Relative toposes for artificial general intelligence

Olivia Caramello

(University of Insubria - Como, and Grothendieck Institute)

MICS Laboratory, CentraleSupélec (Université Paris-Saclay)
Part 2 - 3 July 2025

Geometric logic

Geometric logic is the logic underlying Grothendieck toposes. Indeed, any theory formulated within geometric logic admits a classifying topos and, conversely, any Grothendieck topos is the classifying topos of some (actually, of infinitely many) geometric theories.

In fact, Grothendieck toposes geometrically embody, in a ‘maximally structured’ way, the logical information contained in geometric theories.

Geometric logic is a particular kind of (infinitary) first-order logic (actually not of inferior expressiveness). It is widely considered as the “logic of finite observations”, and is particularly amenable to computation and automated theorem proving.

Indeed, geometric logic is inherently constructive. This is very relevant from a computer science perspective, by the well-known paradigm identifying programs with constructive proofs.

Higher-order and relative logic

Whilst higher-order logic is clearly more expressive than first-order logic, the model theory of geometric theories is much better-behaved and more amenable to computation, as witnessed for instance by the existence theorem for classifying toposes and universal models.

The development of relative geometric logic will allow us to formalize a great number of higher-order notions whilst preserving geometricity and the computational advantages arising from it. (A nice illustration of the expressive power of relative theories is provided by the ongoing work on a refoundation of functional analysis (which is second-order) as algebra (which is first-order) over a suitable base topos by Fields Medalist Peter Scholze and his collaborator Dustin Clausen.)

Also recall that, in the context of 'syntactic learning', knowledge is represented in a stratified way, by means of sequences of relative theories lying at increasing levels of abstraction.

Computational power of Grothendieck topologies

One of the reasons for the computational effectiveness of geometric logic is the connection with Grothendieck toposes, given by the classifying topos construction.

In my 2017 book it is shown that the classical proof system of geometric logic over a given geometric theory is equivalent to new proof systems whose inference rules correspond to the axioms of Grothendieck topologies. These equivalences actually result from topos-theoretic 'bridges' between different presentations of the classifying topos of the theory.

Interestingly, these alternative proof systems turn out to be computationally much better-behaved than the classical one. In fact, instead of having several axioms and inference rules, they only have *two* inference rules, making it much more manegeable to compute inside them. In fact, there is even a *formula characterizing the 'theorems' provable in such systems*!

'Bridges' involving Grothendieck topologies have also proved very useful for building a great variety of structures presented by generators and relations in most explicit ways.

Model theory in toposes

We can consider models of arbitrary first-order theories in any Grothendieck topos, thanks to the rich categorical structure present on it.

The notion of model of a first-order theory in a topos is a natural generalization of the usual Taskian definition of a (set-based) model of the theory.

Let Σ be a (possibly multi-sorted) first-order signature. A *structure* M over Σ in a category $\mathscr{E}$ with finite products is specified by the following data:

- any sort A of Σ is interpreted by an *object* MA of $\mathscr{E}$
- any function symbol $f : A_1, \ldots, A_n \to B$ of Σ is interpreted as an *arrow* $Mf : MA_1 \times \cdots \times MA_n \to MB$ in $\mathscr{E}$
- any relation symbol $R \rightarrowtail A_1, \ldots, A_n$ of Σ is interpreted as a *subobject* $MR \rightarrowtail MA_1 \times \cdots \times MA_n$ in $\mathscr{E}$

Any formula $\{\vec{x} \,.\, \phi\}$ in a given context $\vec{x}$ over Σ is interpreted as a subobject $[[\vec{x} \,.\, \phi]]_M \rightarrowtail MA_1 \times \cdots \times MA_n$ defined recursively on the structure of the formula.
A model of a theory $\mathbb{T}$ over a first-order signature Σ is a structure over Σ in which all the axioms of $\mathbb{T}$ are satisfied.

Geometric theories

Definition

A geometric theory $\mathbb{T}$ is a theory over a first-order signature Σ whose axioms can be presented in the form $(\phi \vdash_{\vec{x}} \psi)$, where ϕ and ψ are *geometric formulae*, that is formulae in the context $\vec{x}$ built up from atomic formulae over Σ by only using finitary conjunctions, infinitary disjunctions and existential quantifications.

Remark

Inverse image functors of geometric morphisms of toposes always preserve models of a geometric theory (but in general not those of an arbitrary first-order theory).

Most of the first-order theories naturally arising in Mathematics are geometric; anyway, if a finitary first-order theory is not geometric, one can always canonically associate with it a geometric theory, called its *Morleyization*, having the same set-based models.

Classifying toposes

It was realized in the seventies (thanks to the work of several people, notably including W. Lawvere, A. Joyal, G. Reyes and M. Makkai) that:

- Every geometric theory $\mathbb{T}$ has a classifying topos $\mathscr{E}_{\mathbb{T}}$ which is characterized by the following representability property: for any Grothendieck topos $\mathcal{E}$ we have an equivalence of categories
$$\mathbf{Geom}(\mathcal{E}, \mathcal{E}_{\mathbb{T}}) \simeq \mathbb{T}\text{-mod}(\mathcal{E})$$
natural in $\mathcal{E}$, where
  - $\mathbf{Geom}(\mathcal{E}, \mathcal{E}_{\mathbb{T}})$ is the category of geometric morphisms $\mathcal{E} \to \mathcal{E}_{\mathbb{T}}$ and
  - $\mathbb{T}$-mod$(\mathcal{E})$ is the category of $\mathbb{T}$-models in $\mathcal{E}$.
- The classifying topos of a geometric theory $\mathbb{T}$ can be canonically built as the category $\mathbf{Sh}(\mathcal{C}_{\mathbb{T}}, J_{\mathbb{T}})$ of sheaves on the syntactic site $(\mathcal{C}_{\mathbb{T}}, J_{\mathbb{T}})$ of $\mathbb{T}$.

The syntactic category of a geometric theory

Definition (Makkai and Reyes 1977)

- Let $\mathbb{T}$ be a geometric theory over a signature Σ. The syntactic category $\mathcal{C}_{\mathbb{T}}$ of $\mathbb{T}$ has as objects the 'renaming'-equivalence classes of geometric formulae-in-context $\{\vec{x} . \phi\}$ over Σ and as arrows $\{\vec{x} . \phi\} \to \{\vec{y} . \psi\}$ (where the contexts $\vec{x}$ and $\vec{y}$ are supposed to be disjoint without loss of generality) the $\mathbb{T}$-provable-equivalence classes $[\theta]$ of geometric formulae $\theta(\vec{x}, \vec{y})$ which are $\mathbb{T}$-provably functional i.e. such that the sequents
$$(\phi \vdash_{\vec{x}} (\exists y)\theta),$$
$$(\theta \vdash_{\vec{x},\vec{y}} \phi \wedge \psi), \text{ and}$$
$$((\theta \wedge \theta[\vec{z}/\vec{y}]) \vdash_{\vec{x},\vec{y},\vec{z}} (\vec{y} = \vec{z}))$$
are provable in $\mathbb{T}$.
- The composite of two arrows
$$\{\vec{x} . \phi\} \xrightarrow{[\theta]} \{\vec{y} . \psi\} \xrightarrow{[\gamma]} \{\vec{z} . \chi\}$$
is defined as the $\mathbb{T}$-provable-equivalence class of the formula $(\exists \vec{y})\theta \wedge \gamma$.
- The identity arrow on an object $\{\vec{x} . \phi\}$ is the arrow
$$\{\vec{x} . \phi\} \xrightarrow{[\phi \wedge \vec{x'} = \vec{x}]} \{\vec{x'} . \phi[\vec{x'}/\vec{x}]\}$$

The syntactic site

On the syntactic category of a geometric theory it is natural to put the Grothendieck topology defined as follows:

Definition

The syntactic topology $J_{\mathbb{T}}$ on the syntactic category $\mathcal{C}_{\mathbb{T}}$ of a geometric theory $\mathbb{T}$ is given by:

a small family $\{[\theta_i] : \{\vec{x_i} . \phi_i\} \to \{\vec{y} . \psi\}\}$ in $\mathcal{C}_{\mathbb{T}}$ is $J_{\mathbb{T}}$-covering

if and only if

the sequent $(\psi \vdash_{\vec{y}} \bigvee_{i \in I} (\exists \vec{x_i})\theta_i)$ is provable in $\mathbb{T}$.

This notion is instrumental for identifying the models of the theory $\mathbb{T}$ in any geometric category $\mathcal{C}$ (and in particular in any Grothendieck topos) as suitable functors defined on the syntactic category $\mathcal{C}_{\mathbb{T}}$ with values in $\mathcal{C}$; indeed, these are precisely the $J_{\mathbb{T}}$-continuous cartesian functors $\mathcal{C}_{\mathbb{T}} \to \mathcal{C}$. So if $\mathcal{C}$ is a Grothendieck topos they correspond precisely to the geometric morphisms from $\mathcal{C}$ to $\mathbf{Sh}(\mathcal{C}_{\mathbb{T}}, J_{\mathbb{T}})$. This topos therefore classifies $\mathbb{T}$.

Interpretations and expansions

Definition

An interpretation (resp. a generalized interpretation) of a geometric theory $\mathbb{S}$ into a geometric theory $\mathbb{T}$ is a geometric functor $\mathcal{C}_{\mathbb{S}} \to \mathcal{C}_{\mathbb{T}}$ (resp. a geometric functor $\mathcal{C}_{\mathbb{S}} \to \mathcal{E}_{\mathbb{T}}$).

[Note that the objects of $\mathcal{E}_{\mathbb{T}}$ can be represented as "definable" quotients of coproducts of objects coming from $\mathcal{C}_{\mathbb{T}}$.]

Remark

Any expansion $\mathbb{T}$ of a geometric theory $\mathbb{S}$ (i.e. a theory written over a larger signature in which every axiom of $\mathbb{S}$ is provable) yields an interpretation of $\mathbb{S}$ into $\mathbb{T}$.

Any (generalized) interpretation I of $\mathbb{S}$ into $\mathbb{T}$ induces a geometric morphism $f_I : \mathcal{E}_{\mathbb{T}} \to \mathcal{E}_{\mathbb{S}}$ (and, conversely, any morphism $f : \mathcal{E}_{\mathbb{T}} \to \mathcal{E}_{\mathbb{S}}$ arises uniquely from a generalized interpretation I).

Semantically, composition with f_I yields, for any topos $\mathcal{E}$, a functor

$$s_I^{\mathcal{E}} : \mathbb{T}\text{-mod}(\mathcal{E}) \to \mathbb{S}\text{-mod}(\mathcal{E})$$

such that, for any geometric formula $\phi(\vec{x})$ over the signature of $\mathbb{S}$, for any $\mathbb{T}$-model M in $\mathcal{E}$, the interpretation of $I(\phi(\vec{x}))$ in M coincides with the interpretation of $\phi(\vec{x})$ in $s_I^{\mathcal{E}}(M)$.

A theorem for interwining theories

The following result allows us to turn any morphism f between the classifying toposes of two theories $\mathbb{T}$ and $\mathbb{S}$ into a richer theory $\mathbb{T}^f_{\mathbb{S}}$ "interwining $\mathbb{T}$ and $\mathbb{S}$ along f": indeed, the theory $\mathbb{T}^f_{\mathbb{S}}$ extends the theory $\mathbb{T}$ through the addition of new sorts, function and relation symbols corresponding to the way in which $\mathbb{S}$ is interpreted in the classifying topos of $\mathbb{T}$ via the morphism f.

Theorem

Let $\mathbb{T}$ and $\mathbb{S}$ geometric theories and $f : \mathcal{E}_{\mathbb{T}} \to \mathcal{E}_{\mathbb{S}}$ a morphism between their classifying toposes. Then there exists an expansion $\mathbb{T}^f_{\mathbb{S}}$ of the theory $\mathbb{T}$ which is also an expansion of the theory $\mathbb{S}$ such that the canonical interpretations $H : (\mathcal{C}_{\mathbb{T}}, J_{\mathbb{T}}) \to (\mathcal{C}_{\mathbb{T}^f_{\mathbb{S}}}, J_{\mathbb{T}^f_{\mathbb{S}}})$ and $K : (\mathcal{C}_{\mathbb{S}}, J_{\mathbb{S}}) \to (\mathcal{C}_{\mathbb{T}^f_{\mathbb{S}}}, J_{\mathbb{T}^f_{\mathbb{S}}})$ make the following diagram commute and f_H is an equivalence of toposes:

$$
\begin{tikzcd}
\mathbf{Sh}(\mathcal{C}_{\mathbb{T}^f_{\mathbb{S}}}, J_{\mathbb{T}^f_{\mathbb{S}}}) \arrow[d, "f_H", "\wr"'] \arrow[dr, "f_K"] & \\
\mathbf{Sh}(\mathcal{C}_{\mathbb{T}}, J_{\mathbb{T}}) \arrow[r, "f"] & \mathbf{Sh}(\mathcal{C}_{\mathbb{S}}, J_{\mathbb{S}})
\end{tikzcd}
$$

Examples

The theorem is naturally applied in the context of morphisms f induced by (generalized) interpretations of $\mathbb{S}$ into $\mathbb{T}$, e.g.

- Given an "alphabet" $\mathcal{A}$, the propositional theory $\mathbb{A}$ with one symbol P_a for each $a \in \mathcal{A}$ and the axioms $(\top \vdash \bigvee_{a\in\mathcal{A}} P_a)$ and $(P_a \wedge P_{a'} \vdash \bot)$ for any $a \neq a'$ in $\mathcal{A}$. The (one-sorted) abstract theory $\mathbb{L}$ of "letters" interprets in $\mathbb{A}$ by sending $\{x^L . \top\}$ to the coproduct $\coprod_{a\in\mathcal{A}} P_a$. 'The' resulting theory $\mathbb{A}_{\mathbb{L}}$ can be seen as the theory of "letters in the alphabet $\mathcal{A}$"; indeed, it extends the theory obtained from $\mathbb{L}$ by adding a constant c_a for each $a \in \mathcal{A}$ and the following axioms:

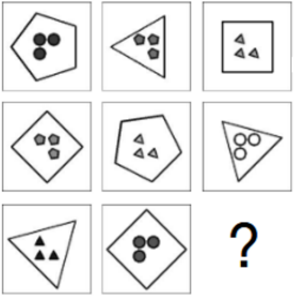
$$c_a = c_{a'} \vdash \bot \text{ for any } a \neq a';$$

$$(T \vdash_x \bigvee_{a\in\mathcal{A}} x = c_a) .$$

- The theory $\mathbb{W}$ of "words" is obtained from the theory $\mathbb{L}$ of "letters" by the same method, using the interpretation of $\mathbb{W}$ into $\mathbb{L}$ sending $\{x^W . \top\}$ to the list object $L(\{x^L . \top\})$.

Note the role of the interpretation functor I as a geometric "constructor" of higher-level entities from lower-level ones.

Modelling of Raven matrices I



We may use the theorem also to model the different levels of abstractions involved in Raven matrices problems.

Solving a Raven matrix requires the identification of certain invariances, whose logical expression lies entirely within the framework of (relative) geometric logic.

We can identify four relevant theories, defined on top of each other, as follows:

- the theory $\mathbb{M}$ of matrices.
- the theory $\mathbb{R}$ of rows;
- the theory $\mathbb{C}$ of cells;
- the theory $\mathbb{I}$ of the internal structure of a cell.

Modelling of Raven matrices II

- The theory $\mathbb{I}$ is meant to axiomatize the internal structure of cells: so, it should have unary predicates corresponding to the usual attributes of individual objects lying inside cells (e.g. Type, Size, Color, Position) as well as predicates for expressing the relations between them (for instance, the difference relations - useful for counting the number of elements inside a cell - or the interior/exterior relations).
- The theory $\mathbb{C}$ is the one that axiomatizes cells, i.e. the structures obtained by putting together, in a single set, the objects in the inner structure of a cell (that is, the elements of a model of the theory $\mathbb{I}$). For building it, one interprets the empty theory with one sort in the theory $\mathbb{I}$ by sending $\{x^C . \top\}$ to $S(\{x^I . \top\})$, where S is the geometric constructor assigning to an object A the coproduct (over the natural numbers) of the quotients A^n by the permutation group on n elements.

Modelling of Raven matrices III

- The theory $\mathbb{R}$ is meant to axiomatize the rows of the matrix, intended as lists of cells. Accordingly, this theory arises by considering the interpretation of the empty theory with one sort in the theory $\mathbb{C}$ which sends $\{x^R . \top\}$ to $L(\{x^C . \top\})$. If one wants to consider only rows with a fixed lenght n, then one considers the interpretation sending $\{x^R . \top\}$ to $\{x^C . \top\}^n$.
- The theory $\mathbb{M}$ is the one that axiomatizes matrices, viewed as lists of rows. It can be obtained from the theory $\mathbb{R}$ of rows by using the interpretation of the empty theory with one sort in the theory $\mathbb{R}$ which sends $\{x^M . \top\}$ to $L(\{x^R . \top\})$. Matrices with a fixed number of rows correspond to quotients of this theory (we have predicates in it, indexed by the natural numbers, which correspond to the subobjects $\{x^R . \top\}^n$ of $L(\{x^R . \top\})$).

Global and local properties

Note that, in applying the theorem, instead of considering interpretations from empty theories, one could consider interpretations from richer theories, in case there are natural "global" operations expressible in their abstract language (such as, in the case of the theory of cells, the rotations of all the elements of a cell, or, in the case of the theory of rows, the permutation of cells of a row).

Within an "interwined" theory, we can notably distinguish between properties which are inherently global and properties which instead are local (i.e. whose definition requires referring to elements lying at lower levels): indeed, by identifying properties with subtoposes, the former correspond precisely to the subtoposes obtainable by taking the fibered products of subtoposes of the classifying topos of $\mathbb{S}$ along the interpretation geometric morphism $f : \mathcal{E}_{\mathbb{T}} \to \mathcal{E}_{\mathbb{S}}$.

The language of invariances

In our formalization, the theories resulting from n successive applications of the theorem have stratified vocabularies with n levels.

Starting from from predicates $\{T_i \mid i \in I\}$ lying at level n and tuples of terms $\vec{t_i}$, one can inductively generate predicates $P_{\{T_i \mid i \in I\}}$ lying at level $n+1$ by the following very general scheme:

$$P_{\{T_i \mid i \in I\}}(\vec{y}) \dashv\vdash_{\vec{y}} \bigvee_{i \in I} \exists \vec{x_i}(\vec{y} = \vec{t_i}(\vec{x_i}) \wedge T_i(\vec{x_i})) \ .$$

(the terms $\vec{t_i}$ are typically those which arise in the construction of the elements $\vec{y}$ at level $n+1$ in terms of the elements $\vec{x_i}$ at level n according to the interpretation morphism I; so, they belong to the vocabulary of the interwined theory constructed by applying the theorem).

Solving Raven matrices

The classical Raven matrix problems require the identification of properties of rows, i.e. properties lying at level 2 in our formalization (recall that a Raven matrix is constructed in such a way that all the rows share a characteristic property given by the application of certain rules acting on the attributes of the innner structures of cells composing it).

These rules lie at level 2 but arise from functions or predicates lying at the ground level, i.e. involving the sets which parametrize the predicates of the ground theory $\mathbb{I}$. Recall that, at level 0, we have predicates that are indexed by certain (finite) sets, endowed with certain operations, the so-called rules for attributes (e.g. *Constant*, *Arithmetic*, *Progression*, *Distribute*).

Solving a Raven matrix thus requires making an abstraction leap from level 0 (that of inner structures of cells) to level 2 (that of rows) so as to identify the right geometric formula in the language of the theory $\mathbb{R}$ which expresses the given invariance.

An example

As an example, suppose that the matrix is a 3×3 one, and that the rule to be identified, acting on the attribute Position, is the *Arithmetic* one. Let

$$g : \mathrm{Pos} \times \mathrm{Pos} \to \mathrm{Pos} .$$

be the function formalizing this rule.

For any predicate R_i indexed by $i \in \mathrm{Pos}$, we have, by following the previous generation scheme, a unary predicate $P_{R_i}(x^C)$ given by $\exists \vec{z}_i(x = \pi_i(\vec{z}_i) \wedge R_i(\vec{z}_i))$, where π_i is the function symbol in the interwined theory corresponding to the canonical projection from $\{x^I \,.\, \top\}^3$ to the quotient by the permutation group.

Then the invariance property is expressed by the geometric formula in the language of the theory $\mathbb{R}$

$$\bigvee_{(i_1, i_2, i_3) \in \mathrm{Graph}(g)} \exists x_1 \exists x_2 \exists x_3 (\vec{y} = (x_1, x_2, x_3) \wedge P_{R_{i_1}}(x_1) \wedge P_{R_{i_2}}(x_2) \wedge P_{R_{i_3}}(x_3))$$

Integration with deep learning techniques

We have seen that Raven matrices problems can be reformulated in terms of the identification of a certain geometric formula lying at level 2 (namely, the one which expresses the given invariance).

More generally, ARC-type problems can be reinterpreted in terms of generation of formal stratified vocabularies and identification of geometric formulas written in them which express the given regularities.

Our techniques can thus be very useful in significantly reducing the space of parameters to be tested in training artificial systems, leading to AI systems that are much more efficient, but also more conceptually inspired and explainable.

They could be swiftly integrated with LLMs, through algorithms for “lifting” the statistically discovered invariances to sequents provable in geometric theories. The likelihood of a sequent being true could be estimated on the basis of the available data, as well as on its logical complexity (one can start testing the sequents which are simpler from the arithmetic or logical point of view).

Modelling of stratified phenomena

These techniques can be used in all situations where we need to model stratified phenomena:

Raven matrices	Natural language	Biology
Matrices	Texts	Living being
Rows	Words	Organs
Cells	Letters	Tissues
Cell inner structure	Alphabet	Cells

They could also be applied to the study of games, formalized in such a way that "tactics" and even "playing styles" can be expressed as geometric sequents in stratified vocabularies. The resulting AI systems would then be able to infer and formally express all sorts of (meta-)rules relevant for game playing.

In music, the formal identification of the invariances which make a piece of music beautiful would be crucial for designing systems for artificially composing good music (as well as for better understanding and emulating the styles of great composers).