THE REVIEW OF SYMBOLIC LOGIC
Volume 10, Number 1, March 2017

GENERALIZING BOOLOS' THEOREM

GRAHAM LEACH-KROUSE
Department of Philosophy, College of Arts and Science, Kansas State University

Abstract. It's well known that it's possible to extract, from Frege's Grudgesetze, an interpretation of second-order Peano Arithmetic in the theory HP^2, whose sole axiom is Hume's principle. What's less well known is that, in *Die Grundlagen Der Arithmetic §82–83* Boolos (2011), George Boolos provided a converse interpretation of HP^2 in PA^2. Boolos' interpretation can be used to show that the Frege's construction allows for any model of PA^2 to be recovered from some model of HP^2. So the space of possible arithmetical universes is precisely characterized by Hume's principle.

In this paper, I show that a large class of second-order theories admit characterization by an abstraction principle in this sense. The proof makes use of structural abstraction principles, a class of abstraction principles of considerable intrinsic interest, and categories of interpretations in the sense of Visser (2003).

§1. Introduction.

1.1. Background. One deep result of Frege's Grundgesetze is this: there exists an interpretation—a theorem-preserving translation—of second order Peano Arithmetic PA^2 into the theory HP^2, where HP^2 consists of nothing but the axioms of second order logic and Hume's principle, the axiom asserting that two concepts have the same number if and only if they are equipollent. In the appendix to *Die Grundlagen Der Arithmetic §82–83* Boolos (2011), Boolos offers a proof of the converse result: there exists an interpretation of HP^2 in PA^2.[1] By combining these two results, it's possible to show that HP^2 and PA^2 are equivalent theories, in the following sense: If ϕ, ψ are sentences in the languages of PA^2 and HP^2, respectively, and ϕ_F, ψ_B are their images under the Frege and Boolos interpretations, respectively, then

$$\mathsf{PA}^2 \vdash \phi \Rightarrow \mathsf{HP}^2 \vdash \phi_F, \tag{F}$$

$$\mathsf{HP}^2 \vdash \psi \Rightarrow \mathsf{PA}^2 \vdash \psi_B. \tag{B}$$

Theorem-hood is preserved by both Frege's translation of Peano arithmetic into Frege arithmetic, and Boolos' translation of Frege arithmetic back into Peano arithmetic. And, what's more, we have the following:

$$\mathsf{PA}^2 \vdash \phi \leftrightarrow (\phi_F)_B. \tag{FB}$$

So, in the process of translation from talk about Peano numbers to talk about Frege numbers and back, theorems are neither lost nor gained.

What accounts for this convergence? In what follows, I venture a partial answer, by isolating a set of properties of PA^2 which give a sufficient condition for this type of

Received: February 19, 2015.

[1] Richard Heck is a co-author, but he attributes the contents of the appendix where the result is to be found strictly to Boolos (Heck, Jr., 2011, xii).

© Association for Symbolic Logic, 2016

doi:10.1017/S1755020316000332

https://doi.org/10.1017/S1755020316000332 Published online by Cambridge University Press

characterizability: any theory T bearing these properties can be characterized by an abstraction principle A_T, in the same way that PA^2 is characterized by Hume's Principle. We'll find that many more theories T than are commonly supposed are characterizable by an abstraction principle in this way. The problem of finding a set of conditions both sufficient and *necessary* for this type of characterizability remains open.

1.2. Aspects of interpretability. My claim that we can isolate properties sufficient for a theory T to be characterized by an abstraction principle, in the same sense that PA^2 is characterized by Hume's Principle, naturally invites the question: "and in what sense is that?" Let me answer that question in two passes. I'll first give a quick and somewhat loose description of how the Boolos interpretation works, with pictures, explain how this can be used to establish the biconditional (FB) above, and indicate how this result will be generalized. I'll then provide the definitions and other technology necessary for a precise statement of Boolos' result and of its promised generalization.

1.2.1. A first pass. Frege's theorem provides us with an interpretation of PA^2 in HP^2. What this means—speaking loosely—is that Frege's methods give us a theorem-preserving translation of the language of PA^2 into the language of HP^2. The Boolos result provides a converse translation. However, although the general statement of the Frege and Boolos results is at the level of the languages of *theories*, it's initially natural to conceptualize these interpretations in terms of their action on *models*. That will be what I do here.

So, consider a standard model of PA^2:

$$0 \qquad 1 \qquad 2 \qquad 3 \qquad 4 \qquad \cdots$$

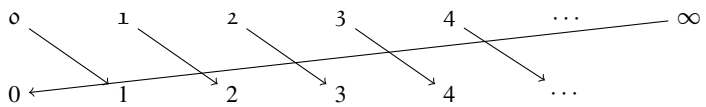
To prove Boolos' theorem, one needs to find a way to make room in this model for the abstracts introduced by Hume's principle. That means, at the very least, all of the finite Frege numbers, and the number of finite Frege numbers: ∞. So, consider the minimal model of Hume's principle:

$$\mathfrak{0} \qquad \mathfrak{1} \qquad \mathfrak{2} \qquad 3 \qquad 4 \qquad \cdots \qquad \infty$$

where

$$\begin{aligned} \mathfrak{0} &= \#\{m : m \neq m\} \\ \mathfrak{n} &= \#\{m : m < n\} \\ \infty &= \#\{x : x = x\}. \end{aligned}$$

There's at least one way of fitting the latter inside the former which is obvious to what logicians might loosely call common sense:



From that idea—which is, essentially, the key idea of the Boolos construction—it's easy to see that one can define a minimal model of the theory HP^2 inside the standard model of PA^2. One needs only to define (in the language of PA^2) a function $\#'$ which sends each second-order entity to the appropriate surrogate Frege number. One can then interpret assertions about the cardinality operator # and Frege numbers in terms of #' and the surrogate numbers. By following a suitable precisification of this recipe, for every statement ϕ in the

https://doi.org/10.1017/S1755020316000332 Published online by Cambridge University Press

language of HP^2, one can write down a corresponding statement ϕ_B in the language of PA^2 which is true in the standard model of PA^2 just in case ϕ is true in the minimal model of HP^2.

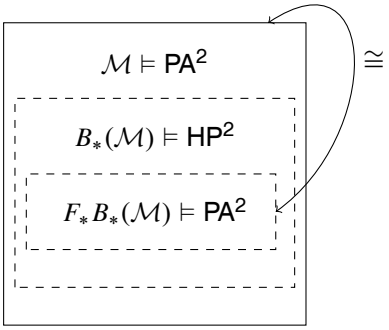
Every model $\mathcal{M}$ of PA^2 (standard or nonstandard) contains a model of HP^2, as above. Since this model is induced by Boolos' construction, let's call it $B_*(\mathcal{M})$. If some statement ϕ in the language of HP^2 is provable (and so true in every model of HP^2), then the corresponding statement ϕ_B is true in every model of PA^2, and in particular, in every model $B_*(\mathcal{M})$. It follows that:

$$\mathsf{HP}^2 \vdash \phi \Rightarrow \mathsf{PA}^2 \vdash \phi_B.$$

Similarly, by Frege's interpretation, every model $\mathcal{M}$ of HP^2 contains a model $F_*(\mathcal{M})$ of PA^2. So we have

$$\mathsf{PA}^2 \vdash \phi \Rightarrow \mathsf{HP}^2 \vdash \phi_F.$$

The model $B_*(\mathcal{M})$ that we build inside of a model $\mathcal{M}$ of PA^2 is, of course, a model of HP^2. So it has inside of it, *another* model of PA^2: $F_*B_*(\mathcal{M})$. PA^2, however, is a categorical theory: any two "standard" models of PA^2—models in which the second-order quantifiers range over all the subsets of the domain—are isomorphic. As it turns out, $F_*B_*(\mathcal{M})$ is standard if $\mathcal{M}$ is. So, when $\mathcal{M}$ is standard, $\mathcal{M}$ is isomorphic to $F_*B_*(\mathcal{M})$:



This result can, in fact, be improved even further. PA^2 is not only categorical. It is, *internally categorical*—it *proves* that every standard model of PA^2 is isomorphic to the natural numbers.[2] As a result, an appropriate formalization of the statement "there's an isomorphism between $\mathbb{N}$ and $F_*B_*(\mathbb{N})$" is a theorem of PA^2. It's this fact that gives us the result mentioned in the introduction:

$$\mathsf{PA}^2 \vdash \phi \leftrightarrow (\phi_F)_B.$$

Because PA^2 recognizes the isomorphism between $\mathbb{N}$ and $F_*B_*(\mathbb{N})$, it recognizes the equivalence of statements ϕ making assertions about numbers and statements $(\phi_F)_B$ making assertions about the numbers-representatives residing within $F_*B_*(\mathbb{N})$.

It's this situation—in which the models of a nice theory define models of an abstraction principle, from which an isomorphic copy of the original model can be uniformly

[2] Since the first-order quantifiers of PA^2 range over numbers, PA^2's theorems about "all models" need to be understood as speaking of "every standard model of PA^2 with a domain contained in the natural numbers."

https://doi.org/10.1017/S1755020316000332 Published online by Cambridge University Press

recovered—that we wish to understand better.[3] The main result of this paper will show that, in fact, this situation can be reproduced with many nice theories. The approach will have two main ingredients. In §1.3, I'll describe a kind of "structural" abstraction principle—an abstraction principle which provides abstract representatives for the isomorphism types of models—which will play the role in our construction that Hume's principle plays in the construction above. In §2, I'll show how to define, within a model of a sufficiently nice theory, a structural abstraction operator, and how to use an idea like Boolos' to recover an isomorphic copy of the ambient model within the induced model of the structural abstraction principle.

1.2.2. A second pass. The previous discussion made use of an unspecified notion of "translating" the language of one theory into another, or of interpreting one theory in another. Let's now make this more precise.

We can begin by saying what languages we intend for our translations to be translations between. In what follows, $\overline{x_n}$, $\overline{X_n}$ will abbreviate sequences of n variables $x_0, \ldots, x_{n-1}$ and $X_0, \ldots, X_{n-1}$. Given a formula ϕ, $\phi(\mathcal{M})$ will denote the set or relation defined in $\mathcal{M}$ by ϕ. By the "type" of a relation, I'll mean the sequence of sorts that the variables occupying the argument places of that relation must range over in order for the resulting atomic formula to be well-formed. S^n will abbreviate a sequence of n sorts, $\underbrace{S \times S \times S \times \cdots \times S}_{n}$.

DEFINITION 1.1.

(1) *$\mathcal{L}_0$ is the language of second-order logic, regarded as a many-sorted language consisting of*
   (a) *Variables ranging over an* object *sort O.*
   (b) *For every n, variables ranging over an n-ary "concept" sort S_n.*
   (c) *An equality relation symbol "=", of type $O \times O$.*
   (d) *For every n, a predication relation symbol* pred_n *of type $O^n \times S_n$. Normally* $\text{pred}_n(\overline{x_n}, X)$ *will be abbreviated by $X\overline{x_n}$.*

(2) *A second-order language is a language extending $\mathcal{L}_0$ by arbitrarily many new relation symbols of any combination of types, but no new sorts.*

(3) *Second order languages in which all relation symbols other than the* pred_n *are of type O^n for some n, we call* standard. *We call nonstandard second-order languages* enriched *languages.*

(4) *For a given theory T, $\mathcal{L}_T$ is the language of T.*

(5) *T is second-order if $\mathcal{L}_T$ is second-order, and T contains the full complement of comprehension axioms. T is standard/enriched according to whether $\mathcal{L}_T$ is.*

This definition disallows constant and function symbols. This is for convenience only—nothing rests on this convention, and we'll sometimes speak as if the language contains these symbols, with the understanding that they can be eliminated in the standard way.

DEFINITION 1.2 (Interpretation). *If Θ, Σ are two second-order languages, then*

(1) *A* translation Γ *of Σ in Θ consists of*
   (a) *A monadic "domain" formula ∂_Γ of Θ.*
   (b) *For each n-ary predicate symbol P of Σ, a Θ-formula $P_\Gamma(\overline{x_n})$.*

[3] As far as I know, attention was first called to this phenomenon in Walsh, Cook, & Visser (2013); the authors there introduce the term "abstraction characterizable" for a model which can be encoded by an abstraction principle in this way.

https://doi.org/10.1017/S1755020316000332 Published online by Cambridge University Press

We indicate that Γ *is such a translation by writing* $\Gamma : \Sigma \to \Theta$. *The translation of predicate symbols* $P \mapsto P_\Gamma$ *can be extended to all formulas of* Σ *in the obvious way, commuting with Boolean connectives, restricting first-order quantifiers to elements of the set defined by* ∂_Γ, *and restricting second-order quantifiers to sub-concepts of* ∂_Γ.[4]

(2) *If* $\mathcal{M}, \mathcal{N}$ *are* Σ/Θ*-models, respectively, then a translation* $\Gamma : \Sigma \to \Theta$ *is an* interpretation *of* $\mathcal{M}$ *in* $\mathcal{N}$ *if there exists a "coordinate map"* $f_\Gamma : \mathrm{dom}(\mathcal{M}) \to \partial_\Gamma(\mathcal{N})$ *such that for all* $\overline{a_n}$ *in* $\mathrm{dom}(\mathcal{M})$,

$$\mathcal{M} \vDash P(\overline{a_n}) \Leftrightarrow \mathcal{N} \vDash P_\Gamma(f_\Gamma(a_0), \ldots, f_\Gamma(a_{n-1})),$$

we write $\Gamma : \mathcal{M} \to \mathcal{N}$ *to indicate that* Γ *is an interpretation of* $\mathcal{M}$ *in* $\mathcal{N}$.

(3) *For translations* $\Gamma : \Sigma \to \Theta$ *we let* $\Gamma_*(\mathcal{N})$ *denote the* Σ*-model induced by taking identity as a coordinate map.*[5] *So,* Γ *interprets* $\mathcal{M}$ *in* $\mathcal{N}$ *if and only if* $\Gamma_*(\mathcal{N}) \cong \mathcal{M}$.

(4) *If* S, T *are* Σ/Θ*-theories, respectively, then a translation* $\Gamma : \Sigma \to \Theta$ *is an interpretation* $\Gamma : S \to T$ *of* S *in* T *if for every* $\mathcal{N} \vDash T$, $\Gamma_*(\mathcal{N}) \vDash S$.

(5) T *interprets* S *if there exists an interpretation of* S *in* T.

There are many obvious examples of interpretation, in the sense described here, to be found in mathematics. For example, ZFC interprets PA, because each model V of ZFC has a set of internally finite ordinals—the objects satisfying an appropriate regimentation of "x is a finite ordinal" in the language $\{\in\}$—which, given appropriate definitions of addition, multiplication, zero, and one, constitute a model of PA. Conversely, PA interprets the theory ZFC − Inf via what's sometimes called the Ackermann coding, defining $n \in m$ by "the n^{th} digit in the binary numeral for m is a 1."

It's an easy consequence of Gödel's completeness theorem that if Γ is an interpretation of S in T, then for every theorem ϕ of S, we have $T \vdash \phi_\Gamma$. Hence, the results (F) and (B) we mentioned at the outset follow from the existence of an interpretation F of PA^2 in HP^2 (the construction of which is due to Frege) and B of HP^2 in PA^2 (the construction of which is due to Boolos).

As our notation suggests, it makes sense to consider the composition of two translations, or of two interpretations:

DEFINITION 1.3.

(1) *If* $\Gamma : \Sigma \to \Theta, \Delta : \Theta \to \Xi$ *are two translations, then the composite translation* $\Delta \circ \Gamma : \Sigma \to \Xi$ *is given by*

$$\partial_{\Delta\circ\Gamma} = (\partial_\Gamma)_\Delta \wedge \partial_\Delta,$$
$$P_{\Delta\circ\Gamma} = (P_\Gamma)_\Delta.$$

Note that relative to this definition, if $\mathcal{W}$ *is a* Ξ*-model, then* $(\Delta \circ \Gamma)_*(\mathcal{W}) = \Gamma_*(\Delta_*(\mathcal{W}))$.

(2) *The identity interpretation* Id_T *of* T *in* T *takes* ∂_{Id_T} *to be* $x = x$ *and* P_{Id_T} *to be* P.

[4] Much more general notions of interpretation are available. For simplicity, we'll take this one to be the default.

[5] That is, the unique model $\mathcal{M}$ such that $\mathrm{dom}(\mathcal{M})$ is the set defined by ∂_Γ in $\mathcal{N}$, and such that for all $\overline{a_n}$ in $\mathrm{dom}(\mathcal{M})^n$, and P of Σ,

$$\mathcal{M} \vDash P(\overline{a_n}) \Leftrightarrow \mathcal{N} \vDash P_\Gamma(\overline{a_n}).$$

https://doi.org/10.1017/S1755020316000332 Published online by Cambridge University Press

(3) *Similarly, the identity interpretation* $\mathrm{Id}_{\mathcal{M}}$ *of* $\mathcal{M}$ *in* $\mathcal{M}$ *takes* $\partial_{\mathrm{Id}_{\mathcal{M}}}$ *to be* $x = x$ *and* $P_{\mathrm{Id}_{\mathcal{M}}}$ *to be* P.

One can easily verify that the composition of interpretations (of models or theories) is an interpretation.[6] One detail that we have left unsettled so far is the conditions under which two interpretations are to be identified. We offer the following criterion:

DEFINITION 1.4.

(1) *Two interpretations* $\Gamma, \Delta : \mathcal{M} \to \mathcal{N}$ *are equal* (*qua interpretations of* $\mathcal{M}$ *in* $\mathcal{N}$) *if there's an isomorphism* $f : \Gamma_*(\mathcal{N}) \to \Delta_*(\mathcal{N})$ *definable in* $\mathcal{N}$.[7]

(2) *Two interpretations* $\Gamma, \Delta : S \to T$ *are equal* (*qua interpretations of* S *in* T) *if there's a formula* ϕ *defining, in every model* $\mathcal{N}$ *of* T, *an isomorphism between* $\Gamma_*(\mathcal{N})$ *and* $\Delta_*(\mathcal{N})$.

Relative to this notion of identity, the Boolos result admits a new characterization. Not only do we have interpretations F and B, but furthermore—by the internal categoricity of PA^2, as in the argument sketch of the previous section—for any model $\mathcal{M}$ of PA^2 we have $B \circ F = \mathrm{Id}_{\mathcal{M}}$.[8] Moreover, the isomorphism witnessing this equality is definable in a uniform way. So we have, in fact, that $B \circ F = \mathrm{Id}_{\mathsf{PA}^2}$ when F and B are viewed as interpretations at the level of theories. Thus, in an important sense, Peano arithmetic is completely characterized by Hume's principle. The shape of any possible arithmetical universe can be recovered by means of Frege's construction, without loss, from an abstraction operator constructed along the lines of Boolos's theorem.

It's this phenomenon connected with PA^2 that we wish to show how to reproduce for a wider class of theories. Specifically, we wish to show that for a fairly wide class of theories T, there exist abstraction principles A_T, and translations $\Gamma : \mathcal{L}_{A_T} \to \mathcal{L}_T$ (analogous to Boolos' construction) and $\Delta : \mathcal{L}_T \to \mathcal{L}_{A_T}$ (analogous to Frege's) such that for standard models $\mathcal{M}$ of T, $\Gamma \circ \Delta$ is an interpretation of $\mathcal{M}$ in $\mathcal{M}$, which is equal (qua interpretation of $\mathcal{M}$ in $\mathcal{M}$) to $\mathrm{Id}_{\mathcal{M}}$.[9] On the basis of this result, we'll show that the same holds at the level of theories (and so of all models, standard and nonstandard), for a slightly narrower class of T.

1.3. Structural abstraction. Our goal was to describe a generalization of the Frege/Boolos result, in such a way as to show that a wide range of theories can be captured, in the sense above, by an associated abstraction principle. How shall we find appropriate abstraction principles that correspond to the theories we consider?

---

[6] Indeed, interpretations, together with this notion of composition, have a rich category-theoretic structure. See Visser (2003).

[7] Here, and later, "definable in $\mathcal{N}$" means "parameter-free-definable in $\mathcal{N}$." So a relation is definable (in $\mathcal{N}$) if there's a formula ϕ satisfied (in $\mathcal{N}$) by all and only the tuples in the relation, and a function is definable (in $\mathcal{N}$) if its graph is.

[8] One possible source of confusion here is the fact that, in our sketch above, we applied the model construction associated with Boolos' interpretation first; so it seems like the identity should be $F \circ B$. But remember, the move from interpretation to model-construction is contravariant: $(B \circ F)_* = F_* \circ B_*$.

[9] In this situation, $\mathcal{M}$ is sometimes said to be "abstraction characterizable," e.g., in Walsh *et al.*, (2013), and Δ is standardly is said to be a "split mono" or "section," when regarded as a morphism $\mathcal{M} \to \Gamma_*(\mathcal{M})$. So, for example, Boolos' interpretation witnesses that Frege's interpretation is a split mono.

https://doi.org/10.1017/S1755020316000332 Published online by Cambridge University Press

The trick will be to use *structural* principles, principles which, intuitively, allow us to introduce abstracts which stand for the isomorphism types of models of a given theory T. More specifically:

Let T be a finitely axiomatized theory in a standard second order language in the signature $\Sigma = \{R_1, \ldots, R_n\}$, where the relation symbols $R_1, \ldots, R_n$ are of adicities $i_1, \ldots, i_n$, respectively.[10] Given T, we wish to construct an open formula ϕ_T of $\mathcal{L}_0$ asserting that a sequence $X_0, \ldots, X_n$ of second-order variables, with X_0 monadic, and $X_1, \ldots, X_n$ of adicities $i_1, \ldots, i_n$, respectively, constitute a model of T. So, let $\phi_T(X_0, \ldots, X_n)$ be the open formula which is obtained by first replacing each occurrence of the predicate symbol R_i in the axioms of T by an occurrence of the variable X_i, then restricting first order quantifiers to objects falling under X_0, and second order quantifiers to subsets of X_0. This sentence asserts that $X_0, \ldots, X_n$ constitute a model of T.[11]

For clarity, we'll now begin to use Fraktur letters (like $\mathfrak{M}, \mathfrak{N}$) to abbreviate our references to tuples $\overline{X_{n+1}}$ of variables which are meant to be intuitively understood as models. So let $\mathfrak{M}$ and $\mathfrak{M}$ be two nonoverlapping sequences of $n+1$ second-order variables $X_0, \ldots, X_n$ and $Y_0, \ldots, Y_n$ of adicities $i_0, \ldots, i_n$, respectively, with $i_0 = 1$. Let $f : \mathfrak{M} \xrightarrow{\sim} \mathfrak{N}$ abbreviate the conjunction of three formulae of $\mathcal{L}_0$: two formulae asserting, respectively, that X_0, Y_0 contain the fields of the relations $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, and one asserting that $\mathfrak{M}$ and $\mathfrak{N}$ are isomorphic when regarded as models in the signature Σ (with second-order quantifiers given a standard interpretation relative to the ambient model).[12] Let $\mathfrak{M} E_T \mathfrak{N}$ abbreviate an open formula of pure second order logic, asserting that $\mathfrak{M}$ and $\mathfrak{N}$ are structurally equivalent (with respect to T), either because they are isomorphic systems modeling T, or because neither is a model of T, in which case both can be assigned to a single "bad abstract" as a matter of bookkeeping:

$$\mathfrak{M} E_T \mathfrak{N} :\equiv \neg(\phi_T(\mathfrak{M}) \vee \phi_T(\mathfrak{N})) \vee (\exists f)[f : \mathfrak{M} \xrightarrow{\sim} \mathfrak{N}]. \tag{4}$$

Finally, let the structural abstraction principle A_T for T be a closed sentence (in the enriched second order language consisting of the language of pure second order logic augmented by a function symbol $\S_T$, of type $S_1 \times S_{i_1} \times \cdots \times S_{i_n} \to O$) which says

[10] The restriction to a finite axiomatization is for simplicity's sake only. It can be considerably relaxed using a method developed by Craig & Vaught (1958). We'll occasionally speak as if these signatures also included function symbols. This talk can be eliminated in favor of talk about purely relational signatures without difficulty.

[11] One can verify that an assignment σ of values to $X_0, \ldots, X_n$ satisfies ϕ_T in $\mathcal{M}$ just in case the second-order model constructed by restricting the various sorts of $\mathcal{M}$ to X_0^σ, and using $X_1^\sigma, \ldots, X_n^\sigma$ to interpret the vocabulary of T, models T.

[12] Explicitly,

$$(\forall x_1 \ldots x_{\max_{m \le n} i_m}) \bigwedge_{j \le n} \left[X_j(x_1, \ldots, x_{i_j}) \to \bigwedge_{k < i_j} X_0(x_k) \right], \tag{1}$$

$$(\forall x_1 \ldots x_{\max_{m \le n} i_m}) \bigwedge_{j \le n} \left[Y_j(x_1, \ldots, x_{i_j}) \to \bigwedge_{k < i_j} Y_0(x_k) \right], \tag{2}$$

$$(\exists f)(\forall x_1 \ldots x_{\max_{m \le n} i_m}) \left[\bigwedge_{j \le n} X_i(x_1, \ldots, x_{i_j}) \leftrightarrow Y_i(f(x_1), \ldots, f(x_{i_j})) \right]. \tag{3}$$

https://doi.org/10.1017/S1755020316000332 Published online by Cambridge University Press

that two systems in the signature Σ are structurally equivalent if and only if they have the same structure:

$$A_T :\equiv (\forall \mathfrak{M})(\forall \mathfrak{N})[\S_T(\mathfrak{M}) = \S_T(\mathfrak{N}) \leftrightarrow \mathfrak{M} E_T \mathfrak{N}]. \tag{5}$$

Let $\mathsf{SA}(T)$ be the second order theory with A_T the only nonlogical axiom.

§2. Generalizing. We now prove the main result of this short paper. We need only one more auxiliary definition.

DEFINITION 2.1.

(1) *Say $X \subseteq \mathcal{M}$ is definably infinite in $\mathcal{M}$ if there's some definable injection $f : X \to X$ in $\mathcal{M}$ which is not onto.*

(2) *If T is a finitely axiomatized standard second-order theory in signature Σ, let T_* be the theory in the signature $\Sigma + P$ (where P is of type O) consisting of T plus a statement "P holds of exactly one object."*

One can think of the models of T_* as "pointed structures": models of T plus a highlighted point. Hence, isomorphism types of T_* correspond to "positions" in isomorphism types of T. This idea—that we can introduce an abstract for each "position" in each isomorphism type of given theory—is the key to the theorem that follows.

Let's now explain how the generalization goes. The construction we sketched in §1.2.1 had two main ingredients. First, interpreting PA^2 in HP^2 required a mechanism that supplied a representative for each element of a standard model of PA^2. This role was played by Hume's Principle. Second, interpreting HP^2 in PA^2 required a way of fitting all the Frege numbers (including ∞) back into a model of PA^2. The trick here was Boolos' "shifting the number line" to free up 0, so that it could be the target of ∞.

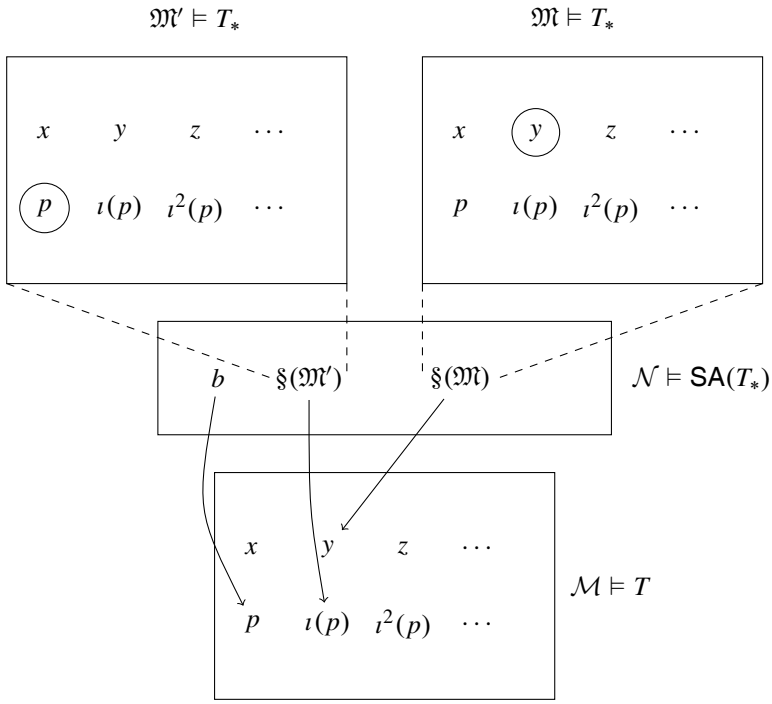
In the construction below, we supply a representative for each element of a given model $\mathcal{M}$ by using the abstract "position" of that element in the $\mathcal{M}$ structure. The isomorphism types of T_* correspond to positions in T-structures, so the theory $\mathsf{SA}(T^*)$ gives us the abstraction principle we need. With a categorical theory T, the internal structure carried by these abstract positions encodes enough information for us to recover the relations of the standard model of T from the position abstracts in a suitable standard model of $\mathsf{SA}(T^*)$—just like the internal structure of the Frege-numbers is rich enough to recover their ordering and arithmetical relations in a standard model of HP^2. So this lets us interpret nice models of T in nice models of $\mathsf{SA}(T_*)$.

In order to interpret $\mathsf{SA}(T^*)$ in T, intuitively, we let each object in a given model $\mathcal{M}$ of T stand for the abstract position that that object happens to occupy. This mostly works. But we have one object in every model of $\mathsf{SA}(T^*)$—the bad abstract[13]—that does not correspond to any abstract position, and we have to find room for that too. To make room for the bad abstract, we "shift the number line," as in the Boolos construction—where the numberline is actually a sequence $p, \iota(p), \iota^2(p), \ldots$ for ι witnessing the existence of a definably infinite set.

In pictures, with pointed structures connected to their abstracts by dashed lines, b standing for the bad abstraction, highlighted positions in structures indicated by circles, and the objects representing a given element of the constructed model $\mathcal{N}$ of $\mathsf{SA}(T_*)$ indicated by arrows:

[13] Recall, the bad abstract is $\S_{T_*}(\mathfrak{M}, X)$ for $\mathfrak{M}$ that do not model T or X that are not singletons within the domain of $\mathfrak{M}$.

https://doi.org/10.1017/S1755020316000332 Published online by Cambridge University Press



THEOREM 2.2. *If $\mathcal{M}$ is a standard model of a finitely axiomatized second order theory T, where*

(1) *T is categorical in the standard semantics,*
(2) *$\mathcal{M}$ has no nontrivial automorphisms,*
(3) *$\mathcal{M}$ has a definably infinite set X given by f, and*
(4) *$\mathcal{M}$ has a definable singleton in $X - f(X)$,*

then there exists translations $\Gamma : \mathcal{L}_{A_{T_*}} \to \mathcal{L}_T$ and $\Delta : \mathcal{L}_T \to \mathcal{L}_{A_{T_*}}$ such that $\Gamma \circ \Delta = \mathrm{Id}_{\mathcal{M}}$, regarded as an interpretation $\mathcal{M} \to \mathcal{M}$, and furthermore such that $\Gamma : \mathsf{SA}(T_*) \to T$ and $\Delta : \mathcal{M} \to \Gamma_*(\mathcal{M})$ are interpretations of theories and models, respectively.[14]

Proof. Because T is categorical, to supply the interpretation $\Gamma : \mathsf{SA}(T_*) \to T$, it is enough to show how to define a translation Γ which is an interpretation of some $\mathcal{N} \vDash \mathsf{SA}(T_*)$ in an arbitrary $\mathcal{M} \vDash T$.

So, let $\mathcal{M}$ be an arbitrary model of T. Let I be the definably infinite set, and let $\iota : I \to I$ be the corresponding definable injection. Let p be the point in the definable singleton contained in $I - \iota(I)$.

Let ∂_Γ simply be $x = x$. Define $(y = \S_{T_*}(\mathfrak{M}, X))_\Gamma$ in three cases.

(1) If $\neg\phi_{T_*}(\mathfrak{M}, X)$, let $y = p$.

In the remaining cases, where $\phi_{T_*}(\mathfrak{M}, X)$, there's an isomorphism $f : \mathfrak{M} \to \mathcal{M}$, (see footnote 11), and X is a singleton; we can refer to the contents of this singleton as c.

[14] Note that Δ is not an interpretation $T \to \mathsf{SA}(T_*)$. Generally, there won't be an interpretation like this, since generally the $\mathcal{L}_{SA(T_*)}$-structure with a single-element first order domain (the bad abstract) and a map taking every element of the second order domain to that object, will be a model of $\mathsf{SA}(T)$ in which no model of T can be defined.

https://doi.org/10.1017/S1755020316000332 Published online by Cambridge University Press

Because $\mathcal{M}$ has standard semantics, the isomorphism f is in $\mathcal{M}$ and because $\mathcal{M}$ has no automorphisms, f is unique, and therefore definable given $\mathfrak{M}$ as a parameter (as the unique isomorphism $\mathfrak{M} \to \mathcal{M}$). So we can set the other two conditions in the definition of $(\S_{T_*})_\Gamma$ as follows:

(2) If $\phi_{T_*}(\mathfrak{M}, X)$, then

(a) if $f(c) = \iota^n(p)$ for some $n \geq 0$, let $y = \iota \circ f(c)$;[15]

(b) if $f(c) \neq \iota^n(p)$ for some $n \geq 0$, let $y = f(c)$.

The relation defined by the resulting formula $(y = \S_T(\mathfrak{M}, X))_\Gamma$ is a function. For suppose both y, y' satisfy this formula with $\mathfrak{M}, X$. We then need to show that $y = y'$. There are two possible cases.

It may be that $\neg\phi_{T_*}(\mathfrak{M}, X)$. In this case, $y, y' = p$, so in this case the claim is immediate. Otherwise, we have a pair of isomorphisms, $f : \mathfrak{M}, X \to \mathcal{M}, A$ and $g : \mathfrak{M}, X \to \mathcal{M}, A'$, given by y, y' respectively, where A, A' are singletons. Let c denote the unique element of X, in this case. $f \circ g^{-1}$ and $g \circ f^{-1}$ are automorphisms of $\mathcal{M}$, so (because $\mathcal{M}$ has no nontrivial automorphisms) $f \circ g^{-1} = g \circ f^{-1} = \mathrm{Id}$. Thus, $f = g$ and $A = A'$. So, if $f(c)$ is a ι-descendent of p, then by definition $y = \iota(f(c)) = \iota(g(c)) = y'$, and otherwise $y = f(c) = g(c) = y'$. Since we've established that $(y = \S_{T_*}(\mathfrak{M}, X))_\Gamma$ defines a function in $\mathcal{M}$, we'll refer to that function as $\S_{T^*}$.

We now verify that $\S_{T_*}$ satisfies the axiom A_{T_*}. First suppose the left side of A_{T_*} holds so $\S_{T_*}(\mathfrak{M}, X) = \S_{T_*}(\mathfrak{M}', X')$. If $\S_{T_*}(\mathfrak{M}, X) = p$, then we must have both $\mathfrak{M}, \mathfrak{M}'$ failing to satisfy ϕ_{T_*}, and hence the right side of A_{T_*} is true. Otherwise, we have isomorphisms $f : \mathfrak{M}, X \to \mathcal{M}, A$ and $g : \mathfrak{M}', X' \to \mathcal{M}, A$, for some singleton A (where, by construction of $\S_{T_*}$, A either contains $\S_{T_*}(\mathfrak{M}, X)$, or contains its ι-predecessor). If we compose these, we have an isomorphism witnessing the right side of the abstraction principle.

Now suppose the right side of the abstraction principle is true, so that either $\mathfrak{M}, X$ and $\mathfrak{M}', X'$ don't satisfy ϕ_{T_*} and both are mapped to p, or $\mathfrak{M}, X$ and $\mathfrak{M}', X'$ are isomorphic standard models of T_*. In the second case, $\mathfrak{M}$ and $\mathfrak{M}'$ are each isomorphic to $\mathcal{M}$, and the resulting isomorphisms f, g must take the contents c, c' of X, X' to the same point (if they did not, then the composition $f \circ g^{-1}$ would be a nontrivial automorphism of $\mathcal{M}$). Hence it's easy to confirm that $\mathfrak{M}, X$ and $\mathfrak{M}, X'$ are mapped to the same point by $\S_{T_*}$.

This establishes that there's an interpretation $\Gamma : \mathsf{SA}(T^*) \to T$. It remains to show that there's an interpretation $\Delta : \mathcal{M} \to \Gamma_*(\mathcal{M})$ such that $\Gamma \circ \Delta = \mathrm{Id}_\mathcal{M}$. Let ∂_Δ be given by the set of abstracts other than p (p is evidently definable in $\Gamma_*(\mathcal{M})$). For simplicity, we show how R_Δ is defined for R dyadic—the generalization to higher adicities should be clear. For each dyadic relation symbol R_i in $\mathcal{L}_T$, let

$$(R_i)_\Delta(\S_{T_*}(\mathfrak{M}, X), \S_{T_*}(\mathfrak{M}', X'))_\Delta = (\exists f)[f : \mathfrak{M}' \xrightarrow{\sim} \mathfrak{M} \wedge X_i(f(c'), c),$$

where c, c' are the unique occupants of X, X', and X_i is the $i + 1^{th}$ variable in the sequence of variables $\mathfrak{M}$. To see that this is well-defined, suppose $a = \S_{T_*}(\mathcal{M}_0, Y_0) = \S_{T_*}(\mathcal{M}_1, Y_1)$, so that there's an isomorphism $g : \mathcal{M}_1, Y_1 \xrightarrow{\sim} \mathcal{M}_0, Y_0$, and similarly suppose $b = \S_{T_*}(\mathcal{M}_0', Y_0') = \S_{T_*}(\mathcal{M}_1', Y_1')$ as witnessed by g'. Then if $(\exists f)[\mathcal{M}_0' \xrightarrow{\sim} \mathcal{M}_0 \wedge X_i(f(c_0'), c_0)$ is witnessed by f, the map $g^{-1} \circ f \circ g' : \mathcal{M}_1' \to \mathcal{M}_1$ can be verified to witness the corresponding statement for $\mathcal{M}_1', Y_1'$ and $\mathcal{M}_1, Y_1$.

[15] As Frege taught us, given an injection f, the relation "$x = f^n(y)$ for some n" is definable in a standard model of a second-order language.

https://doi.org/10.1017/S1755020316000332 Published online by Cambridge University Press

Within $\Gamma_*(\mathcal{M})$, Δ gives us a structure isomorphic to $\mathcal{M}$ via the following map f: $f(x) = x$, for all x which are not ι-descendants of p, and $f(x) = \iota(x)$ for all x (including p) which are ι-descendants of p. This isomorphism $\mathcal{M} \xrightarrow{\sim} (\Delta_*(\Gamma_*(\mathcal{M})) = (\Gamma \circ \Delta)_*(\mathcal{M})$ is definable in $\mathcal{M}$, so this completes the proof. □

The remaining result is an easy corollary, given the right condition on T—in this case, the concept of *internal categoricity*, due to Väänänen (2012).

DEFINITION 2.3. *A finitely axiomatized second order theory T is internally categorical if it is a theorem of second order logic that*

$$\phi_T(\mathfrak{M}) \wedge \phi_T(\mathfrak{M}') \rightarrow (\exists f)[f : \mathfrak{M} \xrightarrow{\sim} \mathfrak{M}'].$$

COROLLARY 2.4. *If T is internally categorical, and T proves*

(1) *There are no automorphisms of the domain,*
(2) *There's a definably infinite set X, witnessed by ι,*
(3) *$X - \iota(X)$ contains a definable singleton,*

then there exists an interpretation $\Gamma : \mathsf{SA}(T_*) \rightarrow T$, and a translation $\Delta : \mathcal{L}_T \rightarrow \mathcal{L}_{\mathsf{SA}(T_*)}$ such that $\Gamma \circ \Delta = \mathrm{Id}_T$ regarded as an interpretation $T \rightarrow T$.

Proof. Make the appropriate substitutions in the previous proof. □

It's worth noting that the hypotheses of this corollary are not very demanding. One can take any theory T satisfying the hypotheses of Theorem 2.2, and add finitely many axioms true in the standard model to get a theory T' satisfying the hypotheses of the corollary.

§3. Conclusion. We've shown, as promised, that many more theories than are commonly supposed can be characterized by abstraction principles, in one of the interesting senses that PA^2 can be thusly characterized. In the sense of characterization requiring that the standard model $\mathcal{M}$ of a categorical theory T can be recovered from the model of an abstraction principle, where that model is definable in $\mathcal{M}$, nothing is required but a few weak assumptions, which are satisfied by a wide range of foundationally important theories—for example second order ZFC plus "there are no inaccessible cardinals," or the second-order theory of $\mathbb{R}$ as a complete ordered field.[16] And any theory T satisfying the assumptions of the main theorem can also satisfy the assumptions of the corollary, so long as we are willing to add a finite number of axioms true in a standard model of T—so any of these foundationally important theories is (modulo a trivial addition) by its own lights isomorphic to a copy of itself defined by means of an abstraction principle.

Strikingly, our interpretation made use of a family of abstraction principles that—in most cases—have virtually no intrinsic deductive strength. In particular, A_T for T with only infinite models will be unable to prove that there exists more than one object.

This suggests an important lesson we might draw from the main result of this paper. Hume's principle plays *two distinguishable roles* in the Frege/Boolos result that we began with. It functions first, to supply an infinite collection of objects out of which to construct a model of PA^2, and second, it makes available a definable choice-like function with which to construct the desired model in a parameter-free way.

[16] Incidentally, this settles a problem raised by Walsh *et al.*, (2013). In that setting, Walsh proves, under the assumption that $V = L$ that $\mathbb{R}$ is characterized in this sense by the abstraction principle *New Burali-Forti*; he then poses the question of whether $\mathbb{R}$ can be characterized without this assumption.

https://doi.org/10.1017/S1755020316000332 Published online by Cambridge University Press

In our arguments above, these two possible roles come apart dramatically. The abstraction principle plays its choice-like function, but it's only the ambient model that supplies the infinite set of objects. This indicates at least two avenues for further research. First—to what extent can we systematically distinguish between these two roles played by abstraction principles? Second—can every theory T satisfying the hypotheses of our corollary be interpreted by $\mathsf{SA}(T_*) + U(T)$, where $U(T)$ is some second-order statement about the cardinality of the universe?

§4. Acknowledgements. Thanks are due to the audience at the March 2016 ASL meeting in Chicago, where I first presented this material, for their encouragement and many helpful suggestions—especially to John Baldwin, Roy Cook, and Sean Ebbels-Duggan. Thanks are also due to Shay Logan, for comments on an earlier version of this paper, and to Richard Heck, for helping me get my attributions right. Special thanks are due to Sean Walsh, whose conjectures and observations suggested the whole idea of this paper.

BIBLIOGRAPHY

Boolos, G. S. (2011). Die Grundlagen der Arithmetik §82–83. In Burgess, J. P., editor. *Logic, Logic, and Logic*. Cambridge: Harvard University Press, pp. 315–333.

Craig, W. & Vaught, R. L. (1958). Finite axiomatizability using additional predicates. *The Journal of Symbolic Logic*, **23**(3), 289–308.

Heck, Jr., R. G. (2011). *Frege's Theorem*. Oxford: Oxford University Press.

Väänänen, J. (2012). Second order logic or set theory? *The Bulletin of Symbolic Logic*, **18**(1), 91–121.

Visser, A. (2006). Logic in Tehran. In Enyat, A., Kalantari, I., and Moniri, M., editors. *Proceedings of the Workshop and Conference on Logic, Algebra and Arithmetic*. La Jolla, CA: Association of Symbolic Logic, pp. 284–341.

Walsh, S., Cook, R. T., & Visser, A. (2013). ESSLLI Notes on Abstraction Principles. Unpublished, available on request.

DEPARTMENT OF PHILOSOPHY
KANSAS STATE UNIVERSITY
1116 MID CAMPUS DR NORTH
201 DICKENS HALL
MANHATTAN, KS 66506-0803, USA
E-mail: gleachkr@ksu.edu

https://doi.org/10.1017/S1755020316000332 Published online by Cambridge University Press