Centre de Recherches Mathématiques
CRM Proceedings and Lecture Notes
Volume 53, 2011

Model Theory vs. Categorical Logic: Two Approaches to Pretopos Completion (a.k.a. T^{eq})

Victor Harnik

Dedicated to Michael Makkai on his 70th birthday

Introduction

It was a sunny day in Jerusalem, in the fall of 1980, when Michael Makkai approached me with a question that sounded more like a declaration. "Did you realize that T^{eq} is nothing but the pretopos completion?" Mihaly (pronounced 'Mee-hi') as many of his friends, including myself, call him, spoke with certain excitement, but that was not unusual with him, when the subject was mathematics. This was the first month of a special year on Model Theory organized by Saharon Shelah at the Institute for Advanced Studies of the Hebrew University.

I knew what Shelah's T^{eq} was, but I had no idea about pretoposes. I realized, however, that a remarkable event took place. The very same new mathematical concept came up in two, seemingly totally unrelated, pieces of mathematical research, one of Shelah and the other of Makkai and Reyes. Both works were developed by what we call *theory builders*. In the tradition of good mathematics, both works were seeking to discover *structure*.

Saharon Shelah worked for years to understand the ways in which mathematical models of given theories can be described. His work drastically changed the scene of Model Theory. Especially striking are his results concerning *first order countable complete theories*. He showed that each such theory T belongs to one of two kinds: *either* T has 'many models', meaning that in every uncountable cardinality λ the number $I(\lambda, T)$ of nonisomorphic models of T has the maximal possible value 2^{λ}, *or*, there is a good *structure theory* for the models of T, meaning that the isomorphism type of each model is determined by a system of *cardinal invariants* (a simplest example being the vector spaces over the field of rationals, which are described by one cardinal, i.e. the dimension). As a corollary, Shelah proved the so called Morley conjecture saying that, for T as above, the function $I(\lambda, T)$ is nondecreasing for uncountable values of λ. Shelah's work goes way beyond the first order countable case and gives much information about the class of models of uncountable first order theories, or of theories in infinitary or higher order logics and, even, more

2010 *Mathematics Subject Classification.* Primary 03G03; Secondary 03C45.
This is the final form of the paper.

general classes of structures called by Shelah '*abstract elementary*.' The first detailed account of his impressive work, was published in [21] and substantial additions, including the proof of Morley's conjecture, appeared in the second edition [23]. For an illuminative, non technical description, see [22]. For the results on abstract elementary classes, see [24] and [25].

Shelah had to overcome a great deal of technical difficulties, using imaginative, ingenious ideas. One of these is connected to the present paper. To every first order theory T, he attached an extension T^{eq}, a theory formulated in a richer language. The two theories are closely connected, but Shelah could use T^{eq} to prove new statements about T itself. It is this construction that appeared, in another disguise, in the work of Makkai and Reyes.

Makkai arrived in Montreal in the fall of 1973 and met there Gonzalo Reyes, his friend from the days of a visit in Berkeley. They both were trained as model theorists, but in Montreal is hard to avoid exposure to Category Theory. At the time of Makkai's arrival, Reyes was already won over to categories, especially due to intensive discussions with André Joyal. Mihaly listened, with great reluctance at first, to Gonzalo's 'preaching' but soon found himself engaged in vivid discussion that led, finally, to a massive work summed up in [15].

The main point of this work is that categories, as underlying structures, are much more prevalent in logic and model theory than previously thought. Not only do the models of a given theory form a category (everyone knew that) but the *first order theories* themselves are, or can be seen as, *categories*. The authors take care to emphasize that this idea has been reached under the influence of the work of Lawvere and discussions with Joyal (as well as a previous work of Joyal and Reyes). However, the extent to which this point of view can be elaborated, is the main novelty of this book. Theories are categories and models become functors into the category of sets. Interpretations of theories into each other become also functors between the corresponding categories and can be seen as *generalized* models. An intimate relation between categorical logic and work of Grothendieck in algebraic geometry, esp. the theory of *topoi*, is established.

Each first order theory T is identified by Makkai and Reyes with a category $\mathbb{T}$ of a kind called *logical* (actually, if we work in the framework of *full classical logic*, as model theorists do, $\mathbb{T}$ is what we call a *Boolean* logical category). Makkai and Reyes notice that the class of categories called *pretoposes* by Grothendieck, is a subclass of the logical categories. The pretoposes enjoy a special property called by them *conceptual completeness*. They also prove that each logical category has a closely related pretopos called its *pretopos completion*.

What Makkai was saying in the sunny fall day of Jerusalem was that **the pretopos completion of the logical category $\mathbb{T}$, identified with T, is nothing but the logical category $\mathbb{T}^{\mathrm{eq}}$ which is identified with the first order theory T^{eq}.**

When I was invited to take part in the *Makkaifest*, I seized the opportunity to clarify to myself and convey to others, the two approaches to the same concept that emerged, more or less at the same time, in two contexts that are so vastly different. This expository paper is based on the talk that I gave at the Makkaifest. I thought that this would be a suitable homage to Mihaly on his 70th birthday. It is, as a matter of fact, an homage to the three men involved, Makkai, Reyes and Shelah and to their important contributions to mathematics.

The paper is addressed to a mixed audience of model and category theorists, hoping to give each group a flavor of the way of thinking of the other.

I assume that the prospective reader has some basic knowledge in both fields and yet, I devote Section 1 to preliminaries in both, category theory and model theory. They are brought mainly to establish nomenclature and notations. A special feature is the description of *multisorted* logic which, although very natural, is seldom used in model theory.

In Section 2, I describe the construction of T^{eq} and give a rough, but hopefully understandable, explanation of its usefulness. In Section 3, I spell out the close connection between the category of models of T and those of T^{eq}.

Actually, the second half of Section 3 as well as Sections 4 and 5, can be seen as a *model-theorist's view of the category-theoretical thinking.* To say it in a nutshell, category theory takes with utmost seriousness the universally accepted mathematical principle that *isomorphic structures are the same.* I hope to be able to soften somewhat the reluctance that most of us, model theorists, feel (precisely as Makkai felt initially) when confronted with the mixture of category theory and logic.

Section 4 presents the 'theories as categories' point of view and Section 5 deals with pretoposes. The final Section 6 contains remarks, a brief description of subsequent developments and a discussion concerning the relevance of conceptual completeness to the problem of giving the widest possible definition to the notion of interpretation of theories.

Acknowledgements. First of all, I want to thank Mihaly for his longtime friendship and endless hours of exciting discussions on mathematics. In particular, I thank him for the many explanations concerning the subject of this paper.

I benefited from useful comments and information provided by Bradd Hart, Wilfrid Hodges, Moshe Kamensky and Anand Pillay.

The few diagrams and pictures were drawn using Michael Barr's diagram package.

1. Preliminaries

In this section we describe our terminology and notations. The presentation will be succinct, as we assume the reader to be familiar with the basic concepts of category theory, on one hand, and (first order) model theory, on the other.

Categories. A category $\mathbb{A}$ has *objects* and *morphisms.* We write $f\colon X \to Y$, or $X \xrightarrow{f} Y$, to indicate that f is an $\mathbb{A}$-morphism with *domain* and *codomain* the $\mathbb{A}$-objects X and Y, respectively. Likewise, $F\colon \mathbb{A} \to \mathbb{B}$ and $\mathbb{A} \xrightarrow{F} \mathbb{B}$ will indicate that F is a *functor* from category $\mathbb{A}$ to category $\mathbb{B}$. The partial operation of *composition* of morphisms (or of functors) will be denoted by $\circ$.

For $\mathbb{A}$-objects X and Y, we will let $\mathbb{A}(X, Y)$ be the set of $\mathbb{A}$-morphisms from X to Y (another customary notation is $\mathrm{Hom}_{\mathbb{A}}(X, Y)$).

Let $\mathbb{S}\mathrm{et}$ be the *category* of sets whose class of objects is the class of all sets and whose morphisms are the functions between sets. Thus, the set of functions between given sets A and B, which is designated by logicians as B^A, is also denoted as $\mathrm{Set}(A, B)$ by category theorists. $\mathbb{S}\mathrm{et}$ is a *large* category, i.e., one whose collection of objects is a *proper* class (but the collection of morphisms between any two given objects is a *set*).

A morphism $j\colon X \to Y$ in a category $\mathbb{A}$ is called an *isomorphism* iff it has an *inverse* $k\colon Y \to X$, meaning that both $\mathbb{A}$-morphisms $k \circ j\colon X \to X$ and $j \circ k\colon Y \to Y$ are identity morphisms. Notice that in the category $\mathbb{S}$et, a morphism is an isomorphism iff it is a bijection.

Two objects X and Y are called isomorphic *in* $\mathbb{A}$ iff there exists an $\mathbb{A}$-morphism $j\colon X \to Y$ which is an isomorphism. Likewise, two $\mathbb{A}$-morphisms $X \xrightarrow{f} Y$ and $X' \xrightarrow{f'} Y'$ are called isomorphic in $\mathbb{A}$ if we have isomorphisms $X \xrightarrow{j} X'$ and $Y \xrightarrow{k} Y'$ such that the diagram

$$\begin{array}{ccc} X & \xrightarrow{f} & Y \\ {\scriptstyle j}\downarrow & & \downarrow{\scriptstyle k} \\ X' & \xrightarrow[f']{} & Y' \end{array}$$

commutes (meaning that $k \circ f = f' \circ j$). We define in a similar way isomorphism of arbitrary (finite) *diagrams* in $\mathbb{A}$.

This much for category theory preliminaries, for the moment.

We will use the '$\mapsto$' notation, customary both in logic and category theory, for function description (e.g., '$x \mapsto x^2$' will denote the function $f(x) = x^2$).

Model theory of first order languages. The symbols of a *first order language* are of two kinds: logical and extralogical. The *logical* symbols are common to all first order languages and consist of an infinite sequence of variables as well as of logical connectives (e.g., conjunction or negation), quantifiers, the equality symbol and auxiliary symbols (such as parentheses and commas). The quantifiers and equality symbols operate on variables. What makes the language *first order* is the insistence that variables denote individual elements (rather than more complex objects such as, say, *sets* of elements).

The *extralogical* symbols are peculiar to each particular first order language and consist of sequences of *relation* (or *predicate*), *function* and *individual constant* symbols. Each relation and function symbol has its *arity* (i.e., number of arguments).

A given first order language L allows us to write *formulas* $\varphi(x_1, \ldots, x_n)$ expressing properties of elements (denoted by $x_1, \ldots, x_n$) of any *L-structure*. An L-structure M is specified by a set of elements $d(M)$, the *domain* of M, together with an *interpretation* $\sigma(M)$, *over* $d(M)$, of any extralogical symbol σ of L (e.g., if f is an n-ary function symbol of M, then $f(M)$ is a function $f(M)\colon d(M)^n \to d(M)$).

We will employ the vector notation $\vec{x} = (x_1, \ldots, x_n)$ to denote finite sequences. This will allow abbreviated notations. For example, formulas will appear as $\varphi(\vec{x})$ and $\vec{a} \in d(M)$, will mean that $a_i \in d(M)$, for $1 \leqslant i \leqslant n$, where $\vec{a} = (a_1, \ldots, a_n)$.

We use the notation $M \vDash \varphi[\vec{a}]$ to indicate the fact that a given sequence of elements $\vec{a} \in d(M)$ satisfies, in the structure M, the property expressed by the formula $\varphi(\vec{x})$. This satisfaction relation allows us to define the interpretation of $\varphi(\vec{x})$ in M to be the n-ary relation

$$\varphi(M) = \{\vec{a} \in d(M) : M \vDash \varphi[\vec{a}]\} \subset d(M)^n.$$

Let us mention that what we just described is the most common way of interpreting logical formulas and is called the *classical* interpretation. The first order logic with this interpretation is called *classical first order logic*. There are

alternative ways of interpreting logical formulas, the most notable one being the *intuitionistic* interpretation. In this paper, we deal with classical logic only.

A *sentence* φ (i.e., a formula with no free variables) is either *true* or *false* in M, and we denote this as $M \models \varphi$ and $M \not\models \varphi$, respectively. φ is called *valid* if it is true in every L-structure, and we denote this as $\models \varphi$.

A first order *theory* T in the language L is a set of L-sentences. A *model* of T is an L-structure M such that $M \models \varphi$ for all $\varphi \in T$. We use $M \models T$ to denote the fact that M is a model of T.

If T is a first order theory and φ a sentence in its language, we say that T implies φ, and denote this as $T \models \varphi$, iff $M \models \varphi$ for all $M \models T$ (we also say that φ is a T-theorem).

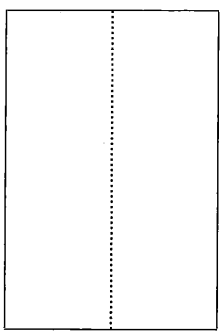
Here is a very simple and illustrative example of a first order theory $\mathcal{E}$ that will accompany our exposition. The language of $\mathcal{E}$ has a binary relation symbol E as its unique extralogical symbol. The axioms of $\mathcal{E}$ will say that: (a) E is an equivalence relation, (b) E has precisely two equivalence classes, and (c) both classes are infinite. Fact (b) is expressed by the first order axiom

$$\exists x_1 \exists x_2 (\neg E(x_1, x_2) \wedge \forall y (E(x_1, y) \vee E(x_2, y))).$$

Fact (c) is expressed by infinitely many first order axioms (for each $n > 0$, one axiom stating that every equivalence class has at least n distinct elements).

A model of $\mathcal{E}$ can be represented schematically as a rectangle split into two classes by a vertical line.

M :



$E(M) \subset d(M) \times d(M)$

The models of a first order theory T form a large category $\mathrm{Mod}\, T$ whose objects are the models of T and morphisms are the *elementary maps* between models. An elementary map $h\colon M \to N$ is a function $h\colon d(M) \to d(N)$ which preserves satisfaction of formulas, i.e., for any L-formula $\varphi(\vec{x})$ and $\vec{a} \in d(M)$, $M \models \varphi[\vec{a}]$ implies $N \models \varphi[h\vec{a}]$ (where $h\vec{a} = (ha_1, \ldots, ha_n)$, if $\vec{a} = (a_1, \ldots, a_n)$).

Multisorted first order languages. Many of the 'everyday' mathematical structures are more conveniently described as having a domain that is a disjoint union of sets of elements of various *sorts*. For instance, a vector space is naturally construed as a structure with two sorts of elements: scalars and vectors. Also, a category is a structure with two sorts of elements: objects and morphisms. In the model theoretical practice, this situation is dealt with by regarding the various sorts as unary predicates. For our purposes, however, it is more convenient to extend the concepts discussed sofar to *multisorted* languages and structures.

A multisorted language L is based on a set $\mathcal{D} = \mathcal{D}(L)$ of *sort* symbols (the *unisorted* languages that we described above, correspond to the case of a singleton

set $\mathcal{D}$). Any L-structure M will have a set $d(M)$ of elements of sort d, for any $d \in \mathcal{D}(L)$. If $\vec{d} = (d_1, \ldots, d_n)$ is a finite sequence of sort symbols, then we denote $\vec{d}(M) = d_1(M) \times \cdots \times d_n(M)$.

The variables of L are sorted. For each $d \in \mathcal{D}$, we have an infinite sequence of variables of sort d. If x is a such, we denote this as $x : d$ and we also write $\vec{x} : \vec{d}$ to mean that $x_i : d_i$ for all $1 \leqslant i \leqslant n$, where $\vec{x} = (x_1, \ldots, x_n)$, $\vec{d} = (d_1, \ldots, d_n)$.

The extralogical symbols are also sorted, and so are their interpretations. Thus, each n-ary function symbol f has arguments of given sorts $\vec{d} = (d_1, \ldots, d_n)$ and values of a sort d_0. We denote this as $f\colon \vec{d} \to d_0$. The interpretation of f in an L-structure M will be a function $f(M)\colon \vec{d}(M) \to d_0(M)$. We let the reader figure out what do we mean by sorting and interpretation of relation and individual constant symbols, and what are the multisorted L-formulas and their interpretations in L-structures. We define in the obvious way the notion of a multisorted first order theory T, its models as well as the concept of elementary map. Thus, we can talk about the category of models $\operatorname{Mod} T$, also in the multisorted context.

2. The theory T^{eq}

Shelah associates with every theory T in a first order language L, a theory T^{eq} in a language L^{eq}. The language L^{eq} is multisorted, even if L was unisorted.

We start the description of L^{eq} with a preliminary definition. An L-formula $\varepsilon(\vec{x}, \vec{y})$ with $\vec{x}, \vec{y} : \vec{d} = (d_1, \ldots, d_n) \in \mathcal{D}^n(L)$, is called an *equivalence formula of* T iff for every $M \models T$, the relation $\varepsilon(M) \subset \vec{d}(M) \times \vec{d}(M)$ is an equivalence relation.

Before going any further, let me remark that any theory T has a plethora of equivalence formulae. Indeed, if $\varphi(\vec{x}, \vec{z})$ is any L-formula with $\vec{x} : \vec{d}$, $\vec{z} : \vec{d'}$, then the following is an equivalence formula:

$$\varepsilon(\vec{x}, \vec{y}) := \forall \vec{z}\big(\varphi(\vec{x}, \vec{z}) \leftrightarrow \varphi(\vec{y}, \vec{z})\big)$$

(notice that this is an equivalence formula in a strong sense, since $\varepsilon(M)$ is an equivalence relation for *all* L-structures M, not just those that are models of T). Actually, it is an easy exercise to show that any equivalence formula of T is equivalent, *in* T, to one of this form.

The language L^{eq} will be an extension of L. It will have, in addition to the symbols of L, a new sort symbol d_ε for every equivalence formula $\varepsilon(\vec{x}, \vec{y})$ of T, as well as a function symbol $f_\varepsilon\colon \vec{d} \to d_\varepsilon$ (where we assume that $\vec{x}, \vec{y} : \vec{d}$).

The theory T^{eq} is, in turn, an extension of T. Its axioms are those of T and, with every equivalence formula $\varepsilon(\vec{x}, \vec{y})$, also the following two axioms saying that the elements of d_ε are (in one-to-one correspondence with) the equivalence classes of the relation defined by $\varepsilon(\vec{x}, \vec{y})$ and that f_ε sends each tuple to its equivalence class:

$$\forall z \exists \vec{x} f_\varepsilon(\vec{x}) = z$$
$$\forall \vec{x} \forall \vec{y}\big(\varepsilon(\vec{x}, \vec{y}) \leftrightarrow f_\varepsilon(\vec{x}) = f_\varepsilon(\vec{y})\big)$$

(it is understood that $\vec{x}, \vec{y} : \vec{d}$ and $z : d_\varepsilon$).

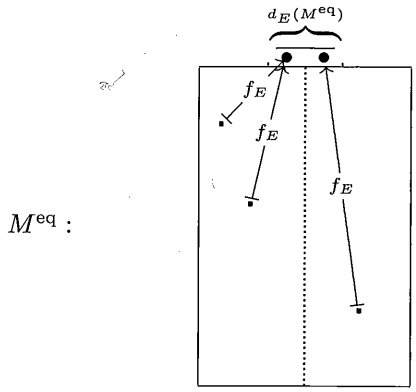
The fact that T^{eq} is an extension of T implies that every model $N \models T^{\mathrm{eq}}$ has a *restriction* to a model $N_0 = N \restriction L \models T$; this is the L-structure obtained from N by ignoring the interpretations of those symbols of L^{eq} that are additional to those of

L. It follows that for all L-sentences φ, if $T \models \varphi$ (i.e., φ is true in all models of T) then $T^{eq} \models \varphi$ as well.

In the case of T^{eq}, we can go also in the converse direction. Every model $M \models T$ has a *canonical* extension to $M^{eq} \models T^{eq}$ in which the interpretations of the additional symbols are:

$d_\varepsilon(M^{eq}) = \{\vec{a}/\varepsilon(M) : \vec{a} \in \vec{d}(M)\}$ (= the set of equivalence classes of $\varepsilon(M)$)
$f_\varepsilon(\vec{a}) = \vec{a}/\varepsilon(M)$.

Recall our illustrative example $\mathcal{E}$, the theory of an equivalence relation E with two equivalence classes, both infinite. $E(x,y)$ is an equivalence formula of $\mathcal{E}$. If $M \models \mathcal{E}$, then the sort $d_E(M^{eq})$ will have two elements and the function $f_E(M^{eq})$ (denoted f_E in the figure below) will send the elements of the left class of M to one of these and those of the right class to the other.



The existence of the canonical extension of models of T implies that T^{eq} is a *conservative* extension of T. This means that T^{eq} has no new theorems expressible in the original language L. In other words, if φ is an L-sentence such that $T^{eq} \models \varphi$, then $T \models \varphi$ as well (indeed, if $M \models T$ and φ is a T^{eq}-theorem, then $M^{eq} \models \varphi$ and, as the truth of the L-sentence φ in M^{eq} depends only on the interpretations of the L-symbols, we have also that $M^{eq}\restriction L \models \varphi$; however, $M^{eq}\restriction L = M$).

Remark. Actually, Shelah's construction in [21] was a little different, because he stuck to unisorted languages. Therefore, T^{eq} is conceived by him as a unisorted theory having a unary predicate (rather than a sort symbol) P_ε, for each equivalence formula ε and additional axioms stating, among others, that distinct P_εs denote disjoint sets. To insure that M^{eq} is an extension of M, the domain of M is identified with $P_=(M^{eq})$, where $P_=$ is the predicate associated with the equivalence formula $x = y$.

For all practical purposes, Shelah's construction is the same with the one that we described above, but one should be aware of the fact that the domain of his M^{eq} is allowed to contain a set of elements that do not belong to any of the sets $P_\varepsilon(M^{eq})$. This set is totally arbitrary and its elements are never looked at, they are just "sitting idle." As a result, M^{eq} is not uniquely defined from M. This is just an aesthetic flaw, which is avoided when we switch to multisorted languages as suggested by Makkai. This suggestion (prompted by his work with Reyes [15]) was adopted by other authors as well (e.g., [1]).

The usefulness of T^{eq}. Shelah showed that the extension of T and $M \models T$ to T^{eq} and M^{eq} is a very useful device. By skilfully switching between T and T^{eq}, he manages to obtain results pertaining to T, that he could not obtain otherwise. In a joint paper with Leo Harrington ([6]), we offer an explanation for the efficacy of Shelah's construction. The details are quite technical, but I will try to give a rough outline that will, hopefully, convey the flavor.

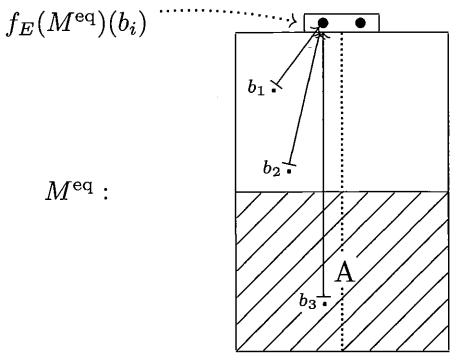
A central notion of model theory is that of a *type of an element over a set.* If M is an L-structure (where, for simplicity, we assume L to be a unisorted language), $A \subset d(M)$ a subset of its domain and $b \in d(M)$ an element, then the type of b over A is the list of all first order properties of b that are expressible in the language (usually denoted as $L(A)$) obtained from L by adding individual constants that denote the elements of A. More formally, the type of b over A *in* M is defined as

$$\mathrm{tp}^M(b, A) = \{\varphi(x, \vec{a}) : \varphi(x, \vec{y}) \in L, \vec{a} \in A, M \models \varphi[b, \vec{a}]\}$$

(following a common practice of model theorists, we do not distinguish between an element of A and the individual constant of $L(A)$ that denotes it). In the sequel, we will refer to $\mathrm{tp}^M(b, A)$ as the *full* type of a over A (because it incorporates *all* $L(A)$-properties).

Shelah came to the conclusion that the notion of full type in the given theory is too crude. To remedy this, he sometimes "steps out of T" and looks at full types in T^{eq}. As we point out in [6], we could achieve the same results by staying inside T, but speaking also about *partial* types that list only *some* of the first order properties of an element over a given set. Rather than giving any precise definition, we'll examine a simple example based on the theory $\mathcal{E}$ that accompanies our exposition.

Assume that $M \models \mathcal{E}$, $A \subset d(M)$ and $b_1, b_2, b_3 \in d(M)$ are equivalent elements such that $b_1, b_2 \notin A$ but $b_3 \in A$. Then we have $\mathrm{tp}^M(b_1, A) = \mathrm{tp}^M(b_2, A) \neq \mathrm{tp}^M(b_3, A)$, the reason for the inequality being that while b_1, b_2 are distinct from all elements of A, b_3 equals one of them. Yet, these three elements have something in common, as they belong to the same equivalence class. We can express this by saying that the *partial* types $\mathrm{tp}^M_E(b_i, A)$ obtained by restricting ourselves to formulas using only E (and *no* equality symbol $=$) are the same for $i = 1, 2, 3$. Rather than doing that, Shelah "works in T^{eq}" and notices that the *full* types $\mathrm{tp}^{M^{\mathrm{eq}}}(f_E(M^{\mathrm{eq}})(b_i), A)$ are the same (simply, because the elements $f_E(M^{\mathrm{eq}})(b_i)$, for $i = 1, 2, 3$, are identical).



Roughly speaking, the T^{eq}-*full* type $tp^{M^{\mathrm{eq}}}(f_E(M^{\mathrm{eq}})(b_i), A)$ that Shelah prefers to consider, is essentially the same as the *partial* type $tp^M_E(b_i, A)$ that we suggest.

We conclude that, after having elaborated the main properties of the notion of full type in a first order theory T, we can proceed in two ways. *Either* work with full types in the extended theory T^{eq} and then return to the original theory T (this is Shelah's option), *or*, stay within T and use a suitably devised notion of *partial* type which is shown to possess the main properties of full types (this is our alternative path suggested in [6]). Shelah's way is conceptually simpler and is universally accepted. Ours is offered as an explanation of what happens when we work in T^{eq}.

3. The models of T^{eq} are essentially the same with those of T

The theory T^{eq} is, as we saw, a non trivial, significant extension of T. Yet, the two theories are closely related. As we noted, each model of T can be canonically extended to one of T^{eq}, a fact that implies that the extension is conservative. *Moreover*, as it turns out, T^{eq} has no other models besides these canonical extensions. To be more precise, every model of T^{eq} is *isomorphic* to the canonical extension of a model of T. Indeed, if $M \vDash T^{\mathrm{eq}}$ and if $M_0 = M\restriction L$ is the restriction of M to a model of T, then a moment of thought will show that M is isomorphic to $M_0^{\mathrm{eq}} = (M\restriction L)^{\mathrm{eq}}$. Thus, we may say that T and T^{eq} have *essentially* the same models. This fact can be stated more perspicuously in terms of the categories of models.

The first observation is that the function $M \mapsto M^{\mathrm{eq}}$, that associates with each model of T its canonical extension to a model of T^{eq}, can be extended to a *functor*

$$\mathrm{Mod}\, T \xrightarrow{(-)^{\mathrm{eq}}} \mathrm{Mod}\, T^{\mathrm{eq}}$$

by sending any map $h\colon M \to N$ in $\mathrm{Mod}\, T$ to its extension $h^{\mathrm{eq}}\colon M^{\mathrm{eq}} \to N^{\mathrm{eq}}$ defined by $h^{\mathrm{eq}}(\vec{a}/\varepsilon(M)) = h\vec{a}/\varepsilon(N)$ (of course, we mean here that $h\vec{a} = (ha_1, \ldots, ha_k)$, if $\vec{a} = (a_1, \ldots, a_k)$).

It is easy to see that, for any two models $M, N \vDash T$, the map $h \mapsto h^{\mathrm{eq}}$ is actually a bijection between the set $\mathrm{Mod}\, T(M, N)$ of elementary maps from M to N and the set $\mathrm{Mod}\, T^{\mathrm{eq}}(M^{\mathrm{eq}}, N^{\mathrm{eq}})$ of elementary maps between the corresponding canonical extensions. In the terminology of category theorists, this means that $(-)^{\mathrm{eq}}\colon \mathrm{Mod}\, T \to \mathrm{Mod}\, T^{\mathrm{eq}}$ is a *full and faithful* functor. In general, a functor $F\colon \mathbb{C} \to \mathbb{D}$ between categories $\mathbb{C}$ and $\mathbb{D}$ is called *full and faithful* iff for any two objects A, B of $\mathbb{C}$, F is a bijection between the set $\mathbb{C}(A, B)$ of $\mathbb{C}$-morphisms from A to B and the corresponding set $\mathbb{D}(FA, FB)$ of $\mathbb{D}$-morphisms.

Is $(-)^{\mathrm{eq}}$ an isomorphism of categories? In other words, is it also bijective on objects? Not quite. The functor $(-)^{\mathrm{eq}}$ happens to be one-to-one on objects but it is *not* surjective. However, it is not *far* from being surjective, as we remarked already, because every model $M \vDash T^{\mathrm{eq}}$ is *isomorphic* to the image $(M\restriction L)^{\mathrm{eq}}$ of its restriction to a model of T. We say that $(-)^{\mathrm{eq}}$ is *essentially surjective on objects.* In general, $F\colon \mathbb{C} \to \mathbb{D}$ is called essentially surjective on objects iff every object B of $\mathbb{D}$ is isomorphic (*in* $\mathbb{D}$) to one of the form FA for some $\mathbb{C}$-object A.

In the terminology of category theorists, a functor, such as $(-)^{\mathrm{eq}}$, which is full and faithful and essentially surjective on objects is called an *equivalence of categories.* This is so, because if we have an equivalence of categories $F\colon \mathbb{C} \to \mathbb{D}$ then the categories $\mathbb{C}$ and $\mathbb{D}$ (are called *equivalent categories* and) are the same from the point of view of category theory. Equivalent categories differ from each other

only in the number of isomorphic copies that each object has. A most illuminative example of equivalent categories is the category $\mathbb{F}\mathbb{G}\mathrm{rp}$ of finite groups (a large category) on one hand and the category $\mathbb{G}_0$ of finite groups whose domain is a set of natural numbers (a small category!) on the other.

Equivalence of categories is a symmetric relation, so we must have also an equivalence functor from $\mathrm{Mod}\,T^{\mathrm{eq}}$ to $\mathrm{Mod}\,T$. Indeed, the map $M \mapsto M{\restriction}L$ can be extended to a *restriction functor* $(-){\restriction}L\colon \mathrm{Mod}\,T^{\mathrm{eq}} \to \mathrm{Mod}\,T$, which is also an equivalence of categories.

Let us introduce the following terminology. If T' is a first order theory in a language L' that extends the L-theory T, we will say that T' is a *tight* extension of T iff the restriction functor $(-){\restriction}L\colon \mathrm{Mod}\,T' \to \mathrm{Mod}\,T$ is an equivalence of categories.

Thus, T^{eq} is a tight extension of T. It has been devised by Shelah with a definite purpose in mind, and it allowed him to overcome certain difficulties. It is, therefore, a *significant* extension. It seems not unreasonable to expect that in the future, someone will extend T^{eq} further, in a significant but still tight fashion, in order to solve some other problem. Or, maybe, one will find a significant but tight extension T' which is inconsistent with T^{eq}, it will go in "another direction," so to speak. Is this possible at all? Or, maybe, T^{eq} is a *largest* tight extension of T?

To begin with, we must formulate these questions in a precise mathematical form. What is an extension of a theory? A simple minded answer to this question would say that a theory T' in a language L' is an extension of an L-theory T if all symbols of L are also symbols of L' and all axioms of T are axioms of T'. This definition, which is adequate for qualifying T^{eq} to be an extension of T, appears to be too restrictive when compared to our intuition arising from mathematical practice. We would like to say that a theory T' is an extension of T when the language of the latter is not necessarily a sub-language of the former but, rather, it can be *translated* into the language of T', such that the translations of the axioms of T become theorems of T'. Two wellknown examples will clarify what we mean.

The field of complex numbers can be defined from that of reals, when each complex number is identified with an ordered pair of reals. This procedure can be generalized to produce an *algebraically closed field* from any *real closed one* (a real closed field is one that satisfies all the first order properties of the field of real numbers). Thus, we may regard the theory of real closed fields as an extension of the (first order) theory of algebraically closed fields (by the way, this would be an instance of a *non tight* extension). An alternative terminology would be to say that there is a *translation* of the theory of algebraically closed fields into that of real closed ones.

Another important example: the standard construction of a model of plane geometry from the field of reals. Here, every geometrical point is identified with an ordered pair of reals and every line with an ordered triple of reals (the coefficients of a "canonical" equation of the line). This construction is easily generalized to all real closed fields showing that the theory of real closed fields is an extension of elementary geometry (i.e. the theory consisting of all *first order sentences* that are true in the Euclidean plane). This would be an instance of a *tight* extension.

One can make up a more general definition of the notion of extension of a first order theory, in the spirit of these two examples. Such a definition requires the existence of a map $\varphi \mapsto \varphi'$ that translates formulas of T into those of T', such that sentences are mapped to sentences and theorems to theorems (meaning that $T' \vDash \varphi'$

whenever $T \vDash \varphi$). Also, this map should allow us to extract a model of T from any one of T'. Definitions of this kind have been produced, the classical reference being [26]. According to the customary terminology, the map $\varphi \mapsto \varphi'$ is called an *interpretation* of T into T'. Hence, authors mostly say that "T is interpretable in T'", rather than "T' is an extension of T." Even [15] uses this language. We will discuss the notion of interpretation of theories and its generalizations in the concluding Section 6.

We can go on and give a general definition of the notion of *tight* extension (using, of course, some category theory). But then, what is a *significant* extension? It seems that model theory, in its traditional form, is not the appropriate context to ask and solve these kind of questions in precise mathematical terms. It is at this point that category theory enters naturally, in a more substantial way.

4. Theories as categories

The work of Makkai and Reyes ([15]) offers a novel, fresh look at the main concepts of model theory. It starts with the innocent looking observation that, given a first order language L, each L-structure M carries with it a function that associates with every L-formula $\varphi(\vec{x})$ its interpretation $\varphi(M)$, which is just an abstract set, i.e. an object of the category Set (which we described in Section 1). Moreover, this function contains all the information needed to describe the structure M, because it specifies the interpretations of all sort, relation, function and individual constant symbols (indeed, if d is a sort symbol, then $d(M) = \varphi(M)$, where $\varphi(x)$ is the formula $x = x$, with $x : d$). Thus, we may *identify* M with the function $\varphi(\vec{x}) \mapsto \varphi(M)$. Actually, Makkai and Reyes go further and identify the structures that happen to be models of a first order theory T, with functions that are *functors*. The first task of the program they developed in [15] can be formulated as follows:

To associate with every first order L-theory T, a category $\mathbb{T}$ whose objects are (essentially) *the L-formulas such that, for any model $M \vDash T$, the function $\varphi(\vec{x}) \mapsto \varphi(M)$ can be extended to a* functor $\mathbb{T} \xrightarrow{M} \mathbb{S}et$.

The construction of $\mathbb{T}$ is very easy to describe:

The objects of $\mathbb{T}$ will not be formulas but, rather, *equivalence* classes $[\varphi(\vec{x})]_T = \varphi(\vec{x})/\sim_T$ of formulas with respect to the equivalence relation defined by

$$\varphi(\vec{x}) \sim_T \varphi'(\vec{x'}) \text{ iff } T \vDash \forall \vec{x}\big(\varphi(\vec{x}) \leftrightarrow \varphi'(\vec{x})\big)$$

where we assume that $\vec{x}, \vec{x'} : \vec{d}$ (notice that in the rightmost formula we changed $\vec{x'}$ to $\vec{x}$, in the argument of φ'). Whenever T is clear from the context, we just write $[\varphi(\vec{x})]_T = [\varphi(\vec{x})]$.

A $\mathbb{T}$-morphism between the objects $[\varphi(\vec{x})]$ and $[\psi(\vec{y})]$ is $[\varphi(\vec{x})] \xrightarrow{[\chi(\vec{x},\vec{y})]} [\psi(\vec{y})]$, where $\chi(\vec{x}, \vec{y})$ is a formula for which

$$T \vDash \forall \vec{x}, \vec{y}(\chi(\vec{x}, \vec{y}) \to \varphi(\vec{x}) \wedge \psi(\vec{y})) \wedge \forall \vec{x}(\varphi(\vec{x}) \to \exists! \vec{y}(\psi(\vec{y}) \wedge \chi(\vec{x}, \vec{y})))$$

(in other words, in every model $M \vDash T$, the formula $\chi(\vec{x}, \vec{y})$ defines the graph of a function $\varphi(M) \to \psi(M)$).

As formulas that are equivalent in the theory T have the same interpretation in any of its models, we see that for $M \vDash T$, we have a well defined function $[\varphi(\vec{x})] \mapsto \varphi(M)$, which extends to a functor $\mathbb{T} \xrightarrow{M} \mathbb{S}et$, in an obvious way. It is this functor that is identified, by Makkai and Reyes, with the model M itself.

Boolean logical categories. Which categories are $\mathbb{T}$ for some first order theory T? A better formulation of this question, in the spirit of category theory, is: which categories are *equivalent to* $\mathbb{T}$ for some T? Makkai and Reyes set out to discover the properties that characterize these categories. They made the crucial observation that all these properties are shared by the category Set. Taking this as a guiding principle, they defined a notion of *Boolean logical category*, in short *b.l.c.* or *b.l. category*, such that the following conditions are met:

(a) Set *is a Boolean logical category.*
(b) $\mathbb{T}$ *is a b.l.c., for all first order theories* T.
(c) *If* $\mathbb{C}$ *is a* small *b.l.c. then it is equivalent to* $\mathbb{T}$, *for some first order theory* T.

Remark. Statement (c) emphasizes the fact that the category that we associate with a given first order theory T does not have to be determined uniquely, but only *uniquely up to equivalence of categories.* This is in accordance with the categorical point of view and it leaves us with some degree of freedom, when deciding how we are going to define the category $\mathbb{T}$. As a matter of fact, Makkai and Reyes' definition in [15] is different from the one that we gave here. They take the objects of $\mathbb{T}$ to be the L-formulas $\varphi(\vec{x})$ themselves, rather than equivalence classes of such. Their morphisms, however, are equivalence classes $[\chi(\vec{x}, \vec{y})]$ of formulas that define functions; this is necessary, in order to insure that the associativity and identity axioms hold for compositions of morphisms.

Rather than going into the technical details of Makkai and Reyes' definition, we will just say that a Boolean logical category is one which is closed under certain *category-theoretical* operations, some of which are also required to have a special property (called "stability under pullbacks"). Each of these category-theoretical operations is a *substitute* for a logical one. A *logical* category is one that is closed only under the category-theoretical operations that substitute the logical operations of *disjunction*, *conjunction* and *existential quantification*, while a *Boolean* logical category is one that is closed under the substitute of *negation* as well.

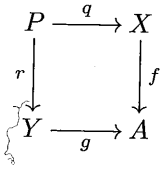
A digression: category-theoretical operations. Our last remarks may sound somewhat cryptical and they call for some clarifications. What are category-theoretical operations? What does it mean that they substitute certain logical operations?

A word of warning: category-theoretical operations are *not* operations in the customary sense of the word, because they are not uniquely defined but defined *uniquely up to isomorphism.*

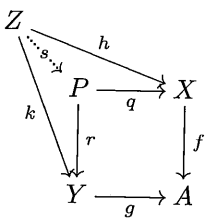
This state of affairs is an outcome of the categorical way of thinking, according to which, category-theoretical properties should be *preserved* and *reflected* by equivalence functors. By this we mean that if F is an equivalence functor then a diagram Γ satisfies a given category theoretical property *iff* so does $F\Gamma$ (category theorists say that a property is reflected by a functor F iff whenever $F\Gamma$ has that property, so does Γ). This principle implies that in category theory, *we cannot define objects uniquely.* Indeed, if X and X' are two objects that are isomorphic in a category $\mathbb{A}$, it is an easy exercise to see that there exists an equivalence functor $F\colon \mathbb{A} \to \mathbb{A}$ which is the identity on all $\mathbb{A}$-objects *except* for X', which is mapped to X, i.e., $FX' = X$. Thus, any category-theoretical property enjoyed by X will be shared by X' as well.

Remark. A syntactical characterization of the first order formulae that express category-theoretical properties has been obtained, independently, by Freyd [5] and Blanc [3].

Let us describe a paradigmatic example of a category-theoretical operation, which is also highly relevant to our context. It is probably known to most readers of this paper. The *pullback* of a diagram $X \xrightarrow{f} A \xleftarrow{g} Y$ is a diagram



which commutes and for all $X \xleftarrow{h} Z \xrightarrow{k} Y$ such that $f \circ h = g \circ k$ there is a *unique* $Z \xrightarrow{s} P$ such that $q \circ s = h$, $r \circ s = k$.



A given diagram may have many pullbacks, however they are all isomorphic to each other (by an isomorphism keeping the original diagram fixed). Yet, it is customary to speak about the *pullback operation* which associates with the given diagram $X \xrightarrow{f} A \xleftarrow{g} Y$ its pullback $X \xleftarrow{q} P \xrightarrow{r} Y$, in spite of the fact that this diagram is not uniquely defined (sometimes, one might even say, carelessly, that the object P alone is *the* pullback of the given diagram!).

A particularly simple example of a pullback occurs in $\mathbb{S}$et, when $X, Y \subset A$ and f, g are the *inclusion* maps, in which case we may take $P = X \cap Y$ with q, r being inclusion maps as well. Thus, the concept of pullback can be seen, not only as a substitute, but as a *generalization* of the *set-theoretical* operation of intersection of subsets of a given set A. It so happens that the same category-theoretical operation of pullback also generalizes the logical operation of *conjunction*. This fact is very important, as we shall see shortly, and is not hard to verify (hint: given formulas $\varphi(x), \psi(x)$ with $x : d$, the object $[\varphi(x) \wedge \psi(x)]$ of $\mathbb{T}$ is, carelessly speaking, the pullback of the diagram

$$[\varphi(x)] \xrightarrow{[\varphi(x) \wedge x = y]} [y = y] \xleftarrow{[\psi(x) \wedge x = y]} [\psi(x)],$$

where $y : d$; remember that $y = y$ is a formula whose interpretation in any model M is the set $d(M)$ of elements of sort d in M).

We conclude this digression with a remark that may look intriguing to some. The identity function from the set of objects of a category to itself is *not* a category theoretical operation! Question: what would be a category-theoretical generalization of this operation?

Natural transformations. We now return to our main theme. Assume that $M, N \vDash T$ are models of T. If we now think of M and N as functors from $\mathbb{T}$ to $\mathbb{S}\mathrm{et}$, it is only natural to ask what is the *category theoretical meaning* of an *elementary map* $h\colon M \to N$. It turns out that h is, essentially, what category theorists call a *natural transformation* between the functors M and N.

In general, given categories $\mathbb{C}, \mathbb{D}$ and functors $F, G\colon \mathbb{C} \to \mathbb{D}$, a *natural transformation* η between F and G, also denoted as $\eta\colon F \to G$, is an indexed family of $\mathbb{D}$-morphisms

$$\{FX \xrightarrow{\eta_X} GX : X \text{ an object of } \mathbb{C}\},$$

such that for every $\mathbb{C}$-morphism $X \xrightarrow{f} Y$, the following diagram (in the category $\mathbb{D}$) commutes

$$\begin{array}{ccc} FX & \xrightarrow{\eta_X} & GX \\ {\scriptstyle Ff}\downarrow & & \downarrow{\scriptstyle Gf} \\ FY & \xrightarrow[\eta_Y]{} & GY \end{array}$$

This notion, defined by Eilenberg and Mac Lane in a seminal paper [4], is surprisingly little known among mathematicians from other fields. It is one of the most important concepts of category theory. The authors just mentioned go as far as to say that "category" has been defined in order to define "functor" and "functor" has been defined in order to define "natural transformation" (this observation is quoted in [10, p. 18]).

We are now in a position that allows us to explain why an elementary map $h\colon M \to N$ can be regarded as a natural transformation. As h preserves the interpretations of first order formulas, we see that for any $\varphi(\vec{x})$, the restriction of h to the set $\varphi(M)$ takes its values in $\varphi(N)$. Obviously, this restriction of h depends on $[\varphi(\vec{x})]$, rather than on $\varphi(\vec{x})$ alone; therefore, we may denote it as $h_{[\varphi]}$. Thus, h determines an indexed family $\{\varphi(M) \xrightarrow{h_{[\varphi]}} \varphi(N) : \varphi \text{ an } L\text{-formula}\}$ of functions (i.e., $\mathbb{S}\mathrm{et}$-morphisms), which is easily seen to be a natural transformation between the functors $M, N\colon \mathbb{T} \to \mathbb{S}\mathrm{et}$ (one should keep in mind that the sets $\varphi(M), \varphi(N)$ are the images, under M and N, of the object $[\varphi(\vec{x})]$ of the category $\mathbb{T}$). Conversely, any natural transformation between the *functors* M and N originates in this way from an elementary map between the *models* M and N.

Categories of functors. With our new point of view, the category $\mathrm{Mod}\,T$ of models of T appears to be one whose objects are functors and morphisms are natural transformations. This is just a particular example of a *category of functors*. It is a remarkable fact that, given any two categories $\mathbb{C}$ and $\mathbb{D}$ we have a category whose objects are the functors from $\mathbb{C}$ to $\mathbb{D}$ and morphisms are the natural transformations between these functors. This category of functors is denoted as $\mathbb{D}^{\mathbb{C}}$.

Having this new concept, we may now say that *the models of the first order theory T are objects of the functor category $\mathbb{S}\mathrm{et}^{\mathbb{T}}$ and the elementary maps between models of T are morphisms of $\mathbb{S}\mathrm{et}^{\mathbb{T}}$*. Actually, for given models $M, N \vDash T$, the elementary maps from M to N are *precisely* the $\mathbb{S}\mathrm{et}^{\mathbb{T}}$-morphisms between M and N (i.e., the natural transformations between the functors $M, N\colon \mathbb{T} \to \mathbb{S}\mathrm{et}$).

In the terminology of category theory, the situation that we just presented is described by saying that $\mathrm{Mod}\,T$ is a *full subcategory* of the functor category $\mathbb{S}\mathrm{et}^{\mathbb{T}}$. In general, a category $\mathbb{C}$ is a *subcategory* of $\mathbb{D}$ iff it is a *substructure* of $\mathbb{D}$, in the

model theoretic sense; this means that the objects and morphisms of $\mathbb{C}$ are objects and morphisms of $\mathbb{D}$ and, moreover, morphisms of $\mathbb{C}$ have the same domain and codomain and compose in the same way in $\mathbb{D}$ as in $\mathbb{C}$. We say that $\mathbb{C}$ is a *full* subcategory iff in addition, for any two objects X, Y of $\mathbb{C}$, $\mathbb{C}(X,Y) = \mathbb{D}(X,Y)$; this means that all $\mathbb{D}$-morphisms from X to Y are already $\mathbb{C}$-morphisms between the same domain and codomain.

It is quite obvious, as we shall explain in a moment, that $\mathrm{Mod}\,T$ is a *proper* subcategory of $\mathbb{S}\mathrm{et}^{\mathbb{T}}$. In other words, there are functors from $\mathbb{T}$ to $\mathbb{S}\mathrm{et}$ that are *not* (identifiable with) models of T. In fact, there are many such functors. This leads to a very natural question: which functors from a b.l. category to $\mathbb{S}\mathrm{et}$ can be seen as models?

Logical functors. Any model $M \vDash T$, seen as a functor, should satisfy some obvious conditions. For example, M sends the objects $[\varphi(\vec{x})], [\psi(\vec{x})], [\varphi(\vec{x}) \wedge \psi(\vec{x})]$ of $\mathbb{T}$, to the sets $\varphi(M), \psi(M), \varphi \wedge \psi(M)$, respectively. These sets should satisfy

$$\varphi \wedge \psi(M) = \varphi(M) \cap \psi(M).$$

In other words, the functor M must transfer the *logical* operation $\wedge$ into the *set theoretical* operation $\cap$. As it so happens that both these operations are generalized by the same category-theoretical pullback operation, it is enough to require M to *preserve pullbacks*. This is just one condition that a functor $\mathbb{T} \to \mathbb{S}\mathrm{et}$ must satisfy in order to be a model of T.

Following this line of thought, Makkai and Reyes defined the notion of *logical functor* between logical categories as being one that preserves the category-theoretical operations under which the *logical* categories are closed. One would be tempted to define also a notion of *Boolean* logical functor. This is not necessary, however. As it turns out, a logical functor whose domain and codomain happen to be *Boolean* logical categories, preserves also the additional, negation-related operations that characterize the Boolean logical categories.

Makkai and Reyes were able to show:

If M is a model of T then the functor $M\colon \mathbb{T} \to \mathbb{S}\mathrm{et}$ which is identified with M is a logical functor.

How about the converse of this statement? Let us call, just for a moment, the functors from $\mathbb{T}$ to $\mathbb{S}\mathrm{et}$ that are identified with models of T by the name "*standard* functors." Is any logical functor from $\mathbb{T}$ to $\mathbb{S}\mathrm{et}$ a standard one? *Essentially* it is. *Essentially*, but not precisely! In fact, any logical functor $N\colon \mathbb{T} \to \mathbb{S}\mathrm{et}$ is *isomorphic* to a standard one. This isomorphism takes place in the functor category $\mathbb{S}\mathrm{et}^{\mathbb{T}}$, meaning that there is a standard functor M such that there exist natural transformations $M \xrightarrow{\eta} N$ and $N \xrightarrow{\theta} M$ which are inverses of each other.

The remarkable statement just made can be said more elegantly and usefully as follows. Let $\mathrm{Mod}'\,T$ be the full subcategory of $\mathbb{S}\mathrm{et}^{\mathbb{T}}$ whose objects are the *logical* functors from $\mathbb{T}$ to $\mathbb{S}\mathrm{et}$. Then $\mathrm{Mod}\,T$ is a full subcategory of $\mathrm{Mod}'\,T$. This means that we have an *embedding* functor $\iota\colon \mathrm{Mod}\,T \to \mathrm{Mod}'\,T$ (i.e., the functor that maps the objects and morphisms of $\mathrm{Mod}\,T$ to themselves). This functor is full and faithful, of course, and the statement that we made above means that it is also essentially surjective on objects. Hence:

The embedding functor $\iota\colon \mathrm{Mod}\,T \to \mathrm{Mod}'\,T$ is an equivalence of categories.

Thus, the category of models of T is the same, from the category theoretical point of view, as the *category of logical functors* from $\mathbb{T}$ to Set.

From now on we will denote this latter category as $\mathrm{Mod}\,T$ *(rather than* $\mathrm{Mod}'\,T$*) and, when saying that* M *is a model of* T*, we will mean that* $M\colon \mathbb{T} \to \mathrm{Set}$ *is a logical functor (and will denote this as* $M \vDash T$*).*

Remark. The fact that the category theoretical characterization of the standard functors is only up to isomorphism, should not come as a surprise. In order to be standard, a logical functor M must satisfy certain *set theoretical* properties that cannot be formulated in category theoretical terms. For example, for all L-formulas $\varphi(x)$ with $x : d$, M should map the object $[\varphi(x)]$ of $\mathbb{T}$ to a *subset* $\varphi(M)$ of $d(M)$ (remember that $d(M)$ is the image under M of the object $[x = x]$). Some of these $\varphi(M)$s should be *proper* subsets of $d(M)$. However, for objects X, Y of Set, there is no way to express, in the language of category theory, the fact that X is a proper subset of Y, simply because this statement is *not* preserved by equivalence functors. The most that we can say, using the concepts of Set-morphism (i.e. function) and morphism composition, is that there exists a *monomorphism* from X into Y (a morphism $f\colon X \to Y$ in a category is called a monomorphism iff $f \circ g = f \circ h$ implies $g = h$, for any two morphisms $f, g\colon Z \to X$; a morphism in Set is a monomorphism iff it is a one-to-one function).

Extensions. We reached a position in which we can produce a precise mathematical definition for the notion of extension of a theory.

Definition 4.1. A theory T_1 is called an *extension* of T, iff there is a logical functor $I\colon \mathbb{T} \to \mathbb{T}_1$. The functor I is called an extension functor or, simply, an extension.

This notion is certainly broad enough to incorporate the one that we outlined roughly at the end of Section 3.

The obvious example, in our context, is that of T^{eq} being an extension of T. The extension functor $I\colon \mathbb{T} \to \mathbb{T}^{\mathrm{eq}}$ is nothing but the *embedding* functor defined, simply, as $I([\varphi(\vec{x})]_T) = [\varphi(\vec{x})]_{T^{\mathrm{eq}}}$ for objects and $I([\chi(\vec{x},\vec{y})]_T) = [\chi(\vec{x},\vec{y})]_{T^{\mathrm{eq}}}$ for morphisms.

At this point, we have a very simple and straightforward mathematical definition of *significance* of an extension, a notion that was not clear how to define in the framework of model theory.

Definition 4.2. An extension $I\colon \mathbb{T} \to \mathbb{T}_1$ is called *significant* iff I is *not* an equivalence of categories.

In Section 3 we made the *intuitive* statement that T^{eq} is a significant extension of T. This can be now stated precisely, by saying that the embedding functor $I\colon \mathbb{T} \to \mathbb{T}^{\mathrm{eq}}$ is not an equivalence of categories. Well, this is true for *most* first order theories, but not for *all* of them.

Definition 4.3. We say that a theory T has *elimination of imaginaries* iff the embedding functor $I\colon \mathbb{T} \to \mathbb{T}^{\mathrm{eq}}$ is an equivalence of categories.

This notion has been defined, in model theoretical terms, by Poizat [19] (or [20], in English translation). His definition says, roughly, that each equivalence relation of T has a first order definable 'quotient function' (i.e., a function under which,

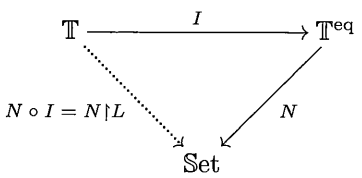
any two tuples have the same value iff they are equivalent). The explanation of this name rests on the fact that Shelah called the equivalence classes $\vec{a}/\varepsilon(M)$ "imaginary elements of the model M."

Theories with elimination of imaginaries do exist. In fact:

Claim 4.4. *For any first order theory T, the theory T^{eq} has elimination of imaginaries. In other words, the embedding functor $I\colon \mathbb{T}^{eq} \to (\mathbb{T}^{eq})^{eq}$ is an equivalence of categories.*

This easily seen fact was certainly known to Shelah, but it seems that Poizat was the one who first stated it explicitly.

How can we define, in category theoretical terms, the notion of *tight* extension? To do this, we have to discuss first the notion of *restriction* of a model. We start with the observation that each model N of T^{eq} (i.e., logical functor $N\colon \mathbb{T}^{eq} \to \mathbb{S}et$) yields one of T by functor composition, namely $N \circ I\colon \mathbb{T} \to \mathbb{S}et$, which is easily seen to be the logical functor that is identified with the restriction $N\restriction L \vDash T$.



More generally, for *any* extension $I\colon \mathbb{T} \to \mathbb{T}_1$, we have a function $N \mapsto N \circ I$ from the objects of $\mathrm{Mod}\, T_1$ to those of $\mathrm{Mod}\, T$. Moreover, this function can be extended to a *functor* $(-) \circ I\colon \mathrm{Mod}\, T_1 \to \mathrm{Mod}\, T$, as it follows by basic category theoretical facts (for the uninitiated, it is an easy but rewarding exercise to work out the definition of $\eta \circ I\colon M \circ I \to N \circ I$ for any natural transformation $\eta\colon M \to N$). It is this functor, denoted as $I^* = (-) \circ I\colon \mathrm{Mod}\, T_1 \to \mathrm{Mod}\, T$ by Makkai and Reyes, that we see as the *generalized* restriction functor *induced* by I.

Having this concept at hand, the following definition is straightforward.

Definition 4.5. An extension $I\colon \mathbb{T} \to \mathbb{T}_1$ is called *tight* iff the restriction functor $I^*\colon \mathrm{Mod}\, T_1 \to \mathrm{Mod}\, T$ is an equivalence of categories.

We can now formulate rigorously the questions raised at the end of Section 3. Does T^{eq} have a tight and significant extension? If not, is T^{eq} a largest tight extension of T? In other words, is T^{eq} an extension of any other tight extension of T? In the next section, we describe the answer given by Makkai and Reyes.

5. The pretopos completion

As we mentioned in the introduction, Makkai and Reyes discovered that the categories called by Grothendieck *pretoposes* form a subclass of the logical categories and enjoy a remarkable property. They also defined a category-theoretical operation that attaches to every logical category a closely related pretopos called its *pretopos completion*. Makkai's remark in the fall of 1980, was that this operation generalizes the construction of T^{eq}. Indeed, if T is a first order complete theory with infinite models (as are the ones considered by Shelah), then the pretopos completion of the category $\mathbb{T}$ turns out to be (equivalent to) the category $\mathbb{T}^{eq}$. These findings will be presented, in some detail, in this section.

As a first step of our exposition, let us ask when is a Boolean logical category equivalent to one of the form $\mathbb{T}^{\mathrm{eq}}$, for a first order theory T. Of course, Makkai and Reyes did not ask themselves this question because they were not aware of Shelah's construction at the time. Yet, we will follow their guiding principle namely, to seek category-theoretical properties, shared by the category $\mathbb{S}\mathrm{et}$ (which is, itself, a pretopos), that will supply the desired characterization. The direction of inquiry should be clear. The categories that we are looking for, should be closed also under category-theoretical operations that generalize the set theoretical operations of taking quotients of equivalence relations. We are faced again with the same problem: how can we describe complex set theoretical constructions, using only category-theoretical concepts? At first sight, the language of category theory seems more restricted than the one of set theory and, in fact, *it is* more restricted (remember that we cannot express even a simple fact such as an object of $\mathbb{S}\mathrm{et}$ being a subset of another). *One can*, nevertheless, *capture a great deal of the richness of the universe of sets*, using this limited language. How to accomplish this task, is a challenging problem well worth pursuing, as it turns out. If this project is pushed far enough, one reaches the notion of *elementary topos*, the subject of an active and vibrant research field. The category $\mathbb{S}\mathrm{et}$ is a topos, but there are many other interesting instances and it would be a mistake to think of topos theory as generalized set theory only. Actually, the origins of this theory are in the work of Grothendieck (to whom the name '*topos*' is due) in algebraic geometry. To get an idea of the broadness of the subject, one could consult, for example, [9] (which ends with an extensive bibliography; the literature of the field is so vast, that one could not provide a single reference without doing injustice to many others).

The properties needed to define pretoposes fall way short of the axioms for an elementary topos. Even less is required to characterize the categories equivalent to some $\mathbb{T}^{\mathrm{eq}}$. We can define category theoretical notions of *equivalence relation* over an object and *quotient* of an equivalence relation, that generalize the set theoretical concepts with the same name. Once this done, we can state:

Claim 5.1. *A small category $\mathbb{C}$ is equivalent to one of the form $\mathbb{T}^{\mathrm{eq}}$ for some first order theory T iff $\mathbb{C}$ is a Boolean logical category which is closed under the operation of taking quotients of equivalence relations.*

Notice that the "only if" direction namely, the fact that $\mathbb{T}^{\mathrm{eq}}$ is closed under quotients of equivalence relations, is established by Claim 4.4.

In order to define the notion of pretopos, we need two more concepts. The first of these is basic and well known.

An object S of a category $\mathbb{C}$ is called *initial*, if for any object X of $\mathbb{C}$, there is precisely one $\mathbb{C}$-morphism $S \to X$. The empty set $\varnothing$ is the unique initial object of the category $\mathbb{S}\mathrm{et}$. An object $[\varphi(\vec{x})]$ of the logical category $\mathbb{T}$ is initial iff $\varphi(\vec{x})$ is a formula which is a contradiction in the theory T (i.e. $T \vDash \forall\vec{x}\neg\varphi(\vec{x})$).

We shall say that two $\mathbb{C}$-objects X and Y *coexist disjointly in* $\mathbb{C}$ iff there exist in $\mathbb{C}$ a diagram $X \xrightarrow{g} A \xleftarrow{h} Y$ where g, h are monomorphisms, which has a pullback

$$\begin{array}{ccc} S & \xrightarrow{q} & Y \\ {\scriptstyle r}\downarrow & & \downarrow{\scriptstyle h} \\ X & \xrightarrow[g]{} & A \end{array}$$

with S an initial object of $\mathbb{C}$. In the category Set, this condition means that g and h are bijections whose images $g(X), h(X)$ are *disjoint* subsets of A. It follows easily than any two objects of Set coexist disjointly.

Definition 5.2. A *pretopos* is a logical category $\mathbb{C}$ which is closed under the operation of taking quotients of equivalence relations and in which any two objects coexist disjointly. A *Boolean pretopos* is a pretopos which is a Boolean logical category.

Remark. This definition is different from, but equivalent to, Grothendieck's original or the Makkai and Reyes definition.

By Claim 5.1, the categories of the form $\mathbb{T}^{eq}$ have to satisfy only the first of the two conditions from the definition of a pretopos. Yet, as we already hinted in the first paragraph of this section, in the cases that are of interest for the present paper, the category $\mathbb{T}^{eq}$ *is* a Boolean pretopos. We are now going to explain this.

If d is a sort symbol of a first order theory T, we say that d *is nonempty* if the set $d(M)$ is nonempty for every model $M \models T$ (in other words, $T \models \exists x(x = x)$, where $x : d$). In unisorted logic, we assume tacitly that the unique sort is nonempty, but in the many sorted context, we are not bound by any nonemptiness assumptions. Let us say that T is a *proper* theory if all its sorts are nonempty and, for some sort symbol d_0, the set $d_0(M)$ has at least two elements for every model $M \models T$ (i.e., $T \models \exists u \exists v(u \neq v)$, where $u, v : d_0$). The theories studied by Shelah are certainly proper, as they are unisorted and have only infinite models. An important simple observation is that *if T is proper, then so is T^{eq}*.

Theorem 5.3. *If T is a proper first order theory, then $\mathbb{T}^{eq}$ is a Boolean pretopos.*

PROOF (SKETCH). We know that $\mathbb{T}^{eq}$ is a Boolean logical category closed under quotients of equivalence relations, so, all we have to show is that any two objects coexist disjointly. Let $X = [\varphi(\vec{x})]$, $Y = [\psi(\vec{y})]$ be two objects in $\mathbb{T}^{eq}$ with $\vec{x} : \vec{d^1}$, $\vec{y} : \vec{d^2}$. As we know, by Claim 4.4, that $\mathbb{T}^{eq}$ is equivalent to $(\mathbb{T}^{eq})^{eq}$, it will be enough to show that X and Y coexist disjointly in this latter category. Let $\vec{d}$ be the sequence $(d_0, d_0, \vec{d^1}, \vec{d^2})$ of sort symbols (where d_0 is the sort that shows that T is proper). With $u, v : d_0$ two distinct variables of the sort d_0, let us denote, for brevity, $\vec{z} = (u, v, \vec{x}, \vec{y}) : \vec{d}$ and take $\vec{z'} = (u', v', \vec{x'}, \vec{y'}) : \vec{d}$ to be a sequence of similar variables *distinct* from their nonprimed counterparts. Let $\varepsilon(\vec{z}, \vec{z'})$ be the formula

$$(u = v \wedge u' = v' \wedge \vec{x} = \vec{x'}) \vee (u \neq v \wedge u' \neq v' \wedge \vec{y} = \vec{y'}).$$

This is an equivalence formula of T^{eq}, hence the theory $(T^{eq})^{eq}$ has a new sort symbol d_ε and a function symbol $f_\varepsilon : \vec{d} \to d_\varepsilon$ which is the quotient map of ε. Remember that we are seeking an object A and monomorphisms $X \xrightarrow{g} A \xleftarrow{h} Y$, *in* $(T^{eq})^{eq}$, showing the disjoint coexistence of $X = [\varphi(\vec{x})]$ and $Y = [\psi(\vec{y})]$. We take $A = [w = w]$, where $w : d_\varepsilon$ (in other words, A is the extension of the sort d_ε), while g, h will be the maps

$$[\varphi(\vec{x})] \xrightarrow{[\gamma]} [w = w] \xleftarrow{[\chi]} [\psi(\vec{y})]$$

where γ and χ are the formulas

$$\varphi(\vec{x}) \wedge \exists u \exists v \exists \vec{y}(u = v \wedge f_\varepsilon(\vec{z}) = w) \text{ and } \psi(\vec{y}) \wedge \exists u \exists v \exists \vec{x}(u \neq v \wedge f_\varepsilon(\vec{z}) = w),$$

respectively.

We let the reader verify that these data show the disjoint coexistence of X and Y. □

Remarks. (1) Notice that the fact that our theories are in the *full* classical logic is crucial here, as the negation and the validity of the classical tautology $u = v \vee u \neq v$ are used in an essential way (by the *full* logic, we mean the one that is allowed to employ all connectives, including negation, and quantifiers, including the universal one; as a result, theories in the full logic are, when regarded as categories, *Boolean* logical categories).

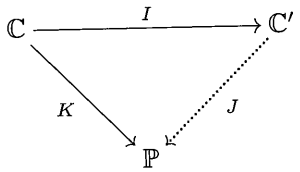
(2) The statement of this theorem is not necessarily true if the theory is not *proper*. Indeed, if T is unisorted and all its models have just one element (i.e. $T \models \exists! x(x = x)$), then a moment of thought will reveal that T is, actually, a theory in *propositional* logic. In this case, $T = T^{\mathrm{eq}}$ and $\mathbb{T}^{\mathrm{eq}}$ is *not* a pretopos.

The following theorem is the independent discovery by Makkai and Reyes of a category theoretical generalization of Shelah's construction of Section 2.

Theorem 5.4 ([15]). *Every logical category* $\mathbb{C}$ *has a tight extension* $I\colon \mathbb{C} \to \mathbb{C}'$ *with the following properties*:

(1) $\mathbb{C}'$ *is a pretopos. If* $\mathbb{C}$ *is Boolean, so is* $\mathbb{C}'$.

(2) *If* $K\colon \mathbb{C} \to \mathbb{P}$ *is a logical functor from* $\mathbb{C}$ *into a pretopos* $\mathbb{P}$, *then there is a logical functor* $J\colon \mathbb{C}' \to \mathbb{P}$ *such that* $K = J \circ I$.



The category $\mathbb{C}'$ or, rather, the extension $I\colon \mathbb{C} \to \mathbb{C}'$ is called, by Makkai and Reyes, *the pretopos completion of* $\mathbb{C}$.

Let us remark that the functor J, of part 2 of Theorem 5.4, is just an object of the functor category $\mathbb{P}^{\mathbb{C}'}$, hence we do not expect it to be *uniquely* defined. Indeed, it is unique *up to an isomorphism* in $\mathbb{P}^{\mathbb{C}'}$.

If $\mathbb{C}$ is a category of the form $\mathbb{T}$ for a *proper* first order theory T, then Shelah's construction of Section 2 furnishes a proof of this theorem. However, Theorem 5.4 is more general, as it covers all logical categories, even not stemming from a proper theory, even not Boolean. Makkai and Reyes' independently found proof is similar to Shelah's construction, the main difference being that it involves an additional step of adding disjoint sums (which is a strong form of disjoint coexistence, used in their definition of pretopos).

Actually, all the results of [15] described in this paper are stated and proven in the general context of *logical* (not necessarily *Boolean*) categories. This is an appropriate place to point out the logical significance of this more general context. The realization that "logical functors are enough", meaning that these functors preserve not only the structure of logical categories, but also the additional structure that occurs in Boolean logical categories and, even, the pretopos structure, focussed the attention of Makkai and Reyes to *coherent logic*, which is the *restricted* fragment of the the *classical* first order logic based only on conjunctions, disjunctions and

existential quantification. Coherent logic is carefully described in [15] in an appropriate formal context (known to logicians as *sequent calculus*), and is shown to be related to the concept of logical category in the same way, in which classical first order logic is related to Boolean logical category.

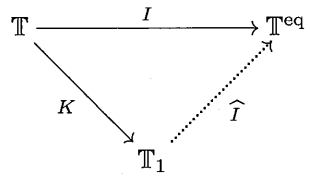
The important result that we present next is the remarkable property of pretoposes that we mentioned in the introduction and in the opening paragraph of the present section. Notice, again, that it refers to arbitrary, not necessarily Boolean, pretoposes.

Theorem 5.5 ([15]). *If $\mathbb{P}$ is a pretopos and $I\colon \mathbb{P} \to \mathbb{P}'$ is a* tight *extension then I is an equivalence of categories.*

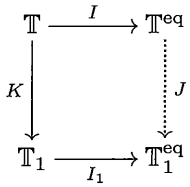
When we take a tight extensions of a first order theory, for example the embedding $\mathbb{T} \to \mathbb{T}^{eq}$, we do not add, in fact, any new models. *However* we do add, *in general*, new *concepts* such as equivalence classes, which now became ordinary *elements* of the models. The theorem that we just presented states that this is not the case if $\mathbb{T}$ was a pretopos, to begin with. Therefore the statement of Theorem 5.5 is called, in [15], **the conceptual completeness of pretoposes**.

The conceptual completeness of *Boolean* pretoposes implies, together with Theorem 5.4, the answer to the question asked at the end of Sections 3 and 4.

Corollary 5.6. *If $\mathbb{T}$ is a Boolean logical category, the embedding $I\colon \mathbb{T} \to \mathbb{T}^{eq}$ is the* largest *tight extension of $\mathbb{T}$. In other words, for any other tight extension $K\colon \mathbb{T} \to \mathbb{T}_1$ there is a logical functor $\widehat{I}\colon \mathbb{T}_1 \to \mathbb{T}^{eq}$ such that $\widehat{I} \circ K = I$.*



PROOF (SKETCH). Let $I_1\colon \mathbb{T}_1 \to \mathbb{T}_1^{eq}$ be the embedding of $\mathbb{T}_1$ into its pretopos completion. Then the composition $I_1 \circ K\colon \mathbb{T} \to \mathbb{T}_1$ is a logical functor into a Boolean pretopos, hence, by Theorem 5.4, there exists a logical functor $J\colon \mathbb{T}^{eq} \to \mathbb{T}_1^{eq}$ such that $J \circ I = I_1 \circ K$.



$I_1 \circ K$ is a *tight* extension, as the composition of two such. Therefore, so is $J \circ I$ and as I is tight as well, we conclude that J *is also tight* (the induced restriction functors satisfy $(J \circ I)^* = I^* \circ J^*$ and is an easy exercise to see that if $I^* \circ J^*$ and I^* are equivalence functors, then so is J^*). But then it follows, by Theorem 5.5, that J is an equivalence of categories and so, $\mathbb{T}^{eq}$ and $\mathbb{T}_1^{eq}$ *are the same.* Therefore, I_1 is *essentially* a logical functor from $\mathbb{T}_1$ into $\mathbb{T}^{eq}$! This means that one can modify (using routine techniques) I_1 into an isomorphic logical functor $\widehat{I}\colon \mathbb{T}_1 \to \mathbb{T}^{eq}$, as desired. □

Remark. Again, one can show that the functor $\widehat{I}$ is unique, up to isomorphism, having the specified property.

6. Concluding remarks

(1) Shelah succeeded to overcome certain difficulties in the development of his classification theory by using an auxiliary construction. He *invented* the extension T^{eq} of a first order theory T.

Makkai and Reyes explored a new framework unveiling the structure behind the categorical approach to logic and model theory. They *discovered* the pretopos completion.

The two contributions, of Shelah on one hand and of Makkai–Reyes on the other, complement each other. The pretopos completion is a concept that *had* to be defined. It lays there as part of the orderly world in which the categories associated with theories and their models live. Once discovered, it is only natural to ask for a context in which this new notion is useful. Shelah's work supplies an instant answer, a gratifying one, to this question.

(2) If $\mathbb{T}$ is a logical category, let $\operatorname{Mod}\mathbb{T}$ be the category of *logical* functors $\mathbb{T} \to \mathrm{Set}$ (in other words, $\operatorname{Mod}\mathbb{T}$ is the same as the category $\operatorname{Mod}T$ of the models of the coherent theory represented by $\mathbb{T}$). The conceptual completeness Theorem 5.5 says, roughly speaking, that if $\mathbb{T}, \mathbb{T}'$ are pretoposes such that $\operatorname{Mod}\mathbb{T}, \operatorname{Mod}\mathbb{T}'$ "are the same," then so are $\mathbb{T}$ and $\mathbb{T}'$. It seems, therefore, that the category $\operatorname{Mod}\mathbb{T}$ *determines* the pretopos $\mathbb{T}$. Or, using a more "model-theory friendly" way of speech, it seems that for a first order theory T, the category of models $\operatorname{Mod}T$ determines the theory T^{eq}. *Is this really so*?

The answer to this question turns out to be positive, *provided* that we phrase it carefully, as done by Makkai in [13] (for a shorter preliminary version, see [11]). When saying that the pretoposes $\mathbb{T}$ and $\mathbb{T}'$ "are the same," we just mean that they are equivalent categories. However, the meaning of the categories of models "being the same," in the hypothesis of Theorem 5.5, is that they are equivalent by a functor $I^*\colon \operatorname{Mod}\mathbb{T}' \to \operatorname{Mod}\mathbb{T}$ *induced* by a logical functor $I\colon \mathbb{T} \to \mathbb{T}'$. How can we recognize which functors between categories of models are of this, induced, kind?

Makkai points out that the *concrete* category $\operatorname{Mod}\mathbb{T}$ carries some extra structure related to the fact that the class of models of a given first order theory is closed under the ultraproduct operation. He defines an abstract notion of *ultracategory*, which is a category endowed with an extra structure of the kind just mentioned, such that the categories of the form $\operatorname{Mod}T$ are ultracategories (and so is Set). A functor between ultracategories that preserves this extra structure is called by Makkai an *ultrafunctor*. He shows that the induced functors I^* are ultrafunctors and that they are, up to isomorphism, the only ultrafunctors $\operatorname{Mod}\mathbb{T}' \to \operatorname{Mod}\mathbb{T}$. Moreover, he shows that the pretopos $\mathbb{T}$ is determined by the *ultracategory* $\operatorname{Mod}\mathbb{T}$.

Actually, [13] describes a concrete way of recovering $\mathbb{T}$ from $\operatorname{Mod}\mathbb{T}$. Let Pt be the category whose objects are the pretoposes and morphisms the logical functors between them, and let UC be the category whose objects are the ultracategories and whose morphisms are the ultrafunctors. Then Mod can be construed as a *contravariant* functor from the category Pt of pretoposes to the category UC of ultracategories (a contravariant functor is one that reverses the direction of morphisms, as $I \mapsto I^*$ does). Makkai constructs a *dual* contravariant functor F from UC to Pt such that for any pretopos $\mathbb{T}$, $F\operatorname{Mod}\mathbb{T}$ is a pretopos that is equivalent

to $\mathbb{T}$ itself. This result implies Theorem 5.5 and is called by Makkai the *strong conceptual completeness of pretoposes.*

Makkai's work that we just described has had several applications, and I will describe two of them.

The first application is due to Makkai himself and concerns conceptual completeness for *Boolean* pretoposes. For $\mathbb{T}$ a Boolean pretopos, let $\mathrm{Mod}^{\mathrm{iso}}\,\mathbb{T}$ be the category whose objects are the models of $\mathbb{T}$ (i.e., logical functors $M\colon \mathbb{T} \to \mathrm{Set}$) and morphisms are only the *isomorphisms* between models. In other words, $\mathrm{Mod}^{\mathrm{iso}}\,\mathbb{T}$ is a subcategory of $\mathrm{Mod}\,\mathbb{T}$ with the same objects but considerably fewer morphisms. Using the methods of [13], Makkai proves in [12] that a Boolean pretopos $\mathbb{T}$ can be recovered from the smaller category $\mathrm{Mod}^{\mathrm{iso}}\,\mathbb{T}$ as well. This is a strong conceptual completeness result for Boolean pretoposes which, as Makkai points out, is closely related to an earlier, unpublished, result of Gaifman, a purely model-theoretical definability theorem.

The second application is due to Zawadowski. Given logical categories $\mathbb{T}, \mathbb{T}_1$ and a logical functor $I\colon \mathbb{T} \to \mathbb{T}_1$, let $\mathrm{Mod}_I\,\mathbb{T}$ be the *full* subcategory of $\mathrm{Mod}\,\mathbb{T}$ whose objects are those models $M\colon \mathbb{T} \to \mathrm{Set}$ of $\mathbb{T}$ that can be *extended* to a model M_1 of $\mathbb{T}_1$ (meaning that $M = M_1 \circ I\colon \mathbb{T} \to \mathrm{Set}$). In the simplest case, in which the two categories represent theories T, T_1 such that T_1 extends T in the ordinary sense (meaning that the language and the set of theorems of the former includes those of the latter), the model theorist will recognize that the set of objects of $\mathrm{Mod}_I\,\mathbb{T}$ is what is called a PC_Δ-class of structures. In general, the category $\mathrm{Mod}_I\,\mathbb{T}$ has fewer objects than the category $\mathrm{Mod}\,\mathbb{T}$ of *all* the models of $\mathbb{T}$. And yet, Zawadowski uses [13] and proves that in certain cases, $\mathbb{T}$ can be recovered from this proper subcategory of $\mathrm{Mod}\,\mathbb{T}$. Indeed, the main result of [27], implies that *if* $\mathbb{T}, \mathbb{T}_1$ *are* pretoposes *and* $I\colon \mathbb{T} \to \mathbb{T}_1$ *is* conservative, *then* $\mathbb{T}$ *can be recovered from* $\mathrm{Mod}_I\,\mathbb{T}$. This beautiful result, which can be called, aptly, *strong conceptual completeness for* PC_Δ*-classes of pretoposes*, generalizes Makkai's result. When translated into the language of pure category theory, the theorem of Zawadowski is called *the descent theorem for pretoposes* and it affirms a conjecture stated earlier by A. Pitts.

After seing Zawadowski's work, Makkai proved, in [14], a descent theorem for *Boolean* pretoposes as well. For a simpler, heavily model-theoretical proof of both, Zawadowski's and Makkai's, descent theorems, see Ballard and Boshuck's [2].

All the results that we mentioned in this discussion deal with classical logic (either in its full or in its coherent version) and were proven using a mixture of category-theoretical and model-theoretical methods. Andrew Pitts succeeded to prove, in [17], by purely category theoretical methods, a stronger, *constructive* version of conceptual completeness of pretoposes. To be more specific: while the original proof of Makkai–Reyes uses freely the axioms of set theory, as is common in modern mathematics, Pitts' arguments use only the *restricted* set of axioms for an elementary topos with a natural number object (the assumption of having such an object is the topos theoretical version of the axiom of infinity); in particular, this weaker set of axioms does not include an analogue of the axiom of choice, hence the adjective 'constructive' that topos theorists use to design such results.

In a subsequent paper [18], written for a logic inclined audience, Pitts used similar methods to prove a conceptual completeness result for *intuitionistic* first order logic. He states his result in both logical and category-theoretical terms, the

latter being what we call conceptual completeness for *Heyting* pretoposes. The validity of a descent theorem for this kind of pretoposes is still an open question.

(3) We mentioned, at the end of Section 3, the notion of *interpretation* of a theory into another, familiar in mathematical logic. As promised there, we are going now to elaborate on this concept, its generalizations and its connection to the notion of extension. This will be our final, and longest, comment and will show that the conceptual completeness of Boolean pretoposes has some relevance to the question whether we have the most general definition.

Given first order languages L, L' and an L'-theory T', an *interpretation of* L *into* T' is a function $\varphi \mapsto \varphi'$ with certain properties that allow us to extract an L-structure M from any model $M' \vDash T'$.

We are now going to spell out the conditions that the function $\varphi \mapsto \varphi'$ must satisfy in order to be an interpretation of a kind that we call *simple*. This is a non customary name and we use it to distinguish between simple interpretations and the more general kind that we will define later.

We assume first that L, L' are *unisorted* languages. If φ is the L-formula $x = x$ then φ' will be an L'-formula $\delta(\vec{x}')$, with $\vec{x}'$ a k-tuple, that plays a very special role: the elements of the L-structure M extracted from $M' \vDash T'$ will be the k-tuples of M' that satisfy $\delta(\vec{x}')$. If $\varphi \equiv \varphi(x_1, \ldots, x_n)$ (where we use '$\equiv$', rather than the equality symbol '$=$', to denote identity of formulas) then $\varphi' \equiv \varphi'(\vec{x}'_1, \ldots, \vec{x}'_n)$ with $\vec{x}'_1, \ldots, \vec{x}'_n$ *distinct* k-tuples and we require also that $T' \vDash \forall \vec{x}'_1 \ldots \forall \vec{x}'_n (\varphi'(\vec{x}'_1, \ldots, \vec{x}'_n) \to \delta(\vec{x}'_i))$ for all $1 \leqslant i \leqslant n$. If $\varphi \equiv R(x_1, \ldots, x_n)$, with R an n-ary relation symbol of L, then the formula $\varphi' \equiv \rho(\vec{x}'_1, \ldots, \vec{x}'_n)$ will define the interpretation $R(M)$ of R in the model M extracted from M' as the set of those n-tuples of k-tuples of elements of M' that satisfy the formula ρ. A very important *simplicity* condition, which makes the interpretation *simple*, requires that if φ is the formula $x = y$ then φ' will be the formula $\vec{x}' = \vec{y}'$ (which is short for $x'_1 = y'_1 \wedge \cdots \wedge x'_k = y'_k$). This condition insures that the equality relation in M will be the *standard* one, i.e. the equality of k-tuples. Next, if $\varphi \equiv f(x_1, \ldots, x_n) = y$, where f is an n-ary function symbol of L, then $\varphi' \equiv \varphi'(\vec{x}'_1, \ldots, \vec{x}'_n, \vec{y}')$ and we require T' to imply that this formula defines $\vec{y}'$ as a function of $(\vec{x}'_1, \ldots, \vec{x}'_n)$. The formula φ' can then be used to define the interpretation $f(M)$ of f in M. The interpretations of individual constants are defined in a similar way. Thus, the values of $\varphi \mapsto \varphi'$ for atomic formulas φ, allow us to extract M from M'. The values of the interpretation for more complex formulas is uniquely determined by additional, very natural, recursive requirements for the function $\varphi \mapsto \varphi'$. It must preserve connectives (e.g., we must have $(\varphi \wedge \psi)' \equiv \varphi' \wedge \psi'$) and translate quantifiers $\forall x, \exists x$ to their versions *relativised* to δ (i.e., $\forall \vec{x}'(\delta(\vec{x})' \to \cdots$ and $\exists \vec{x}'(\delta(\vec{x}') \wedge \cdots)$.

This concludes the definition of a *simple interpretation* of a language L into an L'-theory T', for *unisorted* languages. The extension to the *multisorted* context involves only notational complications, because in this case, we have a formula $x = x$ with $x : d$, for every sort symbol d and this formula is mapped by the simple interpretation to an L'-formula $\delta_d(\vec{x}')$, where the length and the sorting of the sequence of L'-variables $\vec{x}'$ *depend* on d. We skip the obvious details of the extended definition.

If T, T' are theories in the (possibly, multisorted) languages L and L', then a simple interpretation $\varphi \mapsto \varphi'$ of T into T' is a simple interpretation of L into T'

such that for any L-sentence φ, $T \vDash \varphi$ implies that $T' \vDash \varphi'$. It follows, in particular, that for L-formulas φ, ψ, if $\varphi \sim_T \psi$ then $\varphi' \sim_{T'} \psi'$.

Simple interpretations and extensions are essentially the same. Indeed, T has a simple interpretation in T' iff $\mathbb{T}'$ is an extension of $\mathbb{T}$. To be more precise:

Proposition 6.1. *Each simple interpretation $\varphi \mapsto \varphi'$ of T into T' induces an extension functor $I\colon \mathbb{T} \to \mathbb{T}'$ such that I maps any $\mathbb{T}$-morphism $[\varphi]_T \xrightarrow{[\chi]_T} [\psi]_T$ to the $\mathbb{T}'$-morphism $[\varphi']_{T'} \xrightarrow{[\chi']_{T'}} [\psi']_{T'}$. Conversely, every extension functor from $\mathbb{T}$ to $\mathbb{T}'$ is induced by a simple interpretation.*

Remark. Actually, the same extension functor is induced by *several* simple interpretations because, when we construct a simple interpretation inducing a given functor, the value of $\varphi \mapsto \varphi'$ for simple atomic formulas is imposed on us only up to $\sim_{T'}$-equivalence.

Consider a first order theory T in a language L. If T^{eq} is a significant extension of T, as is mostly the case, then $\mathbb{T}$ is *not* an extension of $\mathbb{T}^{eq}$. This means that T^{eq} has no simple interpretation in T. And yet, every model $M \vDash T$ 'carries' with it its canonical extension M^{eq}. The construction of this canonical extension from M is done, actually, via an *interpretation* in a wider sense in which the elements of the 'extracted' model M^{eq} are interpreted, not as k-tuples in M but, rather, as *equivalence classes* of elements of M. This suggests an obvious generalization of the notion of *simple* interpretation, which we are going to describe.

Again, in order to avoid notational complications, we assume first that the languages are unisorted. An *interpretation* $\varphi \mapsto \varphi'$ of a language L into an L'-theory T' is defined like a simple interpretation with one major modification: the simplicity condition (which said that the interpretation of equality is the standard equality of tuples) is replaced by the following:

If φ is the formula $x = y$ then φ' is a formula $\varepsilon(\vec{x}', \vec{y}')$ such that T' implies that ε defines an equivalence relation on the set of k-tuples satisfying $\delta(\vec{x}')$.

This means that each element of the extracted model M is interpreted as an ε-equivalence class of tuples. As a result, we have to impose additional requirements: T' must imply that ε is a congruence relation with respect to the relation defined by the formula ρ describing the interpretation of a relation symbol R. By this we mean that T' implies that whenever $\varepsilon(\vec{x}'_i, \vec{y}'_i)$ for all $1 \leqslant i \leqslant n$, we also have $\rho(\vec{x}'_1, \ldots, \vec{x}'_n) \leftrightarrow \rho(\vec{y}'_1, \ldots, \vec{y}'_n)$. A similar condition is required for the formula φ describing the interpretation of a function symbol f. For this formula we also require a 'functionality' condition, namely, that T' implies that whenever $\varphi(\vec{x}'_1, \ldots, \vec{x}'_n, \vec{y}')$ and $\varphi(\vec{x}'_1, \ldots, \vec{x}'_n, \vec{z}')$ both hold than we also have $\varepsilon(\vec{y}', \vec{z}')$. With these conditions fulfilled , we obviously can, given any model $M' \vDash T'$, build an L-structure whose elements are ε-equivalence classes of k-tuples of M'. The definition of the notion of interpretation is completed by imposing on complex formulas the same recursive requirements as for simple interpretations. This implies that for any L-formula φ, T' will imply that ε is a congruence relation with respect to the relation defined by φ'.

This definition of interpretation, in the unisorted context, is precisely the one given by Wilfrid Hodges in [7], which contains a thorough treatment of interpretations and their uses in logic, especially in model theory (see also [8], where Hodges graciously traces this definition to Mal'cev, [16, Chapter 15]).

It is a routine exercise to generalize this definition to the multisorted context. Once this done, we can make a precise statement to the effect that T^{eq} *has an interpretation in* T, in fact a very simply described one. Let us call it the *canonical interpretation of* T^{eq} *in* T.

Consider, next, an arbitrary interpretation of an L-theory T in an L'-theory T'. It allows, once given a model $M' \models T'$, to extract a model M of T whose elements are equivalence classes of relations definable in M'. But these equivalence classes are just *elements* of the canonical extension M'^{eq}! This means that the given interpretation of T in T' can be represented as a *simple* interpretation of T in T'^{eq} followed by the *canonical interpretation* of T'^{eq} in T'. Thus, interpretations in T' are essentially the same as *simple* interpretations of T in T'^{eq}. This gives a compact definition of the generalized version of interpretability.

Definition 6.2. A theory T is interpretable in another theory T' iff there exists a logical functor $F\colon \mathbb{T} \to \mathbb{T}'^{\mathrm{eq}}$. Such a functor F will be called an interpretation of T in T'.

Remarks. (1) Of course, every theory T is interpretable in itself via the obvious embedding $I\colon \mathbb{T} \to \mathbb{T}^{\mathrm{eq}}$. We will refer to I as the *canonical* interpretation of T into itself.

(2) If we have interpretations $F\colon \mathbb{T} \to \mathbb{T}'^{\mathrm{eq}}$ of T in T' and $G\colon \mathbb{T}' \to \mathbb{T}''^{\mathrm{eq}}$ of T' in T'', then we would like to define their composition $G * F$ as an interpretation of T in T''. This cannot be done as direct functor composition, because the codomain of F does not equal the domain of G. However, as the canonical interpretation $I'\colon \mathbb{T}' \to \mathbb{T}'^{\mathrm{eq}}$ is the pretopos completion of $\mathbb{T}'$, the functor G can be "lifted," by Theorem 5.4, to a functor $K\colon \mathbb{T}'^{\mathrm{eq}} \to \mathbb{T}''^{\mathrm{eq}}$ (which is unique up to isomorphism) such that $G = K \circ I'$. We then define $G * F = K \circ F$.

We now come to the main point of this comment. Had we reached the *most general* definition of the interpretation concept? Or, maybe, there are some more imaginative ways in which L-structures can be extracted from model of a given theory T'? The conceptual completeness of Boolean pretoposes seems to indicate that whatever means of interpretation one may come up with, they are already present in the pretopos completion of T', namely in T'^{eq}. We may hasten to think that *yes*, Definition 6.2 is most general reasonable one. *However*, here is an example that may cast some doubt on this belief. Let T^{EQ} be the *unisorted* version of T^{eq}, as envisaged by Shelah. Remember that we no longer have sorts d_ε but unary predicates P_ε and that we might have, in a model of T^{EQ} a set of elements which are outside the interpretations of all of these new unary predicates. Well, we would expect T^{EQ} to be interpretable in T and yet, there is no logical functor $F\colon \mathbb{T}^{\mathrm{EQ}} \to \mathbb{T}^{\mathrm{eq}}$. The reason for this is that $\mathbb{T}^{\mathrm{EQ}}$ has an 'all inclusive' object $[x = x]_{T^{\mathrm{EQ}}}$ into which any other object of $\mathbb{T}^{\mathrm{EQ}}$ can be embedded (by a monomorphism), while there is no such in the category $\mathbb{T}^{\mathrm{eq}}$.

We may correct this situation by a slight modification of Definition 6.2. For the purpose of the next definition, let me stress that a *finite* language is one whose set of extralogical symbols is finite.

Definition 6.3. An L-theory T is *locally interpretable* in another theory T' iff for every *finite* sublanguage $L_0 \subset L$, the restricted theory $T \restriction L_0$ is interpretable in T'.

It is not hard to see that T^{EQ} is locally interpretable in T. Is local interpretability a reasonable notion? One might argue that the possibility of extracting chunks of models of T from any given model of T', but not the model as a whole, is not really an interpretation. If we adopt this point of view, then we should probably have second thoughts and, contrary to our initial intuition, conclude that T^{EQ} is *not* interpretable in T, in any reasonable sense of the word.

On the other hand, the original use of the concept of interpretation in [26] was for the purpose of proving certain theories to be undecidable. The basic fact is the following:

If T is an undecidable theory which has a recursive *interpretation in T', then T' is undecidable as well.*

This statement remains true if we replace 'interpretation' by 'local interpretation.' From this point of view, local interpretability *is* a reasonable notion. However, this is not very significant, because in most applications of the basic fact, the theory T has a finite language anyhow.

Based on the conceptual completeness of pretoposes, I am ready to venture and propose the following (nonmathematical) conjecture:

For T a first order theory formulated in a finite *language, Definition* 6.2 *provides the most general reasonable notion of interpretability.*

Remark. The referee kindly suggested that I discuss, in this context, the intuitive concept of *bi-interpretability* of theories, as well. We usually say that two theories T and T' are *bi-interpretable* if each of them has an interpretation into the other, such that these two interpretations are, roughly speaking, inverses of each other. Bi-interpretable theories are considered to have the same logical strength. A non trivial example, which we mentioned already, is that of real closed fields vs. elementary geometry.

If we speak about finite-language theories and adopt the conjecture that I just proposed, it seems reasonable to reformulate the rough description above as the following precise mathematical definition (where we use the concepts and notations introduced in the remarks following Definition 6.2): T and T' are bi-interpretable iff we have interpretations $F\colon \mathbb{T} \to \mathbb{T}'^{eq}$ and $G\colon \mathbb{T}' \to \mathbb{T}^{eq}$ such that $G * F \simeq I$ (meaning that the two functors $G * F, I\colon \mathbb{T} \to \mathbb{T}^{eq}$ are isomorphic via a natural transformation) and $F * G \simeq I'$, where I and I' are the canonical interpretations of T and T' into themselves. Not surprisingly, it turns out (as one can see) that *T and T' are bi-interpretable iff the pretoposes $\mathbb{T}^{eq}$ and $\mathbb{T}'^{eq}$ are equivalent categories.*

References

1. J. T. Baldwin, *Fundamentals of stability theory*, Perspectives in Mathematical Logic, Springer-Verlag, Berlin, 1988.
2. D. Ballard and W. Boshuck, *Definability and descent*, J. Symbolic Logic **63** (1998), no. 2, 372–378.
3. G. Blanc, *Équivalence naturelle et formules logiques en théorie des catégories*, Arch. Math. Logik Grundlag. **19** (1978/79), no. 3-4, 131–137.
4. S. Eilenberg and S. Mac Lane, *General theory of natural equivalences*, Trans. Amer. Math. Soc. **58** (1945), 231–294.
5. P. Freyd, *Properties invariant within equivalence types of categories*, Algebra, Topology, and Category Theory: A Collection of Papers in Honor of Samuel Eilenberg (A. Heller and M. Tierney, eds.), Academic Press, New York, 1976, pp. 55–61.

6. V. Harnik and L. Harrington, *Fundamentals of forking*, Ann. Pure Appl. Logic **26** (1984), no. 3, 245–286.
7. W. Hodges, *Model theory*, Encyclopedia Math. Appl., vol. 42, Cambridge Univ. Press, Cambridge, 1993.
8. ———, *A shorter model theory*, Cambridge Univ. Press, Cambridge, 1997.
9. F. W. Lawvere, *Comments on the development of topos theory*, Development of Mathematics 1950–2000 (J.-P. Pier, ed.), Birkhäuser, Basel, 2000, pp. 715–734.
10. S. Mac Lane, *Categories for the working mathematician*, 2nd ed., Grad. Texts in Math., vol. 5, Springer, New York, 1998.
11. M. Makkai, *Stone duality for first order logic*, Proceedings of the Herbrand Symposium (Marseille, 1981) (J. Stern, ed.), Stud. Logic Found. Math., vol. 107, North-Holland, Amsterdam, 1982, pp. 217–232.
12. ———, *Ultraproducts and categorical logic*, Methods in Mathematical Logic (Caracas, 1983) (C. A. Di Prisco, ed.), Lecture Notes in Math., vol. 1130, Springer, Berlin, 1985, pp. 222–309.
13. ———, *Stone duality for first order logic*, Adv. in Math. **65** (1987), no. 2, 97–170.
14. ———, *Duality and definability in first order logic*, Mem. Amer. Math. Soc. **105** (1993), no. 503.
15. M. Makkai and G. E. Reyes, *First order categorical logic: Model-theoretical methods in the theory of topoi and related categories*, Lecture Notes in Math., vol. 611, Springer, Berlin, 1977.
16. A. I. Mal′cev, *The metamathematics of algebraic systems. Collected papers*: 1936–1967, Stud. Logic Found. Math., vol. 66, North-Holland, Amsterdam, 1971. Translated, edited, and provided with supplementary notes by B. F. Wells III.
17. A. M. Pitts, *Interpolation and conceptual completeness for pretoposes via category theory*, Mathematical logic and theoretical computer science (College Park, MD, 1984–1985) (D. W. Kueker, E. G. K. López-Escobar, and C. H. Smith, eds.), Lecture Notes in Pure and Appl. Math., vol. 106, Dekker, New York, 1987, pp. 301–327.
18. ———, *Conceptual completeness for first-order intuitionistic logic: an application of categorical logic*, Ann. Pure Appl. Logic **41** (1989), no. 1, 33–81.
19. B. Poizat, *Cours de théorie des modèles: Une introduction à la logique mathématique contemporaine*, Bruno Poizat, Lyon, 1985.
20. ———, *A course in model theory: An introduction to contemporary mathematical logic*, Universitext, Springer, New York, 2000. Translated by M. Klein and revised by the author.
21. S. Shelah, *Classification theory and the number of nonisomorphic models*, Stud. Logic Found. Math., vol. 92, North-Holland, Amsterdam, 1978.
22. ———, *Classification of first order theories which have a structure theorem*, Bull. Amer. Math. Soc. (N.S.) **12** (1985), no. 2, 227–232.
23. ———, *Classification theory and the number of nonisomorphic models*, 2nd ed., Stud. Logic Found. Math., vol. 92, North-Holland, Amsterdam, 1990.
24. ———, *Classification theory for elementary abstract classes*, Stud. Log. (Lond.), vol. 18, Mathematical Logic and Foundations, College Publications, London, 2009.
25. ———, *Classification theory for abstract elementary classes*. Vol. 2, Stud. Log. (Lond.), vol. 20, Mathematical Logic and Foundations, College Publications, London, 2009.
26. A. Tarski, *Undecidable theories*, Stud. Logic Found. Math., North-Holland, Amsterdam, 1953. In collaboration with A. Mostowski and R. M. Robinson.
27. M. W. Zawadowski, *Descent and duality*, Ann. Pure Appl. Logic **71** (1995), no. 2, 131–188.

Department of Mathematics, University of Haifa, Haifa 31905, Israel
E-mail address: harnik@math.haifa.ac.il