$, meaning that (18) holds. In the second case, $\mathrm{Prf}(\boldsymbol{p}(\overline{n}), \ulcorner F^\circ \urcorner)$ is a false Δ^0_1-sentence, meaning that $\neg\mathrm{Prf}(\boldsymbol{p}(\overline{n}), \ulcorner F^\circ \urcorner)$ is provably equivalent to a true Σ^0_1-sentence. Hence $PA \vdash \neg\mathrm{Prf}(\boldsymbol{p}(\overline{n}), \ulcorner F^\circ \urcorner)$ by Σ^0_1-completeness and (18) again holds.

The foregoing reasoning is uniform in n in the sense that each substitution instance of (17) is true. But also note that given $n \in \mathbb{N}$, we can calculate the Gödel number of a proof of $\mathrm{Prf}(\boldsymbol{p}(\overline{n}), \ulcorner F^\circ \urcorner) \rightarrow F^\circ$ via the following procedure: (1) given n, compute $p(n)$ and check whether $\mathrm{Prf}(\boldsymbol{p}(\overline{n}), \ulcorner F^\circ \urcorner)$ is true (this can be decided because $\mathrm{Prf}(\boldsymbol{p}(\overline{n}), \ulcorner F^\circ \urcorner)$ is Δ^0_1); (2) if so, $p(n)$ really is the Gödel number of a proof of F° in PA and we can hence construct the Gödel number of a proof of this instance of (17) by composing the proof coded by $p(n)$ with a proof of the appropriate instance of the tautology $F \rightarrow (G \rightarrow F)$; (3) if not, then since $p(n)$ does not denote a proof of F°, we can (i) obtain the Gödel number q of a proof of $\neg\mathrm{Prf}(\boldsymbol{p}(\overline{n}), \ulcorner F^\circ \urcorner)$ by simply examining the structure of the proof denoted by $p(n)$ and noting that it does not have $\ulcorner F^\circ \urcorner$ as a conclusion, and we can then (ii) construct the Gödel number of a proof of this instance of (17) by composing q with a proof of the appropriate instance of the tautology $\neg F \rightarrow (F \rightarrow G)$. Since the procedure just described is clearly effective and involves only bounded search, we know that there is a primitive recursive function which for every n, returns the Gödel number of a proof demonstrating (18) in PA. It is this function which we take as $\boldsymbol{a}_{x:F \rightarrow F}(y)$.

It is straightforward but tedious to show that there exist primitive recursive functions which perform the same function for the other axioms of $\mathcal{QLP}_0$, and that the rules of this system preserve arithmetical soundness. Once this is accomplished we have the following.

Theorem 2.8 (Arithmetical soundness) *Suppose that $\mathfrak{P}$ is an axiomatically appropriate primitive term specification and that $\mathcal{QLP}_0(\mathfrak{P}) \vdash F$. Then $N \models F^\circ$ for all $\mathfrak{P}$-sound interpretations $(\cdot)^\circ$.*

Proof By induction on the derivation of F. □

This result is significant partly because it demonstrates the consistency of $\mathcal{QLP}_0$.[34] But recall from [3] that if $(\cdot)^+$ is an $\mathcal{LP}$-interpretation, then not only are the images of all theorems of $\mathcal{LP}$ under $(\cdot)^+$ true in N, but they are also provable in PA. As we will now see, however, there is no hope of extending Theorem 2.8 to an analogous result for $\mathcal{QLP}_0$ where "true in N" is replaced with "provable in PA" in the definition of arithmetical soundness.

In order to see this, first consider the case of the LP reflection axiom with $F \equiv \bot$—that is,

(19) $\quad x : \bot \rightarrow \bot \quad (\equiv \neg x : \bot).$

Via UPG in $\mathcal{QLP}_0$ we then have that

(20) $\quad \mathcal{QLP}_0 \vdash (\forall x)\neg x : \bot.$

The image of this statement under $(\cdot)^\circ$ is

(21) $\quad (\forall y)\neg\mathrm{Prf}\big(\boldsymbol{p}(y), \ulcorner\bot\urcorner\big),$

which can be seen to express "for all y, $p(y)$ is not a proof of a contradiction in PA." Assuming that PA is consistent, this statement is true in the standard model. However, since p enumerates all arithmetical proofs, (21) expresses the consistency of PA and is thus not provable in PA by the second incompleteness theorem (assuming that PA is itself consistent). It thus follows that although the images of all instances of LP3 are true in the standard model for all choices of $(\cdot)^\circ$, the image of (19) will not be provable in PA for any $\mathcal{QLP}$ arithmetical interpretation $(\cdot)^\circ$.

In fact, it is easy to see that in $\mathcal{QLP}$ it is possible to derive statements whose images are false under all arithmetically sound interpretations. To be more precise, it will be useful to record the following form of a familiar result about PA.

Proposition 2.9 (Internalized Löb's theorem) *For all $\mathcal{L}_a$-sentences F,*

$$PA \vdash (\exists y)\mathrm{Prf}\big(\boldsymbol{p}(y), \ulcorner(\exists y)\mathrm{Prf}\big(\boldsymbol{p}(y), \ulcorner F\urcorner\big) \rightarrow F\urcorner\big) \rightarrow (\exists y)\mathrm{Prf}\big(\boldsymbol{p}(y), \ulcorner F\urcorner\big).$$

Proof A straightforward adaption of the traditional proof given in, for example, [49]. □

Proposition 2.10 *Let $\mathfrak{P}$ be an axiomatically appropriate and arithmetically sound primitive term specification. Then assuming that PA is sound, there are sentences F such that $\mathcal{QLP}(\mathfrak{P}) \vdash F$, but $N \not\models F^\circ$.*

Proof Let $\mathfrak{P}$ be as in the hypotheses. Then by LP3 and AxNec,

(22) $\quad \mathcal{QLP}(\mathfrak{P}) \vdash r(x_0) : (x_0 : \bot \rightarrow \bot).$

It then follows by JUG that

(23) $\quad \mathcal{QLP}(\mathfrak{P}) \vdash \langle r(x_0)\forall x_0\rangle : (\forall x_0)(x_0 : \bot \rightarrow \bot),$

and then by QLP2 that

(24) $\quad \mathcal{QLP}(\mathfrak{P}) \vdash (\exists x_1)x_1 : (\forall x_0)(x_0 : \bot \rightarrow \bot).$

Note that the image of (24) under $(\cdot)^\circ$ is equivalent to

(25) $(\exists y)\mathrm{Prf}(\boldsymbol{p}(y), \ulcorner(\exists y)\mathrm{Prf}(\boldsymbol{p}(y), \ulcorner\bot\urcorner) \to \bot\urcorner)$.

Assuming that PA is consistent, it follows that (25) is not true in N. For if it were, it would follow from Proposition 2.9 and the soundness of PA that $N \models (\exists y)\mathrm{Prf}(\boldsymbol{p}(y), \ulcorner\bot\urcorner)$, which is true only if PA is inconsistent. □

Proposition 2.10 shows that it is not possible to extend the arithmetical soundness theorem from $\mathcal{QLP}_0$ to $\mathcal{QLP}$. One obvious reason for this is that JUG allows for the internalization of statements expressing consistency like (20) which are provable in $\mathcal{QLP}_0$. In fact, it is routine to check that the image of (23) under $(\cdot)^\circ$ is already false in N by employing the definition of $(\langle r(x)\forall x\rangle)^\circ$ given above (in particular, as long as PA is consistent, the number denoted by $g(\ulcorner r(x_0)^\circ\urcorner, \ulcorner x_0\urcorner)$ cannot denote a proof in the set $\boldsymbol{P}$). It thus follows that JUG is not an arithmetically sound rule of inference relative to the definition of $g(y_1, y_2)$ given above. However, this fact persists even when we limit our attention to those theorems of $\mathcal{QLP}$ that do not contain the universal verifier symbol.

Proposition 2.11 *Let $\mathfrak{P}$ be axiomatically appropriate, and let*

$$\mathcal{QLP}^- = \{F \mid \mathcal{QLP}(\mathfrak{P}) \vdash F \text{ and } F \text{ does not contain the symbol } \langle\cdot\forall\cdot\rangle\}.$$

Suppose also that $(\cdot)^\circ$ is a $\mathfrak{P}$-arithmetically sound interpretation, and let $(\mathcal{QLP}^-)^\circ$ denote the image of all sentences of $\mathcal{QLP}^-$ under $(\cdot)^\circ$. Then $PA \cup (\mathcal{QLP}^-)^\circ$ is inconsistent.

Proof Note that by (24) and (25) we have that

$$(\exists y)\mathrm{Prf}\big(\boldsymbol{p}(y), \ulcorner(\exists y)\mathrm{Prf}\big(\boldsymbol{p}(y), \ulcorner\bot\urcorner\big) \to \bot\urcorner\big) \in (\mathcal{QLP}^-)^\circ.$$

It follows again by Proposition 2.9 that $PA \cup (\mathcal{QLP}^-)^\circ \vdash (\exists y)\mathrm{Prf}(\boldsymbol{p}(y), \ulcorner\bot\urcorner)$. But, as we have seen, $\mathcal{QLP} \vdash (\exists x)x : \bot \to \bot$ and thus the arithmetic interpretation of this formula—that is, $(\exists y)\mathrm{Prf}(\boldsymbol{p}(y), \ulcorner\bot\urcorner) \to \bot$—is in $(\mathcal{QLP}^-)^\circ$. Putting these two facts together, we have $PA \cup (\mathcal{QLP}^-)^\circ \vdash \bot$. □

A similar argument can now be mounted to show that $\mathcal{QLP}$ is not conservative over $\mathcal{QLP}_0$ for sentences lacking the universal verifier symbol. In fact, we may show that no instance of U_q is derivable in $\mathcal{QLP}_0$ for atomic F.

Proposition 2.12 *For all propositional letters P and axiomatically appropriate primitive term specifications $\mathfrak{P}$, $\mathcal{QLP}_0(\mathfrak{P}) \nvdash (\exists x_0)x_0 : ((\exists x_1)x_1 : P \to P)$.*

Proof Suppose for a contradiction that $\mathcal{QLP}_0(\mathfrak{P}) \vdash (\exists x_0)x_0 : ((\exists x_1)x_1 : P \to P)$. By Theorem 2.8, it follows that

(26) $N \models ((\exists x_0)x_0 : ((\exists x_1)x_1 : P \to P))^\circ$

for all $\mathfrak{P}$-arithmetically sound interpretations $(\cdot)^\circ$. By the diagonal lemma for PA, there exists a closed arithmetical sentence δ such that

$$PA \vdash \delta \leftrightarrow \neg(\exists y_2)\mathrm{Prf}\big(\boldsymbol{p}(y_2), \ulcorner\delta\urcorner\big).$$

Hence by Σ^0_1-completeness,

$$PA \vdash (\exists y_1)\mathrm{Prf}\big(\boldsymbol{p}(y_1), \ulcorner\delta \leftrightarrow \neg(\exists y_2)\mathrm{Prf}\big(\boldsymbol{p}(y_2), \ulcorner\delta\urcorner\big)\urcorner\big).$$

If we let $(\cdot)^\circ$ be such that $P^\circ = \delta$, we may now reach a contradiction by reasoning in the standard model in the same manner as the derivation (13). In particular, since all of the axioms and rules of $\mathcal{QLP}_0$ are arithmetically sound, we reach the conclusion that $N \models \neg\exists y_2 \mathrm{Prf}(\boldsymbol{p}(y_2), \ulcorner\delta\urcorner)$ as at line (13.iii). However, as the foregoing reasoning can be formalized in PA, if we also assume via (26) the existence of a natural number $n = (x_0)^\circ$ such that $N \models \mathrm{Prf}(\boldsymbol{n}, \ulcorner\exists y_2 \mathrm{Prf}(y_2, \ulcorner\delta\urcorner) \rightarrow \delta\urcorner)$, we may then also conclude that $N \models \exists y_2 \mathrm{Prf}(\boldsymbol{p}(y_2), \ulcorner\delta\urcorner)$ in parallel to line (13.v′). □

By a similar method, it is also possible to show that $\mathcal{QLP}_0$ is consistent with the existence of self-referential statements such as (12.a,b).

Proposition 2.13 *Let $\mathfrak{P}$ be an axiomatically appropriate primitive term specification, and assume that PA is sound. Then*

$$\mathcal{QLP}_0(\mathfrak{P}) \cup \big\{(\exists x_0)x_0 : \big(P \leftrightarrow \neg(\exists x_1)x_1 : P\big) \mid P \textit{ a propositional letter}\big\}$$

is consistent.

Proof Suppose for a contradiction that

$$\mathcal{QLP}_0(\mathfrak{P}) \vdash \neg(\exists x_0)x_0 : \big(P \leftrightarrow \neg(\exists x_1)x_1 : P\big).$$

It follows by Theorem 2.8 that $N \models (\neg(\exists x_0)x_0 : (P \leftrightarrow \neg(\exists x_1)x_1 : P))^\circ$. From this it follows that

(27) $$N \models \neg(\exists y_0)\mathrm{Prf}(\boldsymbol{p}(y_0), \ulcorner P^\circ \leftrightarrow \neg(\exists y_1)\mathrm{Prf}(\boldsymbol{p}(y_1), \ulcorner P^{\circ\urcorner})\urcorner).$$

But again by the diagonal lemma and Σ^0_1-completeness,

$$PA \vdash (\exists y_0)\mathrm{Prf}\big(\boldsymbol{p}(y_0), \ulcorner\delta \leftrightarrow \neg(\exists y_1)\mathrm{Prf}\big(\boldsymbol{p}(y_1), \ulcorner\delta\urcorner\big)\urcorner\big)$$

for some δ. But now letting $P^\circ = \delta$, we have

$$N \models (\exists y_0)\mathrm{Prf}\big(\boldsymbol{p}(y_0), \ulcorner P^\circ \leftrightarrow \neg(\exists y_1)\mathrm{Prf}\big(\boldsymbol{p}(y_1), \ulcorner P^{\circ\urcorner}\big)\urcorner\big)$$

in contradiction to (27). □

3 Discussion

The case I presented in Section 1 can be summarized as follows: (1) the distinctive feature of Montague's paradox is its reliance on the internalization principle Nec; (2) this rule can be most readily justified if we understand $P(x)$ as expressing some form of provability; (3) but since to say that φ is *provable* is to say that *there exists* a proof of φ, a yet better interpretation of $P(\ulcorner\varphi\urcorner)$ is as expressing implicit quantification over proofs; (4) this feature of Nec may be made explicit by formulating the derivation of the paradox in a system like $\mathcal{QLP}$ which treats proofs as objects; and (5) in such a setting, the validity of Nec is decomposed into a number of axioms and rules which are individually required to sustain internalization in the sense of Theorem 2.4.

Having redeveloped the proof of Proposition 1.1 in $\mathcal{QLP}$, we are now in a position to examine these principles in more detail. For note that the only place that Nec is applied in this result is to derive $P(\ulcorner\delta\urcorner)$ at step v. If we interpret $P(\ulcorner\varphi\urcorner)$ as expressing that there exists a proof of φ, I argued above that the justification for this conclusion rests on the fact that we ought to be able to internalize the reasoning embodied by steps i–iv leading to the conclusion δ under the scope of $P(x)$. In the reconstructed derivation, however, we realize that in order to do this, we must not

only prove an instance of T_q, but also that this proof must itself be internalized to yield an instance of U_q. The resulting sub-derivations (14) and (15) can be seen to rely on the axioms LP2, LP3, QLP4 and the rules AxNec, UPG, and JUG of $\mathcal{QLP}$.

The central claim for which I will argue in this section is that on *each* of the provability interpretations considered in Section 1—that is, formal, informal, and constructive—plausible arguments can be adduced in favor of the correctness of all of these principles *with the exception of JUG*. However, the situation which we face in assessing such claims is complicated by the fact that it is not entirely clear what status should be assigned to the existence of self-referential statements about provability on the constructive and informal interpretations, as illustrated, for example, by the refutability of (12.b) in $\mathcal{QLP}$. The more refined thesis for which I will argue is thus as follows: to the extent to which Montague's paradox can be seen as threatening the consistency of our intuitions about the notion of provability (conceived either formally, informally, or constructively), consistency can be maintained even in the face of self-reference if we are willing to reject principles like JUG which allow for the internalization of certain forms of reasoning by universal generalization over proofs.

These claims may be evaluated with relative ease in the case of formal provability. For as long as we are willing to accept that the arithmetical proof predicate $\mathrm{Prf}(x, y)$ employed in Section 2.3 adequately expresses the relation which a formal proof bears to the sentences which it demonstrates, then the arithmetical soundness of LP2, LP3, QLP4, AxNec, and UPG suggests that these principles ought to be accepted as uncontroversial features of the relation borne by such proofs to their conclusions. This should not be surprising in the case of LP2 and AxNec, which can be respectively understood as explicit versions of special cases of the first and second Hilbert–Bernays derivability conditions. Additionally, the soundness of QLP4 can be taken to follow from the fact that the axioms of first-order logic are valid for arbitrary domains, including the case where we regard their quantifiers as ranging over formal proofs. And finally, although the images under $(\cdot)^\circ$ of all instances $t : F \to F$ of LP3 are not always derivable in PA (e.g., when $t \equiv x$ and $F \equiv \bot$), they will all be true in the standard model.

The status of JUG with respect to the formal provability interpretation is illuminated by the results of Section 2.3. In particular, we observed there that specific instances of this rule are not arithmetically sound. For instance, if we take the formula $F(x)$ to be the instance $x : \bot \to \bot$ of LP3 and the proof term $t(x)$ to be $r(x)$ (where we assume that $(x : \bot \to \bot) \in \mathfrak{P}(r(x))$), then we have $\mathcal{QLP} \vdash r(x) : (x : \bot \to \bot)$ and also $N \models (r(x) : (x : \bot \to \bot))^\circ$ for all interpretations $(\cdot)^\circ$. But although it then follows via JUG that $\mathcal{QLP} \vdash \langle r(x) \forall x \rangle : (\forall x)(x : \bot \to \bot)$, for no such interpretations can it be the case that $N \models (\langle r(x) \forall x \rangle : (\forall x)(x : \bot \to \bot))^\circ$. As noted above, this fact depends in its details on the precise means by which we go about interpreting the uniform verifier symbol $\langle \cdot \forall \cdot \rangle$ into the language of arithmetic. But as was observed in Proposition 2.10, we also have that $\mathcal{QLP} \vdash (\exists x_1) x_1 : (\forall x_0)(x_0 : \bot \to \bot)$ via existential generalization. The image of this statement under $(\cdot)^\circ$ is equivalent to $(\exists y_1)\mathrm{Prf}(\boldsymbol{p}(y_1), \ulcorner \forall y_0 \neg \mathrm{Prf}(\boldsymbol{p}(y_0), \ulcorner \bot \urcorner) \urcorner)$. As this statement formalizes the fact that PA proves its own consistency, it must hence be false in the standard model in virtue of the second incompleteness theorem (presuming that PA is consistent).

On the formal proof interpretation, JUG can thus be understood to be objectionable in virtue of implying the existence of proofs whose *nonexistence* we take to be a matter of mathematical necessity.[35] One might, however, think that we possess some form of justification for believing that PA is consistent—for example, our belief that its axioms are true and that its rules preserve truth. One might even take such observations to constitute an *informal* proof of the consistency of PA. As such, it might be thought that the arithmetical invalidity of JUG is a consequence of the manner in which the formal proof interpretation requires us to identify proofs with natural numbers. The question thus arises whether JUG ought to be retained as a valid principle about informal or constructive proof.

I will presently argue that JUG also ought to be rejected on both of these interpretations, albeit for reasons quite different than the ones just considered. But before engaging with this issue directly, something more substantive must finally be said about the nature of constructive and informal proof. A thorough discussion of either these notions is beyond the scope of the current paper. The following remarks are thus not intended to resemble a satisfactory analysis of either notion, but merely to highlight what is at stake when we assert or deny that principles expressible in the language of explicit modal logic reflect properties of such proofs.

In elementary expositions of intuitionism (e.g., [13], [53]), the notion of constructive proof is often introduced in contradistinction to that of classical proof. This suggests that one route to identifying what is meant by such a proof is to start from our conventional idea of a "classical" proof and then exclude proofs employing various modes of reasoning which are rejected within intuitionism. At least to a first approximation, we might thus take a constructive proof to be a sequence of mathematical propositions which are axioms of intuitionistic mathematics or which follow constructively from such statements. Similarly, the relevant notion of constructivity is often first illustrated by explaining the sense in which certain classical proofs are nonconstructive—for example, that they demonstrate existential statements without providing a witness. This in turn is used to argue that logical principles such as the law of the excluded middle have a nonconstructive character. On this basis, it might be conjectured that constructive proofs simply are derivations from mathematical axioms in an intuitionistic formal system such as $\mathcal{IQC}$ which omits such principles.

The foregoing proposal is useful as a benchmark as it suggests that inasmuch as constructive proofs can be understood as resembling classical proofs in form, there is no *prima facie* difficulty in regarding them as abstract objects which stand in a certain structural relation to the propositions which they demonstrate. In particular, it seems reasonable to regard the assertion that a constructive proof demonstrates a statement as itself corresponding to a proposition. Such a view is enshrined in explicit modal logics by the decision to treat $t{:}F$ as a formula expressing this relation. Once this is realized, a heuristic case in favor of the axioms of LP2 and LP3 on the constructive proof interpretation can now be provided.

The case for LP2 is almost immediate due to its similarity in form to the clause $\mathrm{BHK}_{\rightarrow}$. In the case of LP3, note that, although intuitionistically we might be tempted to conflate the assertion of F with that of $t{:}F$ (or $(\exists x)x : F$), classically these statements differ in meaning. However, it seems implicit in traditional discussions of constructive proof that mathematical propositions are seen to be true (or become assertable) in virtue of being proven. It hence seems entirely appropriate to adopt

$t : F \to F$ as a basic axiom of a system like $\mathcal{QLP}$ which we seek to use to provide a classical interpretation of intuitionistic logic.[36]

Before we can attempt to account for the status of QLP4, UPG, and JUG, something more must be said about what it means to quantify over constructive proofs. We may begin by noting that it is already evident from the BHK interpretation that constructive proofs should be regarded as having compositional structure. However, it is also standardly remarked that the proof interpretation itself cannot be taken as providing an inductive specification of the proof conditions of a formula in terms of those of its constituents. The underlying problem is exemplified by the impredicativity implicit in $\mathrm{BHK}_{\to}$—that is, since the proof condition of $F \to G$ is understood to be a construction transforming an arbitrary proof of F into a proof of G, the statement of this condition requires quantification over *all* constructive proofs, not just those which may appear in the proof conditions of F and G.

Both Gödel and Kreisel were acutely aware of this problem. In Gödel's case, this came to the fore in the course of his discussion of the potential use of intuitionistic logic for establishing consistency results in light of the incompleteness theorems—for example, by invoking his embedding of PA into HA (see [22]). One of the earliest places he discusses this is in the 1933 lecture [24], in which he stresses the fact that for a consistency proof to carry conviction, it must be conducted in a system A which embodies only "perfectly unobjectionable, constructive methods."

He then went on to write:

> Heyting's axioms differ from those of the system A only by the fact that the substrate on which the constructions 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 [can be seen from] "Given *any* proof for a proposition p, you can construct a *reductio ad absurdum* from 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 applies particularly to the totality of intuitionistic proofs because of the vagueness of the notion of constructivity. [24, p. 53]

In this passage, Gödel makes three points relevant to the current discussion: (i) in the context of foundational uses of intuitionistic logic, constructive proofs should be understood as serving the same role which natural numbers play in classical metamathematics via arithmetization; (ii) but unlike the natural numbers, we should not think of the totality of constructive proofs as being inductively generated; (iii) this is true both in virtue of the impredicativity of the BHK interpretation and also because of the "vagueness" which he takes to be inherent in the notion of constructivity itself.

Taking points ii and iii together with the condition which he places on the use of the word "any," it would seem to follow that Gödel ought to deny that we are *ever* justified in asserting a universally quantified statement about constructive proofs. The question thus arises as to the status which should be assigned on the constructive proof interpretation not only to the $\mathcal{QLP}$ rule UPG, but also to principles like QLP4 which contain universal proof quantifiers. A literal reading of Gödel's recommendation would appear to suggest that we should simply exclude quantifiers from

the object language of any system introduced for purposes of formalizing reasoning about constructive proofs. To some extent, this reflects the original strategy of Artemov [3] and would leave us with a quantifier-free system resembling $\mathcal{LP}$.

Were we to revert to a system of this sort, however, we would not only lose the ability to express the constructive *provability* of statements in the object language, but we would also lose the ability to express statements equivalent to their own unprovability as appear in the proof of Montague's paradox. It is notable, however, that the basis of Gödel's concern about quantification over constructive proofs does not seem to reside in any potentially paradoxical feature of the BHK interpretation, but rather in the general manner in which the class of constructive proofs is characterized.

A somewhat less radical reaction to Gödel's concerns would thus be to allow object language quantification over constructive proofs to the extent permitted by other constraints. For note that our acceptance of axioms like LP2 and LP3 as valid principles of constructive proof seems largely schematic—that is, we accept the axiom $t : (F \to G) \to (s : F \to t \cdot s : G)$ not in virtue of any particular understanding of what t, s, or $\cdot$ denote, but in virtue of the functional relationship between proofs which we understand to be expressed by this statement. But if this is the case, then it seems that we have little basis to demur from accepting quantified statements like $(\forall x)(\forall y)(\exists z)(x : (F \to G) \to (y : F \to z : G))$ which can be derived in $\mathcal{QLP}$ from $\mathcal{LP}$ axioms by the use of UPG and QLP1–QLP4.

In a later lecture, Gödel [23] himself proposed a system containing proof quantifiers which is similar in form to $\mathcal{QLP}$ in which such principles are derivable.[37] On the other hand, Gödel's concerns about the characterization of the totality of *all* constructive proofs seem more serious when we attempt to justify JUG on the constructive proof interpretation. For not only does the BHK interpretation fail to provide an inductive characterization of this class, but it is also a commonplace of intuitionistic mathematics to resist identifying the extent of constructive reasoning with any particular axiomatic system.[38]

This is evident, for instance, in virtue of the fact that there are long-standing debates within intuitionism as to the constructive *bona fides* of nonclassical principles like Markov's principle, Church's thesis, and bar induction which, while arguably constructive in character, lie somewhere outside the core of traditional intuitionistic mathematics.[39] While this may not itself justify Gödel's use of the term "vagueness" to describe the notion of constructivity, it at least suggests that a degree of caution is warranted when making assertions about properties which are claimed to hold universally of constructive proofs.

One way of codifying this observation in a formal system is precisely to abandon principles like JUG which allow for the internalization of certain forms of quantified reasoning about proofs. For suppose that we have derived $F(x)$ in $\mathcal{QLP}_0$ (i.e., without the use of JUG). Then it may be shown directly on the basis of Proposition 1.6 that we will be able to construct a term $t(x)$ such that $\mathcal{QLP}_0 \vdash t(x) : F(x)$. Since in this case x can be understood as denoting an arbitrary constructive proof, it might then be thought that we are justified in concluding that $(\forall x)t(x) : F(x)$ on analogy with our schematic acceptance of the axioms of $\mathcal{LP}$. But the question which we must now consider is whether our recognition of these facts also puts us in a position where we can justifiably conclude that $s : (\forall x)F(x)$ for any proof term s (as we can derive via JUG by taking $s \equiv \langle t \forall x \rangle$).

According to the interpretation we are currently considering, such a statement would express that s is a constructive proof of the universal statement $(\forall x)F(x)$, which in turn expresses that $F(x)$ is true of *all* items falling under the relevant concept of constructive proof. As we have just seen, however, it is standardly acknowledged that the concept of constructive proof is ineffable at least in the sense of being compatible with expansions in our conception of constructivity which we may not be able to currently foresee.[40] And thus one might reasonably doubt that it is within our power to ever rigorously *prove* that any property holds of *all* constructive proofs.[41]

Such doubts may be further substantiated by recalling the BHK clause for the universal quantifier as formulated by Troelstra and van Dalen [55]:

(BHK$_\forall$) a construction which transforms a proof of $d \in \mathcal{D}$ (where $\mathcal{D}$ is the intended range of the variable x) into a proof of $\varphi(d)$.

On a strict reading of this clause, we cannot accept that the proof term $\langle t\,\forall x\rangle$ denotes a proof of $(\forall x)F(x)$ for it fails to contain x free and thus cannot denote a construction of the correct type. This problem aside, however, BHK$_\forall$ can also be understood as stipulating that before we can characterize the conditions under which such a term ought to be regarded as denoting a proof of a universally quantified statement about constructive proofs, we must be able to provide a suitable mathematical characterization of the domain $\mathcal{D}$ of all such proofs. And it is, of course, our apparent inability to provide such an account which is presently at issue.[42]

If we accept that the foregoing considerations represent a compelling reason to reject JUG on the constructive proof interpretation, then the results of Section 2.3 can now be invoked to give a principled explanation of why Montague's paradox does not arise when we interpret $P(\ulcorner\varphi\urcorner)$ as expressing that φ is constructively provable. For on the one hand, we have seen that the construction of $t(y)$ at step v of (13) is blocked if we do not assume JUG (as the corresponding instance of U_q cannot be derived as a consequence of Proposition 2.12). And on the other, not only is the base system $\mathcal{QLP}_0$ consistent (as a consequence of Theorem 2.8), but it also remains so when self-referential statements of the form (12.b) are adjoined (as a consequence of Proposition 2.13).[43]

The final question to which we must attend is to what extent these conclusions carry over to the informal proof interpretation of $\mathcal{QLP}$. Note first that whereas we have seen that constructive proofs are often characterized in contradistinction to classical ones, informal proofs are customarily characterized in contradistinction to formal proofs—that is, derivations in particular axiom systems such as PA or ZF. The primary characteristic which most authors take to distinguish informal proofs from formal ones is thus that while the former may rely on any evidently "correct" mathematical principles and modes of inference, the latter may only rely on axioms and rules from a fixed axiom system.

For instance, Myhill [44] suggests that we may correctly infer the formalized consistency statement for PA (i.e., $\mathrm{Con}(PA)$) from our understanding of its axioms, despite the fact that this inference cannot be formalized in PA itself. Rav [46] goes one step beyond this by suggesting that many branches of contemporary mathematics possess stable and rigorous standards of proof but have developed without any clear delineation of which principles are to be understood as axioms. Leitgeb [39] goes even further in suggesting that informal proofs may have significant *nonpropositional* constituents—for example, imperative components calling for a particular

construction to be performed, or "intuitive" components calling for nonconceptual representations of mathematical objects to be entertained.

The consensus thus seems to be that inasmuch as we can coherently talk about a unified conception of informal proof, such proofs will lack at least some of the structural properties normally associated with formal proofs. For not only may we be unable to identify a recursive set of axioms and rules on which a given informal proof is based, but it may even be that it is misleading to think of all informal proofs as having the general form of conventional deductive proofs—that is, a sequence or tree of sentences in a fixed mathematical language. The prospect for providing a structural or combinatorial analysis of informal provability thus seems yet again more daunting than in the case of constructive provability. And it is for this reason that Myhill proposed that the best (and perhaps only) way to study informal provability is by axiomatizing its propositional properties.

Myhill suggests that this can be carried out in one of two ways: (1) by treating informal provability as a propositional operator $\Box$; or (2) by treating it as a predicate $P(x)$ of sentences. In the first case, Myhill approvingly cites Gödel's proposal that the $\mathcal{S}4$ axioms reflect valid principles of informal provability. But although this system allows consistency to be demonstrated in the form $\Box\neg\Box\bot$, Myhill rejects modal systems as an adequate medium for reasoning about informal provability in mathematics as the operator $\Box$ is applicable to propositions and not sentences (which he takes to be the objects of mathematical proofs). In the course of developing the second option, Myhill at first proposes that principles like T, K, and Nec—analogous to T, K, and Nec of $\mathcal{S}4$—be added to PA, yielding a theory T_4 extending that of Proposition 1.1. As mentioned in Section 1, however, he then anticipates Montague's paradox by demonstrating that T_4 is inconsistent.

What is of more interest, however, is Myhill's proposal that the inconsistency should be resolved by restricting T, K, and Nec to purely arithmetical substitution instances—that is, sentences in which the predicate $P(x)$ does not itself appear. This resolves Montague's paradox in a technical sense because the sentence δ obtained in the conventional manner from the diagonal lemma in Proposition 1.1 will contain $P(x)$ as a subformula, which in turn means that the inference from iv to v via Nec in the derivation of Proposition 1.1 will be blocked in the relevant system.

But in a conceptual sense, Myhill suggests that this restriction is necessary if our intention is to view $P(x)$ as expressing informal provability in mathematics. For while our intuitions about informal provability may be sufficient to motivate the adoption of principles like T or K, such intuitions presumably do not extend to statements involving the notion of informal provability itself, as it is not evident that this is a mathematical notion.[44] Myhill thus argues that it would be improper to adopt a theory for reasoning about this notion which allowed for iterated applications of $P(x)$.[45]

We may finally observe that the motivation for adopting either of these resolutions to Montague's paradox is in both conceptual and technical accord with the rationale I have presented for rejecting JUG as a valid principle about constructive proof. For as we concluded above, we face an in-principle problem about constructively justifying any principle which asserts that *all* constructive proofs have a given property. If we abandon JUG on this basis, then we move from the theory $\mathcal{QLP}$, which satisfies full internalization (i.e., Theorem 2.4) but which is incompatible with self-reference (i.e.,

Proposition 2.11), to the theory $\mathcal{QLP}_0$, which lacks full internalization but which is compatible with self-reference (i.e., Proposition 2.13).

As we have seen, however, unlike the arithmetical rule Nec and the traditional modal rule Nec, internalization for explicit modal logic comes in degrees. For instance, if we abandon JUG we can no longer internalize the derivation of T_q to yield U_q, which is required for the reconstruction of Montague's paradox in $\mathcal{QLP}$. However, by reasoning in $\mathcal{QLP}_0$ we will still be able to internalize reasoning about proofs which do not involve universal generalization. And we will thus also be able to record our schematic acceptance of various general principles about proofs such as LP2–LP4.

Both abilities represent features intrinsic to our intuitive notion of constructive or informal proof which we would be unable to account for if we either banished proof quantifiers altogether or abandoned principles which allow us to internalize theorems of $\mathcal{QLP}_0$ whose derivation does not require UPG. The adoption of such a system for reasoning about provability can thus be taken to provide a means of resolving Montague's paradox which is at once compatible with the observation that the derivation of the contradiction in Proposition 1.1 originates with the unrestricted application of the rule Nec and also the goal of retaining as many of the desirable aspects of internalization as possible.

Notes

1. Montague's criticism also appears to be grounded in the traditional axiomatic approach to modal logic. In this context, principles like T or Nec might be taken to be essential to our understanding of a notion such as logical or metaphysical necessity. But such an understanding has now largely given way to the model theoretic approach to modal logic heralded by Kripke's "Semantical considerations on modal logic" [36] (which originally appeared in the same volume of *Acta Philosophica Fennica* as Montague's paper). In this context, a wide range of systems are studied, not all of which include axioms analogous to T.

2. Montague does not specify a particular axiomatization of first-order logic, but stipulates that the class of logical axioms be recursive, complete for first-order validity, syntactically closed (i.e., containing no free variables), and closed under modus ponens.

3. More precisely, K can be understood as formalizing the fact that the property expressed by $P(x)$ is preserved under implication, thus allowing for the internalization of conditional reasoning within T. Similarly, 4—which will be recognized as a first-order analogue of the familiar "KK" (or "positive introspection") principle of epistemic logic—can be understood as an axiomatic characterization of the fact that the background system is sufficiently strong to internally develop the argument (3) discussed below.

4. Leitgeb [38] also suggests that Propositions 1.1 and 1.3 are part of a larger family of inconsistency results which arise from dropping the right-to-left direction of TS. He proposes that this family further subdivides into those results requiring T and those not requiring T. It is notable, however, that all of the results in the larger class either require Nec or assume principles like U which subsume its applicability to particular axioms. Although I will concentrate on the role of Nec itself in this paper, many of the conclusions I will draw will apply equally to other principles like U or K which may be used to achieve a similar effect (see, e.g., note 29 below).

5. If S is an arithmetical theory with a recursive set of primitive symbols, we may take "φ" to correspond to the Gödel number of φ. The explicit modal systems which we will consider below circumvent the need for such a device by introducing terms intended to denote proofs in their object language and treating the relation between a proof and the proposition it is taken to demonstrate as primitive.

6. Logical demonstrability thus implies (but is not necessarily implied by) both *mathematical demonstrability* (i.e., provability from true mathematical axioms by mathematically valid means) and also *logical necessity* (i.e., truth in virtue of logical form in the sense presumably intended by Quine and Montague). See Burgess [8] for additional discussion of the difference between these notions and the logical principles we should expect them to satisfy.

7. Proposition 1.1 is sometimes presented as a simplified form of the so-called *Knower paradox*, in which case $P(x)$ is interpreted as expressing some form of idealized knowledge. However, the original derivation given by Kaplan and Montague [32] is based on principles which are somewhat different than those assumed in Proposition 1.1—in particular, rather than Nec, U and an epistemic closure principle (conventionally labeled I) are assumed. But although U is not traditionally taken as a basic modal axiom concerning provability, it is this principle which is most often rejected in the context of the Knower (the locus classicus being Anderson [1]). Although the considerations which constrain the interpretation of $P(x)$ are somewhat different if we understand this predicate to express knowledge instead of provability, most of what I say below will be consistent with adopting this resolution.

8. This highlights the sense in which the assumption of Nec imposes only relatively weak requirements on the possible interpretations of $P(x)$. To adopt Montague's terminology, for instance, Nec merely requires that $P(x)$ *supernumerates* provability in S—that is, $S \vdash P(\ulcorner\varphi\urcorner)$ whenever $\varphi \in \{\psi \mid S \vdash \psi\}$—not that, for example, it semantically represents derivability relative to the standard model of arithmetic. This condition is, of course, compatible with S also entailing that $P(\ulcorner\varphi\urcorner)$ in cases where φ is not derivable in S—for example, in cases where we might think that φ was still knowable or valid for reasons which cannot be formalized in S (or even potentially in any other recursively axiomatizable theory).

9. This observation also highlights the deductive role which is played by Nec in Proposition 1.1. This result can be taken to show that an inconsistency can still be obtained if the right-to-left direction of TS—that is, "if φ, then φ is true"—is replaced with the weaker principle "if φ is provable, then it is true." It is for this reason that Friedman and Sheard dub Nec a "partial approximation" to $\varphi \rightarrow T(\ulcorner\varphi\urcorner)$.

10. This is clear, for instance, from his observations (see [43, pp. 292, 295]) that results like Proposition 1.2 can be understood as generalizing Gödel's observation in [20] that if $P(x)$ is taken to coincide with the extension of $\mathrm{Prov}_S(x)$ for some sufficiently strong theory S extending Q (e.g., $S = PA$), then the principle U will be false for $\varphi \equiv (0 = 1)$. For in this case, the corresponding instance of this schema can be seen to express the fact that the formal consistency of S is provable in S itself, in violation of the second incompleteness theorem. However, it is now known that the generality of this result depends at least to some extent on the strength of S and the precise definition of $\mathrm{Prov}_S(x)$. In particular, we must assume that $\mathrm{Prov}_S(x)$ is not defined in a "nonstandard" manner as discussed by Feferman [15]. We must also assume that S is strong enough to satisfy the other Hilbert–Bernays derivability conditions on $\mathrm{Prov}_S(x)$ needed for the proof of

the second incompleteness theorem. These include not only the Σ^0_1-completeness of S, but also the derivability in S of $\mathrm{Prov}_S(\ulcorner\varphi\urcorner) \rightarrow \mathrm{Prov}_S(\ulcorner\mathrm{Prov}_S(\ulcorner\varphi\urcorner)\urcorner)$ (which can be understood as a special case of the fact that S proves is own Σ^0_1-completeness). The latter feature is known to hold for theories S extending $\mathrm{I}\Delta_0 + \mathrm{exp}$. However, in light of Bezboruah and Sheperdson [6], Pudlák [45], and Berarducci and Verbrugge [5] (see also Franks [18]), the situation for Q and for "weak" theories like $\mathrm{I}\Delta_0 + \Omega_1$ is more complex. It is notable, however, that the sorts of issues raised in the literature on Montague's paradox (or more generally on the Knower paradox or axiomatic theories of truth) do not appear to turn on the amount of induction which S is assumed to satisfy. For this reason, I will follow convention and assume that S can be taken to correspond to PA for the rest of this paper.

11. Gödel claimed the left-to-right direction and conjectured that the converse held as well (which was later confirmed by McKinsey and Tarski [42]). The result is quite robust in the sense that several different embeddings are known—the simplest being simply "append a $\Box$ to every subformula of F." $\mathcal{S}4$ is also by no means the unique modal logic into which $\mathcal{INT}$ can be embedded. For instance, the embedding just described is also an embedding into $\mathcal{S}4 + \mathcal{GRZ}$ (i.e., the closure under Nec of $\mathcal{S}4$ together with the Grzegorczyk axiom). And the embedding "replace every subformula G of F with $G \wedge \Box G$" is an embedding of $\mathcal{INT}$ into $\mathcal{GL}$ (see [9]), which notably lacks the modal reflection axiom T.

12. Proposition 1.4 is historically significant, however, in that it represents one of the first results which demonstrate that it is possible to provide a classical account of intuitionistic validity. In understanding Proposition 1.4 it should thus be taken into account that although $\mathcal{S}4$ proves statements such as $P \vee \neg P$ or $(P \rightarrow Q) \vee (Q \rightarrow P)$ which are not theorems of $\mathcal{INT}$, such statements are not in the range of $(\cdot)^g$. Although this is also true of $\Box P \vee \neg\Box P$, statements such as this can be understood as expressing the decidability of the relation which holds between a constructive proof and a statement which it is claimed to demonstrate.

13. For simplified expositions of $\mathcal{C}$ see [53] and [50].

14. $\mathcal{C}$ was originally formulated to contain a ternary function symbol $\pi(t, u, v)$ such that $\pi(t, u, v) = 0$ is intended to express that t is a constructive proof that terms u and v (which may, in the general case, denote functions of higher type on constructive proofs) are extensionally identical. Kreisel proposed that this latter notion be taken as basic and suggested a means by which statements of the form $\Pi(t; F)$ can be analyzed in terms of $\pi(t, u, v)$ and other operations on proof terms—for example, if $F \equiv P_i(a)$ is atomic, then $\Pi(t, F)$ is defined as $s_i(t)$ (where $s_i(t) = 0$ is intended to express that t is a primitive proof of the proposition $P_i(a)$), if $F \equiv G \wedge H$, then $\Pi(t; F)$ is defined as $\Pi(t_1; G) \cup \Pi(t_1; H)$ (where t_1 and t_2 respectively denote the projection of t onto its first and second components and where $\cup$ is a functional expression denoting classical conjunction), and if $F \equiv G \rightarrow H$, then $\Pi(t, F)$ is defined to be $\pi(t_1, \lambda x.\Pi(x; G) \supset \Pi(t_2(x); H), \lambda x.0)$ (where $\supset$ is a functional expression denoting classical implication). Kreisel argued that the use of the classical connectives in formulating these clauses was justified by the fact that the relation $\Pi(t; F)$ ought to be regarded as decidable, meaning that it is intuitionistically justifiable to apply classical logic to statements of the form $\Pi(t; F) = 0$. This motivates his proposed definition of $\Pi(t; G \rightarrow H)$, which can be understood to strengthen the clause $\mathrm{BHK}_{\rightarrow}$ by requiring not only that t embody a construction (t_2) for transforming proofs of G into proofs of H, but that it also embody a proof (t_1) that this other component operates in this manner.

Without this added requirement, the implicit universal quantification over constructive proofs in BHK$_{\rightarrow}$ might be thought to correspond to an undecidable proof condition. I will return to the basis for this concern in Section 3 below.

15. A caveat is necessary here because Kreisel [34] observes that the converse of Proposition 1.5 does not hold—that is, there are formulas F of $\mathcal{IQC}$ such that $\mathcal{C} \vdash \Pi(t; F) = 0$ for some t but $\mathcal{IQC} \nvdash F$. But the equation of intuitionistic validity with derivability in $\mathcal{IQC}$ can also be challenged. For instance, Kreisel [35] demonstrated that the set of intuitionistically valid formulas in the language $\mathcal{IQC}$ is not recursively enumerable. McCarty [40] showed that this result can be strengthened to show that this set is not even arithmetically definable if weak forms of Church's thesis and Markov's principle are assumed.

16. Goodman attributes the discovery of the paradox independently to Kreisel. And in fact Kreisel [33] gives several indications that he was aware of the danger that $\mathcal{C}$ would become inconsistent were the explicit reflection principle to be included among its axioms.

17. Such a similarity was first observed by Weinstein [57] who explicitly likens the Kreisel–Goodman paradox to a paradox of absolute provability akin to Myhill's.

18. Or more accurately, a term d which may be proved in $\mathcal{C}^*$ to be equal to 0 just in case there exists no proof term t such that $\pi(t, d, 0) = 0$ is provable.

19. In a technical sense, Goodman's preferred resolution to the paradox can thus be compared to typed resolutions of the Knower or liar paradoxes which employ a hierarchy of knowledge or truth predicates (see, e.g., [1] or [52]). However, the general conclusions he draws about the status of quantification over constructive proofs are broadly in line with those for which I will advocate below.

20. The use of explicit modal logic for reasoning about informal provability is thus presumably compatible with Leitgeb's proposal (see [39]) that we need not begin a study of this notion by presenting a conceptual or mathematical analysis of what we take informal proofs to be. Rather, we may begin by surveying our practices and intuitions concerning informal provability and then attempting to formulate an axiomatic theory which reconstructs as large a fragment of these as possible. Leitgeb argues that one of the reasons why we must distinguish between formal and informal proofs is that the latter may contain steps which are not propositional in character (e.g., commands or various sorts of "intuitive" or "nonconceptual" representations) which we would not conventionally consider to be legitimate components of a formal proof. These differences aside, he stresses (e.g., [39, p. 266]) that our practices and intuitions largely bear out the view that informal proofs are properly regarded as abstract objects. Once this is acknowledged, there seems to be no conceptual obstacle to refining our axiomatic treatment of informal provability to include a device for expressing quantification over informal proofs. One of my subsidiary goals in this paper is to illustrate not only the extent to which this approach can be substantiated by using explicit modal logic, but also how Montague's paradox can be understood as enforcing a natural constraint on how quantification over informal proofs must be understood.

21. For instance, axiom LP2 codifies the fact that a proof t of a conditional $F \rightarrow G$ should be such that if s is any proof of F, then the result of applying t to s (denoted by the proof term $t \cdot s$) is a proof of G. Axiom LP4 reflects the fact that we regard the relation

expressed by ":" to be internally verifiable—that is, if t is a proof of F, then we should be able to construct another object (denoted by $!t$) to serve as a proof of this fact. Axiom LP5 reflects the fact that we regard this relation as monotonic in both of its arguments—for example, that if t is a proof of F, then the result of adjoining a proof s to t is still a proof of F. Although these axioms will not play a direct role in the reconstruction of Montague's paradox in $\mathcal{QLP}$, they are needed for Theorems 2.4 and 2.6 below.

22. It is easy to see that we may take $t \equiv b(x) \cdot !a(x)$, where $b(x)$ is a primitive proof term justifying the instance of LP2 corresponding to step iii.

23. In particular, not only is the inclusion of this rule required for the constructive necessitation theorem to extend to $\mathcal{QLP}$, but it is also required if we wish the resulting system to be complete with respect to the Kripke semantics described by Fitting [17]. (Note that in this paper, the name JUG is used for a version of this rule which does not allow for hypotheses. See [11] for discussion of the relation between the two formulations.)

24. It is also instructive to compare JUG to the ω-rule of first-order arithmetic and the Barcan formula of quantified modal logic. (See respectively notes 35 and 40 below.)

25. Upon so doing we obtain $\vdash_{\mathcal{QLP}} \langle b(x) \cdot !a(x) \forall y \rangle : (\forall y) y : ((x : F \wedge G) \rightarrow a(x) \cdot y : G)$.

26. In fact (10.b) is already a theorem of the modal logic $\mathcal{T}$, which differs from $\mathcal{S}4$ in lacking the axiom 4 (i.e., $\Box F \rightarrow \Box\Box F$). But although both $\mathcal{T}$ and $\mathcal{S}4$ satisfy the basic modal necessitation rule Nec, S4Nec is not an admissible rule of $\mathcal{T}$. This suggests that although the sentential principles on which Proposition 1.1 is based correspond most closely to those of $\mathcal{T}$, the first-order derivation given above is most readily compared with a modal derivation in $\mathcal{S}4$. The statement (10.b) is refutable by a similar argument in the modal logic $\mathcal{KD}4$, which is the modal analogue of the theory of T_3 of Proposition 1.3.

27. Another distinction between the derivations is highlighted by the role played by the outer occurrence of $\Box$ which distinguishes (10.b) from (10.a). If this operator were not present, then the application of S4Nec at step v would not be licensed (as otherwise we would generally validate inferences of the form $F \vdash \Box F$). On the provability interpretation of $\Box$, this can be taken to reflect the fact that we should only expect a statement F to be mathematically provable if the premises from which it follows are themselves provable. At the same time, however, the distinction between the material truth of a premise and its provability is difficult to capture in arithmetical theories such as T_1, especially in the case of self-referential statements like (1) which are derivable in Q (and may hence be internalized via Nec). But at least at the conceptual level, it seems possible to imagine the existence of self-referential statements which only are contingently true and thus not provable mathematically. (In fact both the self-referential statements employed in the two paradoxes discussed by Kaplan and Montague in [32]—that is, those of the surprise exam and the Knower—are of this character as are the examples of "accidental" self-reference discussed by Kripke [37].) This suggests that propositional modal logic is in fact a reasonable tool for reconstructing the reasoning of paradoxes which require internalization principles like Nec precisely because it allows us to distinguish the provability of self-referential statements from their truth.

28. It should be noted, however, that such self-referential statements *are* derivable in provability logics such as $\mathcal{GL}$ in virtue of the de Jongh–Sambin fixed point theorem (see [7, Chapter 8]).

29. Similar observations apply to the explicit reconstruction of the derivation of $\neg\Box(D \leftrightarrow \neg\Box D)$ in $\mathcal{KD}4$ mentioned in note 26 above. In particular, although it is possible to derive (13.ix) in the variant of $\mathcal{QLP}$ in which the axiom LP3 is replaced with $x : \neg F \rightarrow \neg y : F$, doing so requires that we derive $(\exists x)x : \neg F \rightarrow \neg(\exists y)y : F$ by two applications of UPG. These must then be internalized using JUG in the same manner as (15).

30. Respectively: $a_1 : (A \leftrightarrow \neg B) \rightarrow (B \rightarrow \neg A), a_2 : (A \leftrightarrow \neg B) \rightarrow (A \rightarrow \neg B)$ and $b : (B \rightarrow A) \rightarrow ((B \rightarrow \neg A) \rightarrow \neg B)$.

31. The reasoning of (13) can also be reconstructed in theories like Goodman's $\mathcal{C}^*$. In particular, the construction of a term playing a role analogous to $t(y)$ is also required in the proof of the Kreisel–Goodman paradox. Although Goodman [25, p. 108] observes this, he neither constructs the relevant term explicitly nor does he note its dependence on the internalization of quantified reasoning about proofs.

32. Condition (16.c) codifies the fact that if we ultimately want to interpret sentences of the form $t{:}F$ as $\mathrm{Prf}(t, \ulcorner F^{\circ} \urcorner)$, then $\mathrm{Prf}(x, y)$ must be a *multiconclusion* proof predicate. This is required because there are formulas F, G such that $t{:}F$ and $t{:}G$ (e.g., if F and G are axioms such that $\mathcal{QLP} \vdash t{:}F$ and $\mathcal{QLP} \vdash s{:}G$, then $\mathcal{QLP}$ also proves by $(t + s) : F$ and $(t + s) : G$ by axiom LP5). Conditions b and c respectively require that $\mathrm{Prf}(x, y)$ is defined in a manner such that (the Gödel numbers of) arithmetical proofs are conjoinable but that a given (Gödel number of an) arithmetical proof only serves to demonstrate finitely many formulas.

33. Note that unlike the other functions used to interpret proof terms, $\boldsymbol{g}$ is defined so that it applies not to the result of interpreting its arguments relative to $(\cdot)^{\circ}$, but rather to their Gödel numbers. This reflects the fact that occurrences of x in the uniform verifier symbol $\langle t \forall x \rangle$ are treated as bound. This manner of treating $\langle \cdot \forall \cdot \rangle$ is in some sense arbitrary, as we will see below that it does not lead to an arithmetically sound interpretation of JUG. However, we will also see that this is inevitable because JUG can be used to mediate an arithmetically unsound inference between statements not involving the universal verifier symbol.

34. It can also be shown that the full system $\mathcal{QLP}$ is consistent by showing how the Kripke semantics for $\mathcal{LP}$ originally proposed by Fitting can be extended to $\mathcal{QLP}$ (see [17], [11]). These semantics also provide a means of interpreting proof quantifiers as ranging over a domain of objects which may be understood as distinct from proof terms themselves. It follows from results like Proposition 2.10 below, however, that this form of semantics is not compatible with viewing these objects as proofs in a formal system such as PA. Theorem 2.8 can, however, be compared with Kreisel's result (sketched in [33, Section 7]) that $\mathcal{C}$ is consistent in virtue of the existence of a formal provability interpretation of statements of the form $\Pi(t; F)$, which is similar in form to the definition of $t{:}F^{\circ}$ given here.

35. It is this observation which most sharply distinguishes JUG from the most familiar form of the arithmetical ω-rule—that is, from $\vdash \varphi(\overline{0}), \vdash \varphi(\overline{1}), \ldots$ conclude $\vdash \forall y \varphi(y)$ for all arithmetical predicates $\varphi(x)$. Although the ω-rule is not admissible with respect to PA (as otherwise we would be able to infer from $PA \vdash \neg\mathrm{Prf}(\overline{0}, \ulcorner \bot \urcorner)$, $PA \vdash \neg\mathrm{Prf}(\overline{1}, \ulcorner \bot \urcorner)$, $\ldots$ to $PA \vdash \forall x \neg\mathrm{Prf}(x, \ulcorner \bot \urcorner)$), it is arithmetically sound in the sense that all of its consequences are true in the standard model N (in fact, together with PA the form of the

rule just cited allows for the derivation of all sentences true in the standard model; see [31]). Since the language of $\mathcal{QLP}$ does not contain expressions which serve the same role as numerals, it is not possible to formulate a precise analogue of this principle for $\mathcal{QLP}$. However, the rule UPG can be understood to approximate the role of the ω-rule in the sense that provability of $F(x)$ in $\mathcal{QLP}$ implies the uniform provability of $F(t)$ for any proof term t (whose arithmetical interpretation will be a numeral). While the images under $(\cdot)^\circ$ of $\mathcal{QLP}_0$-theorems derived via UPG need not always be provable in PA, they will again be true in N. But as we have seen, JUG allows us to derive $\langle r \forall x\rangle : ((\forall x)\neg x : \bot)$ from $r(x) : (\neg x : \bot)$. An arithmetical analogue of this would be an inference from $\vdash \mathrm{Prf}(\boldsymbol{r}(\overline{0}), \ulcorner \neg \mathrm{Prf}(\overline{0}, \ulcorner \bot \urcorner)\urcorner)$, $\vdash \mathrm{Prf}(\boldsymbol{r}(\overline{1}), \ulcorner \neg \mathrm{Prf}(\overline{1}, \ulcorner \bot \urcorner)\urcorner), \ldots$ to $\vdash \mathrm{Prf}(\boldsymbol{r}^*, \ulcorner \forall x \neg \mathrm{Prf}(x, \ulcorner \bot \urcorner)\urcorner)$, where $\boldsymbol{r}^*$ is some closed term which we might imagine is effectively constructed from $\boldsymbol{r}(x)$. But the conclusion of such an inference will not only be false in N but also *refutable* in PA (this time in virtue of the internal provability of the second incompleteness theorem) for any potential denotation of $\boldsymbol{r}^*$.

36. It is more complicated to provide constructive justification for the status of the rule AxNec in $\mathcal{QLP}$. As we have seen, the role of this rule in explicit modal logic is to provide a means of introducing atomic symbols which can be understood as names for proofs of axioms of the system. Note, however, that since $\mathcal{QLP}$ is based on classical logic, it is possible to derive $a : (F \vee \neg F)$ via axiom necessitation for an arbitrary F. Such statements will not, however, appear in the image of $\mathcal{INT}$ under the embedding provided by Theorem 2.5.

37. This system is described in [23, pp. 101–3] and contains atomic formulas of the form zBp, q with the intended interpretation "z is a derivation of q from p." Using this notation, Gödel formulates axioms analogous to LP2, LP3, and LP4 and a single internalization rule ("if q has been proved and a is the proof, [then] 'aBq' "). He then notes that a formula analogous in form to (23) can be derived (in his notation: $aB(\forall u)\neg uB(0 = 1)$) and observes that this expresses the consistency of his system. (He does not, however, present quantifier axioms for proof variables or explain how an explicit modality can be permuted across a proof quantifier in the manner of JUG.)

38. For example, Heyting observed: "Of course, one is never sure that the formal system represents fully any domain of mathematical thought; at any moment the discovering of new methods of reasoning may force us to extend the formal system" [29, p. 5].

39. The adoption or denial of such principles is often taken to characterize distinct notions of constructivity which are studied separately within different branches of intuitionistic mathematics. Thus in seeking to characterize an overall domain of constructive proofs, one cannot simply adopt a "maximalist" strategy by seeking to include all of the non-classical principles which have been considered. This may be borne out technically by observing that certain pairs of these principles—Church's thesis and bar induction—are formally inconsistent (see [55]).

40. One way to make the relationship between JUG and the potential for extending our conception of proof precise is to note a relationship between an axiomatic form of this principle—that is, $(\forall x)t(x) : F(x) \rightarrow \langle t(x)\forall x\rangle : (\forall x)F(x)$—and the traditional Barcan formula (BF) of first-order modal logic—that is, $(\forall x)\Box F(x) \rightarrow \Box(\forall x)F(x)$. In [16], Fitting referred to the former principle as the uniform Barcan formula (UBF) and showed that it is valid in all constant-domain Kripke models for $\mathcal{QLP}$ (this is in accord with the validity of BF in antimonotonic models). Such models can be thought

of as ones in which new proof-objects do not come into existence when we move from world to world. If we were to attempt to interpret $\mathcal{QLP}$ using an arithmetical form of the relational semantics (i.e., one in which worlds are taken to be models of PA), then the arithmetical interpretation of UBF would be invalid for a similar reason—that is, it would incorrectly predict that nonstandard elements satisfying the arithmetical formula formalizing "x is a Gödel number of a proof" cannot come into existence when we move from a world based on the standard model to a world based on a nonstandard one. (This situation can in turn be compared to that which we face with respect to BF in classical quantified provability logic (cf. [7, p. 225].) See [11] for additional results and discussion about constant versus varying domain Kripke semantics for $\mathcal{QLP}$.

41. Recall, for instance, that the property which we must show holds of all constructive proofs x in the course of reconstructing (13) is that x satisfies the explicit reflection principle $x : F \rightarrow F$. I suggested above that the validity of such a principle is evident in cases where we have already acknowledged that a constructive proof t is a proof of F. Note, however, that before we can assert with warrant that this property holds universally, it seems that we must not only agree which objects we will ultimately acknowledge as legitimate constructive proofs, but also whether we are willing to accept them as sufficient evidence for various nonclassical principles. As we may not have made up our mind about a given principle F, it seems that we are not in a position to accept certain items (as we might take to be denoted by terms of the form $\langle r(x)\forall x\rangle$) as a *proof* of its constructive legitimacy.

42. In light of results like Theorem 2.5, we might initially take the inductive definition of the class of $\mathcal{QLP}$ proof terms to provide a specification of $\mathcal{D}$. It is unclear, however, whether definitions of this sort which are specified using classical mathematics are acceptable intuitionistically. On this basis Kreisel [33] refers to such characterizations as "nonstandard" interpretations of the domain of constructive proofs and proposes instead that a "standard" interpretation would have to be a proper extension of such an inductively defined class. Although he provides little indication of what such an interpretation might look like, it seems clear that it would violate Gödel's condition on finite generability.

43. It is also easy to see using the Kripke semantics of [17] that no statement of this form is derivable in the system. $\mathcal{QLP}_0$ can hence be taken to be neutral on the issue of self-reference. On the assumption that the other principles embodying $\mathcal{QLP}_0$ which we have not discussed here (e.g., LP4, LP5) themselves correspond to valid principles about constructive proof, the question remains as to what theory expressible in the language of $\mathcal{QLP}$ best represents a maximally consistent set of principles about this notion. It seems appropriate to leave this question open here not only in virtue of the foregoing observations about the potential open-endedness of constructive provability, but also because of the expressive weakness of the language of $\mathcal{QLP}$ itself. A more thoroughgoing analysis of constructive provability could presumably be conducted by adding the apparatus of explicit modal logic to a first-order arithmetical system such as PA or HA. This would result in an explicit analogue of Shapiro's epistemic arithmetic (see [48]), wherein conventional metamathematical techniques could be used to provide an arithmetization of syntax, thereby guaranteeing the existence of self-referential statements about formal provability predicates in the traditional manner.

44. The evidence that Myhill cites for this is largely drawn from episodes in the history of mathematics where we have, over the course of time, come to accept a principle which we originally regarded with suspicion—for example, the existence of infinite sets, the axiom

of choice, large cardinal axioms. While Myhill observes that the basis for our acceptance of such principles originates from inside mathematics, he also concedes that there may be no means of distinguishing between principles we ultimately adopt for conventional (or even "sociological" or "aesthetic") reasons and those which are somehow forced on us by more intrinsic concerns: "[T]here is no easy answer to the question of what is conventional and what requires proof in mathematics: it does not seem possible sharply to classify proposed new hypotheses into the 'antiseptic' kind, where we may make whatever conventions we see fit with a clear conscience, and the 'real' kind, where what is called for is the discovery of new, hitherto unformalized methods of correct inference" ([44, p. 466]).

45. This proposal also receives technical support from Solovay's well-known result that if we interpret $\Box F$ as $\mathrm{Prov}_{PA}(\ulcorner F \urcorner)$, then the set of modal principles which are true in the standard model for all arithmetical realizations $(\cdot)^*$ coincides with the modal logic $\mathcal{GLS}$. $\mathcal{GLS}$ is derived from the more familiar provability logic $\mathcal{GL}$ (whose axioms are K and the Löb axiom $\Box(\Box F \to F) \to \Box F$, and is thus capable of deriving self-referential sentences such as (10.b) by the de Jongh–Sambin fixed-point theorem as noted above) by adding all instances of the reflection axiom T and closing under modus ponens but *not* the modal necessitation rule Nec. Égré [14] has proposed that when $P(x)$ is understood as expressing knowledge (i.e., in the sense of the knower paradox), Montague's paradox can be avoided by taking the properties satisfied by $P(x)$ to coincide with the modal principles of $\mathcal{GLS}$.

References

[1] Anderson, C. A., "The paradox of the knower," *Journal of Philosophy*, vol. 80 (1984), pp. 338–55. MR 0875995. DOI 10.2307/2026335. 186, 188

[2] Artemov, S., *Operational Modal Logic*, Technical Report no. 95-29, Mathematical Sciences Institute, Cornell University, 1995. 165, 166

[3] Artemov, S., "Explicit provability and constructive semantics," *Bulletin of Symbolic Logic*, vol. 7 (2001), pp. 1–36. MR 1836474. DOI 10.2307/2687821. 159, 165, 166, 167, 169, 172, 173, 174, 176, 182

[4] Beeson, M. J., *Foundations of Constructive Mathematics: Metamathematical Studies*, vol. 6 of *Ergebnisse der Mathematik und ihrer Grenzgebiete 3*, Springer, Berlin, 1985. MR 0786465. 159, 164

[5] Berarducci, A., and R. Verbrugge, "On the provability logic of bounded arithmetic," pp. 75–93 in *Provability, Interpretability and Arithmetic Symposium (Utrecht, 1991)*, vol. 61 of *Annals of Pure and Applied Logic*, 1993. MR 1218656. DOI 10.1016/0168-0072(93)90199-N. 187

[6] Bezboruah, A., and J. C. Shepherdson, "Gödel's second incompleteness theorem for Q," *Journal of Symbolic Logic*, vol. 41 (1976), pp. 503–12. MR 0403947. 187

[7] Boolos, G., *The Logic of Provability*, Cambridge University Press, Cambridge, 1993. MR 1260008. 159, 172, 174, 189, 192

[8] Burgess, J. P., "Which modal logic is the right one?" *Notre Dame Journal of Formal Logic*, vol. 40 (1999), pp. 81–93. MR 1811204. DOI 10.1305/ndjfl/1039096306. 163, 186

[9] Chagrov, A., and M. Zakharyashchev, "Modal companions of intermediate propositional logics," *Studia Logica*, vol. 51 (1992), pp. 49–82. MR 1171642. DOI 10.1007/BF00370331. 187

[10] Curry, H. B., and R. Feys, *Combinatory Logic, Vol. 1: Studies in Logic and the Foundations of Mathematics*, North-Holland, Amsterdam, 1958. MR 0094298. 164

[11] Dean, W., "Epistemic paradox and explicit modal logic," Ph.D. dissertation, City University of New York, New York, 2010. 166, 168, 189, 190, 192, 196

[12] Dean, W., and H. Kurokawa, "Knowledge, proof and the Knower," pp. 81–90 in *Proceedings of the 12th Conference on Theoretical Aspects of Rationality and Knowledge*, ACM, 2009. 196

[13] Dummett, M., *Elements of Intuitionism*, 2nd edition, vol. 39 of *Oxford Logic Guides*, Oxford University Press, New York, 2000. MR 1815706. 180

[14] Égré, P., "The Knower paradox in the light of provability interpretations of modal logic," *Journal of Logic, Language and Information*, vol. 14 (2005), pp. 13–48. MR 2122891. DOI 10.1007/s10849-005-6406-6. 193

[15] Feferman, S., "Arithmetization of metamathematics in a general setting," *Fundamenta Mathematicae*, vol. 49 (1960/1961), pp. 35–92. MR 0147397. 186

[16] Fitting, M., "Quantified LP," Technical Report no. TR-2004019, CUNY Ph.D. Program in Computer Science, City University of New York, N. Y., 2004. 191

[17] Fitting, M., "A quantified logic of evidence," *Annals of Pure and Applied Logic*, vol. 152 (2008), pp. 67–83. MR 2397490. DOI 10.1016/j.apal.2007.11.003. 166, 168, 169, 189, 190, 192

[18] Franks, C., *The Autonomy of Mathematical Knowledge: Hilbert's Program Revisited*, Cambridge University Press, Cambridge, 2009. MR 2597046. DOI 10.1017/CBO9780511642098. 187

[19] Friedman, H., and M. Sheard, "An axiomatic approach to self-referential truth," *Annals of Pure and Applied Logic*, vol. 33 (1987), pp. 1–21. MR 0870684. DOI 10.1016/0168-0072(87)90073-X. 160

[20] Gödel, K., "An interpretation of the intuitionistic propositional calculus," pp. 301–3 in *Kurt Gödel: Collected Works, Vol. I, Publications 1929-1936*, edited by S. Feferman, Oxford University Press, Oxford, 1986. MR 0831941. 159, 162, 163, 167, 186

[21] Gödel, K., "On formally undecidable propositions of *Principia Mathematica* and related systems, I," pp. 144–95 in *Kurt Gödel: Collected Works, Vol. I, Publications 1929-1936*, edited by S. Feferman, Oxford University Press, Oxford, 1986. MR 0831941. 161

[22] Gödel, K., "On intuitionistic arithmetic and number theory," pp. 287–95 in *Kurt Gödel: Collected Works, Vol. I, Publications 1929-1936*, edited by S. Feferman, Oxford University Press, Oxford, 1986. MR 0831941. 181

[23] Gödel, K., "Lecture at Zilsel's," pp. 62–113 in *Kurt Gödel: Collected Works, Vol. III, Unpublished Lectures and Essays*, edited by S. Feferman, Oxford University Press, Oxford, 1995. MR 1332489. 159, 163, 165, 182, 191

[24] Gödel, K., "The present situation in the foundations of mathematics," pp. 36–53 in *Kurt Gödel: Collected Works, Vol. III, Unpublished Lectures and Essays*, edited by S. Feferman, Oxford University Press, Oxford, 1995. MR 1332489. 181

[25] Goodman, N. D., "A theory of constructions equivalent to arithmetic," pp. 101–20 in *Intuitionism and Proof Theory (Buffalo, N.Y., 1968)*, edited by J. Kino and R. Vesley, North-Holland, Amsterdam, 1970. MR 0282797. 164, 190

[26] Halbach, V., *Axiomatic Theories of Truth*, Cambridge University Press, Cambridge, 2011. MR 2778692. DOI 10.1017/CBO9780511921049. 160

[27] Halldén, S., "A Pragmatic Approach to Modal Theory," *Acta Philosophica Fennica*, vol. 16 (1963), pp. 53–64. 159, 163

[28] Heyting, A., *Mathematische Grundlagenforschung: Intuitionismus, Beweistheorie*, Springer, Berlin, 1934. 162

[29] Heyting, A., *Intuitionism: An Introduction*. North-Holland, Amsterdam, 1956. MR 0075147. 162, 191

[30] Hilbert, D., and P. Bernays, *Grundlagen der Mathematik, I*, vol. 40 of *Grundlehren der Mathematischen Wissenschaften*, Springer, Berlin, 1968. MR 0237246. 162

[31] Isaacson, D., "Some considerations on arithmetical truth and the ω-rule," pp. 94–138

in *Proof, Logic and Formalization*, edited by M. Detlefsen, Routledge, London, 1992. MR 1367783. 191

[32] Kaplan, D., and R. Montague, "A paradox regained," *Notre Dame Journal of Formal Logic*, vol. 1 (1960), pp. 79–90. 186, 189

[33] Kreisel, G., "Foundations of intuitionistic logic," pp. 198–210 in *Logic, Methodology and Philosophy of Science (Proc. 1960 Internat. Cong.)*, vol. 44 of *Studies in Logic and the Foundations of Mathematics*, Stanford University Press, Stanford, Calif., 1962. MR 0153565. 158, 159, 163, 164, 188, 190, 192

[34] Kreisel, G., "Mathematical logic," pp. 95–195 in *Lectures on Modern Mathematics, Vol. III*, edited by T. Saaty, Wiley, New York, 1965. MR 0177866. 159, 163, 188

[35] Kreisel, G., "Church's thesis: A kind of reducibility axiom for constructive mathematics," pp. 121–150 in *Intuitionism and Proof Theory (Buffalo, N.Y., 1968)*, North-Holland, Amsterdam, 1970. MR 0278903. 188

[36] Kripke, S. A., "Semantical considerations on modal logic," *Acta Philosophica Fennica*, vol. 16 (1963), pp. 83–94. MR 0170800. 185

[37] Kripke, S. A., "Outline of a theory of truth," *Journal of Philosophy*, vol. 72 (1975), pp. 690–716. 189

[38] Leitgeb, H., "Theories of truth which have no standard models," *Studia Logica*, vol. 68 (2001), pp. 69–87. MR 1950034. DOI 10.1023/A:1011950105814. 160, 185

[39] Leitgeb, H., "On formal and informal provability," pp. 263–299 in *New Waves in Philosophy of Mathematics*, edited by O. Bueno and O. Linnebo, Palgrave Macmillan, 2009. 159, 163, 183, 188

[40] McCarty, C., "Constructive validity is nonarithmetic," *Journal of Symbolic Logic*, vol. 53 (1988), pp. 1036–41. MR 0973099. DOI 10.2307/2274603. 188

[41] McGee, V., "How truthlike can a predicate be? A negative result," *Journal of Philosophical Logic*, vol. 14 (1985), pp. 399–410. MR 0816243. DOI 10.1007/BF00649483. 160

[42] McKinsey, J. C. C., and A. Tarski, "Some theorems about the sentential calculi of Lewis and Heyting," *Journal of Symbolic Logic*, vol. 13 (1948), pp. 1–15. MR 0024396. 187

[43] Montague, R., "Syntactical treatments of modality, with corollaries on reflexion principles and finite axiomatizability," *Acta Philosophica Fennica*, vol. 16 (1963), pp. 153–67. MR 0163841. 157, 158, 159, 162, 186

[44] Myhill, J., "Some remarks on the notion of proof," *Journal of Philosophy*, vol. 57 (1960), pp. 461–71. MR 0111677. 158, 159, 163, 183, 193

[45] Pudlák, P., "On the lengths of proofs of consistency: A survey of results," pp. 65–86 in *Collegium Logicum, Vol. 2*, vol. 2 of *Collegium Logicum, Annals of the Kurt-Gödel-Society*, Springer, Vienna, 1996. MR 1410777. DOI 10.1007/978-3-7091-9461-4_5. 187

[46] Rav, Y., "Why do we prove theorems? Mathematical proof," *Philosophia Mathematica, Series III*, vol. 7 (1999), pp. 5–41. MR 1697024. DOI 10.1093/philmat/7.1.5. 183

[47] Reinhardt, W. N., "Absolute versions of incompleteness theorems," *Noûs*, vol. 19 (1985), pp. 317–46. MR 0824155. DOI 10.2307/2214945. 159

[48] Shapiro, S., "Epistemic and intuitionistic arithmetic," pp. 11–46 in *Intensional Mathematics*, edited by S. Shapiro, vol. 113 of *Studies in Logic and the Foundations of Mathematics*, North-Holland, Amsterdam, 1985. MR 0783643. DOI 10.1016/S0049-237X(08)70138-1. 192

[49] Smoryński, C., *Self-reference and Modal Logic*, Springer, New York, 1985. MR 0807778. DOI 10.1007/978-1-4613-8601-8. 159, 173, 176

[50] Sundholm, G., "Constructions, proofs and the meaning of logical constants," *Journal of Philosophical Logic*, vol. 12 (1983), pp. 151–72. MR 0715827. DOI 10.1007/BF00247187. 159, 162, 187

[51] Tait, W. W., "Gödel's interpretation of intuitionism," *Philosophia Mathematica, Series*

III, vol. 14 (2006), pp. 208–28. MR 2245400. DOI 10.1093/philmat/nkj004. 163

[52] Tarski, A., "The concept of truth in formalized languages," pp. 152–278 in *Logic, Semantics, Metamathematics*, edited by A. Tarski, Oxford University Press, Oxford, 1956. 160, 188

[53] Troelstra, A. S., *Principles of Intuitionism*, vol. 95 of *Lecture Notes in Mathematics*, Springer, Berlin, 1969. MR 0244003. 180, 187

[54] Troelstra, A. S., and H. Schwichtenberg, *Basic Proof Theory*, 2nd edition, vol. 43 of *Cambridge Tracts in Theoretical Computer Science*, Cambridge University Press, Cambridge, 2000. MR 1776976. 168

[55] Troelstra, A. S., and D. van Dalen, *Constructivism in Mathematics, Vol. I: An Introduction*, vol. 121 of *Studies in Logic and the Foundations of Mathematics*, North-Holland, Amsterdam, 1988. MR 0966421. 162, 183, 191

[56] van Atten, M., "The development of intuitionistic logic," *Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta, published electronically April 1, 2009. 162

[57] Weinstein, S., "The intended interpretation of intuitionistic logic," *Journal of Philosophical Logic*, vol. 12 (1983), pp. 261–70. MR 0715830. DOI 10.1007/BF00247190. 164, 188

Acknowledgments

This paper derives from results initially obtained in Dean and Kurokawa [12] and Dean [11]. Thanks are owed especially to Hidenori Kurokawa for drawing my attention to the potential difference between $\mathcal{QLP}$ and its subsystem $\mathcal{QLP}_0$ and for suggesting that the former may not be conservative over the latter. Thanks also to Sergei Artemov, Melvin Fitting, Leon Horsten, and Susan Schweitzer for discussion.

Department of Philosophy
University of Warwick
Coventry CV47AL
United Kingdom
w.h.dean@warwick.ac.uk
http://go.warwick.ac.uk/whdean