Toposes and 'bridges' for artificial general intelligence

Olivia Caramello

(University of Insubria - Como, Grothendieck Institute and Mines Paris - PSL)

EDF, Paris-Saclay, 21 May 2026

Plan of the talk

- Motivation
- General principles for a future AI
  - Syntactic learning
  - Relative toposes and meta-learning
- Modelling with relative toposes
  - Linguistic meta-learning
  - The logic of games
  - A financial application
  - Chains of relative theories
- Classiying toposes and 'bridges'
- Modelling of Raven matrices and ARC-type problems
- The Grothendieck Institute

The key features of true intelligence

Human intelligence enjoys a number of fundamental features that are responsible for its effectiveness:

- the ability to identify invariants, giving a qualitative, rather than numerical or quantitative, understanding of complex phenomena;
- the ability to combine and synthetize different points of view on a given theme, corresponding to different invariants;
- the ability to function at different levels of abstraction, and to build on previously accumulated knowledge;
- the use of languages of different kinds to describe and represent pieces of reality, so that findings about the world can be expressed in them, stored and communicated to other agents.

Despite the spectacular advances in deep learning systems of the last years, we are still far from a form of artificial intelligence enjoying these features. In fact, new mathematical foundations are needed to achieve these goals.

The next-generation AI

The new generation of artificial learning systems should be based on computations that are not blind, but conceptually inspired and meaningful to the human mind.

In particular, we should shift to a new conception of information which is based on semantics and invariants rather than on the classical set-theoretic foundations for mathematics.

Topos theory is bound to play a key role in this kind of developments.

In this talk, I will focus on one particular aspect of the theory, that is, the possibility of studying toposes in relation to each other, and discuss its relevance for AI. In particular, I will argue about the role that relative toposes can play in connection with the development of systems for meta-learning (i.e. learning taking place at different levels of abstraction constructed on top of each other).

Syntactic learning

To achieve the above-mentioned goals, artificial learning systems should be equipped with *extensive formal vocabularies* that will be used to express the concepts that the system will learn from data.

Indeed, learning should take place at the abstract level of syntax, rather than at that of a particular semantics, as is currently the case. Although a basic vocabulary should be provided at the beginning, we should let the system discover by itself all the logical rules that can be expressed in that vocabulary, by using the usual statistical techniques. We should also let the system suggest extensions and refinements of the vocabulary on the basis of empirically discovered invariances in the data, thus achieving a form of 'concept emergence'.

In this way, we will obtain systems capable of extracting all sorts of 'syntactic rules' from data (e.g. the grammatical rules of a language from a large number of text samples, or the rules of a game from a large collection of matches, etc.).

Making AI systems 'reason'

All this points to the construction of logical languages for AI agents, enabling them, in a sense, to *'reason' through language*. After all, how can we expect a system to learn robustly if we do not give it the possibility to reason linguistically?

Think, for instance, about the process of learning a natural language. Young children have no other way of learning a language than to rely on data (note, however, that their brains already possess many a priori structures that are used to organise and categorise the knowledge they gather from the environment). However, their knowledge of language remains fragile until they come into contact, mostly at school, with grammar, which represents syntax in this context.

Note that the role of grammar is crucial in making *explicit* much of the *implicit* information accumulated during the previous 'bottom-up' learning process and in taking learning to a higher level, especially in terms of explanation skills.

Towards artificial general intelligence

Our program of topos-empowered 'syntactic learning' points to the development of new AI systems that are:

- explanatory (thanks to the *use of symbolic logic* and the coding of *invariants*)
- secure (as they allow for a *certification* of the generated results thanks to the integration of the usual, data-driven statistical AI techniques with *formal proofs* in relative geometric logic)
- modular (as they natively support the *interoperability* between different knowledge representation systems thanks to the role of toposes as 'bridges')
- energy-saving and robust, i.e. requiring lesser amounts of data for the training and achieving better generalization capabilities (thanks to the *equivariance constraints* in the optimization induced by the logical vocabularies embedded in the system)
- able to *explicitly* operate across different levels of abstraction i.e. capable of meta-learning.

Relative toposes and meta-learning

Since toposes are spaces that embody information, in a way which is invariant with respect to the different knowledge representations, the theory of relative toposes is crucial in relation to the development of a theory of semantic information.

A relative topos (i.e. *a topos defined above another topos*) corresponds to a way of organising a given piece of information by incorporating part of it in the basic topos and the 'remaining part' in the topos defined over it.

Modelling learning through chains of relative toposes thus allows us to formalise the crucial capacity of human intelligence to reason at different levels of abstraction and to learn on top of existing knowledge, in other words, to realise a form of meta-learning.

Linguistic meta-learning

Think, for example, of learning a language: before one can write good prose or poetry, one needs to master *grammar* perfectly. But to learn grammar, it is necessary to first familiarize with the *vocabulary* of the language, which, in turn, requires knowledge of the *alphabet* in which it is written.

In a sense, knowledge of the alphabet lies at level 0, vocabulary at level 1 and grammar at level 2, while prose and poetry lie at much higher levels, defined each on top of the formers.

Learning of music follows a similar pattern, as does the learning of any art or science.

The logic of games

Games can be modelled very naturally in terms of chains of relative toposes, with the base topos formalising the '*hardware*' of the game, the level 1 topos the *basic, first-order rules of the game*, and the higher-level relative toposes the higher-level rules (i.e. *tactics* or *strategies*, up to *playing styles*).

Each morphism in the chain encodes the way in which the rules at that level are concretely realized in terms of the rules at a lower level; in particular, the morphism from the level 1 topos to the base topos expresses the way in which the basic rules (e.g. those governing the movement of the game pieces) are arithmetically encoded in the game hardware.

A topos-inspired trading system

Since 2013, the Italian company EOS S.r.l. has developed an award-winning stock trading system inspired by the above-mentioned topos-theoretic ideas. The most recent developments of the system involve the construction of meta-models of order up to 10, which perform significantly better than ordinary (or lower-order) models in terms of robustness and resilience, as shown by the following table (courtesy of EOS):

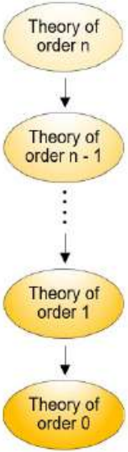
LEVEL	Performance		Differential with respect to the benchmark 2003-2023		
	Back-test period 2003-2014	Offline period 2015-2023	% positive years	Best annual differential	Worst annual differential
Level 0	18.72	21.70	90.79%	54.28%	-5.45%
Level 1	14.08	14.11	90.48%	36.43%	-3.88%
Level 2	14.04	18.07	94.18%	39.05%	-2.83%
Level 3	14.36	16.94	93.65%	34.63%	-3.61%
Level 4	14.53	17.74	98.41%	36.24%	0.50%
Level 5	14.80	18.46	96.30%	38.14%	-0.32%
Level 6	13.46	18.65	96.30%	33.12%	-0.44%
Level 7	14.18	19.97	97.88%	34.81%	-0.50%
Level 8	14.06	18.56	96.30%	32.60%	-0.03%
Level 9	13.05	19.04	98.94%	31.87%	1.22%
Level 10	13.54	19.68	98.94%	34.85%	1.01%

The table reports the results of theoretical operativity, for the period 2003-2023, of models and meta-models applied to a basket of large-cap stocks traded on the Italian market. The benchmark is the FTSE ITALIA All-Share Index. The total period is divided into two sub-periods: back-test (2003-2014) and offline (2015-2023).

Chains of relative theories

Learning can be modelled in terms of chains of relative toposes, each defined over the previous one, with the ground one corresponding to 'raw data'. Whereas ordinary toposes can only classify first-order theories, relative toposes can classify theories of an arbitrarily high order. In this way, the complexity of a situation to be modelled can be decomposed across different layers of increasing degree.

For example, a theory of order 1 with respect to the ground topos can become a theory of order 0 with respect to a topos which is richer than the ground topos. More generally, a theory of order n can be regarded as a theory of order $n - k$ with respect to a k-chain of relative toposes.



Classifying toposes

A fundamental connection between topos theory and logic is realized by the theory of classifying toposes.

With any first-order mathematical theory $\mathbb{T}$ (of a very general form) one can canonically associate a topos $\mathcal{E}_{\mathbb{T}}$, called its classifying topos, which represents its 'semantical core'.

The classifying topos embodies the semantic 'essence' of the theory, an essence which is invariant with respect to its different syntactic (axiomatic) presentations.

The classifying topos of a theory is constructed by means of completion process of the theory, with respect, in a sense, to all the concepts that it is potentially capable to express.

It is at the level of these completed objects that invariants actually live.

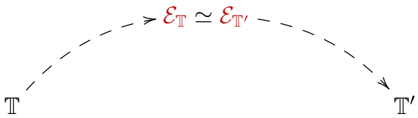
From a sketch to reality

Every language, in its attempt to express a reality that is much richer, can be compared to a sketch; the transition from a linguistic expression to its meaning is thus a kind of completion, similar to the automatic one performed by our brain as we watch the artist drawing it:

The classifying topos construction provides a mathematical formalization of this completion process. Indeed, the classifying topos makes *explicit* whatever is *implicit* in the theory.

Toposes as *bridges*

- The existence of theories which are sematically equivalent to each other translates into the existence of different presentations for the same Grothendieck topos.
- Grothendieck toposes can be effectively used as 'bridges' for transferring notions, properties and results across different sematically equivalent theories:



- The transfer of information takes place by expressing topos-theoretic invariants in terms of the different sites of definition (or, more generally, presentations) for the given topos.
- As such, different properties (resp. constructions) arising in the context of theories classified by the same topos are seen to be different *manifestations* of a *unique* property (resp. construction) lying at the topos-theoretic level.

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 $\mathbb{T}_{\mathbb{S}}^{f}$ "interwining $\mathbb{T}$ and $\mathbb{S}$ along f": indeed, the theory $\mathbb{T}_{\mathbb{S}}^{f}$ 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 (O.C.)

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 $\mathbb{T}_{\mathbb{S}}^{f}$ of the theory $\mathbb{T}$ which is also an expansion of the theory $\mathbb{S}$ such that the canonical interpretations $H : (\mathcal{C}_{\mathbb{T}}, J_{\mathbb{T}}) \to (\mathcal{C}_{\mathbb{T}_{\mathbb{S}}^{f}}, J_{\mathbb{T}_{\mathbb{S}}^{f}})$ and $K : (\mathcal{C}_{\mathbb{S}}, J_{\mathbb{S}}) \to (\mathcal{C}_{\mathbb{T}_{\mathbb{S}}^{f}}, J_{\mathbb{T}_{\mathbb{S}}^{f}})$ make the following diagram commute and f_H is an equivalence of toposes:

$$
\begin{array}{ccc}
\mathbf{Sh}(\mathcal{C}_{\mathbb{T}_{\mathbb{S}}^{f}}, J_{\mathbb{T}_{\mathbb{S}}^{f}}) & & \\
\wr\downarrow f_H & \searrow^{f_K} & \\
\mathbf{Sh}(\mathcal{C}_{\mathbb{T}}, J_{\mathbb{T}}) & \xrightarrow{\ f\ } & \mathbf{Sh}(\mathcal{C}_{\mathbb{S}}, J_{\mathbb{S}})
\end{array}
$$

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 (partial) constant c_a for each $a \in \mathcal{A}$ and the following axioms:

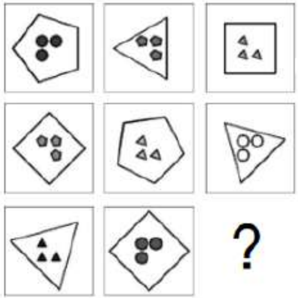
$$c_a = c_{a'} \vdash \bot \text{ for any } a \neq a';$$

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

- 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:

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

(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

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

be the function formalizing this rule.

For any predicate R_i indexed by $i \in \text{Pos}$, 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 $\mathbb{R}$

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

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).

The Grothendieck Institute

- These projects are notably developed at the Grothendieck Institute, an international foundation devoted to cutting-edge research in mathematics and its interactions with other disciplines.
- Established in 2022, the Institute is named after Alexander Grothendieck, whose work it is committed to valorize and disseminate.

www.igrothendieck.org

- The philosophical aspects of Grothendieck's work are notably investigated at the Institute's Centre for Grothendieckian Studies: www.csg.igrothendieck.org

If you wish to collaborate, don't hesitate to get in touch!