Relative toposes for artificial general intelligence

Olivia Caramello

(University of Insubria - Como, and Grothendieck Institute)

MICS Laboratory, CentraleSupélec (Université Paris-Saclay)

Part 1 - 26 June 2025

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 course, 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).

The relativity method

- In Mathematics the relativity method consists in trying to state notions and results in terms of morphisms, rather than objects, of a given category, so that they can be 'relativized' to an arbitrary base object.
- One works in the new base universe as if it were the 'classical' one, and then interprets the obtained results from the point of view of the original universe. This process is usually called *externalization*.
- Relativity techniques can be thought of as general 'change of base techniques', allowing one to choose the universe relatively to which one works according to one's needs.
- The relativity method has been pioneered by Grothendieck, in particular for schemes, in his categorical refoundation of Algebraic Geometry, and has played a key role in his work.
- Since the theory of toposes is more general and technically better-behaved than the theory of schemes, it is particularly important to develop relativity techniques for toposes.

Applications to mathematics

- By definition, a relative topos is a morphism between toposes. The codomain topos is regarded as the base over which the domain topos is defined.
- The relativity method provides an extremely high degree of technical flexibility, resulting from the possibility of 'encapsulating' part of the complexity of a situation in the base topos, so that the given notions acquire a simpler (e.g. a lower-degree) expression with respect to it.
- For example, a first-order theory becomes a propositional one when regarded relative to a suitable base topos.
- Several mathematicians have already started using relativity techniques for toposes:
  - Scholze and Clausen's condensed mathematics;
  - Tao and Jamneshan's topos-theoretic measure theory;
  - Tomasic's topos-theoretic difference algebra.

The theory of relative toposes is currently being systematically developed, from the geometric, algebraic and logical points of view, by our research team.

Relative toposes for AI

- Since toposes are spaces which embody information, in a way which is invariant with respect to the different syntactic presentations, the theory of relative toposes is also crucial in connection with the development of a theory of semantic information.
- A relative topos corresponds to a way of organizing a given piece of information by incorporating part of it in the base topos and the ‘remaining part’ in the topos defined over it.
- Modelling learning through chains of relative toposes thus allows us to formalize the crucial ability of human intelligence to reason at different levels of abstraction and learn on top of existing knowledge, that is, as we shall see, to realize a form of meta-learning.

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 $\mathscr{E}_{\mathbb{T}}$, 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.

Empowering learning systems with logic

Artificial learning systems, in particular deep neural networks, should be empowered by building on the notions of mathematical theory and proof:

- Any intelligent agent must, in order to get an effective understanding of (aspects of) the world, derive knowledge starting from certain 'sensory' inputs, which play a similar role to that of axioms for a mathematical theory, by following certain dynamical rules, which correspond to the inference rules of the logical system inside which the mathematical theory is formulated.
- As every mathematical theory can be enriched by the addition of new axioms, so the functioning of an agent can be updated by the integration of new information which becomes available to it.
- The functioning of a learning system would thus correspond to a sequence of a mathematical theories, each of which richer or more refined than the previous ones.

Learning processes via proofs

- In principle, any learning sequence, as any process 'approximating truth', is infinite, but for practical purposes one is normally satisfied by the result of a learning process when the last theory in the sequence leaves a degree of ambiguity which is sufficiently low for the desired applications.
- All the theories in the sequence should extend (that is, be defined over) the basic theory of the agent formalizing its essential features. More generally, every theory in the sequence can be seen as theory defined *over* each of the preceding ones. This corresponds, through the classifying topos construction, to a sequence of relative toposes.
- Note that any constraints embedded in the logical formalism (or integrated at some step of the sequence of theories) will allow to significantly reduce the space of parameters that the agent has to explore and hence correspondingly decrease the computational complexity of the learning process.

The stratified nature of knowledge

Learning is a process that goes through a sequence of steps: any new, learned information is stored in the brain according to pre-existing structures.

In fact, human learning is essentially relational, i.e. it consists in relating the given new piece of information with the previously accumulated knowledge. This is actually what makes it robust; in fact, only structured forms of learning can enjoy strong forms of robustness and resilience.

The stratified nature of human knowledge parallels the stratified nature of things themselves (think, for instance, of biological phenomena, which unfold at different levels, such as that of cells, or tissues, or organs...).

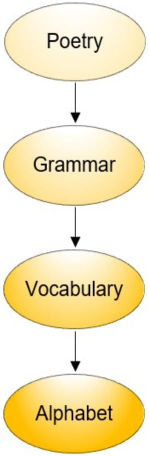
The 'relative nature' of learning is precisely what enables us to arrive at arbitrarily high degrees of complexity and abstraction in our learning/understanding: indeed, this is achieved by means of a structured composition of successive fundamental steps, whose nature can be more easily investigated.

Learning a language

Think, for instance, about learning a language: before being able to write good prose or poetry, one needs to have a perfect command of the grammar. Now, in order to be able to properly learn a grammar, one needs to have a familiarity with the vocabulary of the language, which in turn requires a knowledge of the alphabet in which it is written.

So, we can regard the knowledge of the alphabet as lying at level 0, that of the vocabulary as lying at level 1 and that of grammar lying at level 2, while prose and poetry lie at much higher levels defined over the preceding ones.

The learning of music follows a similar scheme, as well as that of any art or science.



Stratified knowledge and relativity principles

The theory of relative toposes provides a mathematical formalisation of this learning architecture.

Given a piece of information embodied by a certain topos, finding a 'base topos' over which the given topos is defined corresponds to an abstraction operation, which consists in organizing the given piece of information by incorporating part of it in the base topos and the remaining part in the topos defined over it.
For example, one can reduce complex notions to simpler ones by changing the base category relative to which one works:

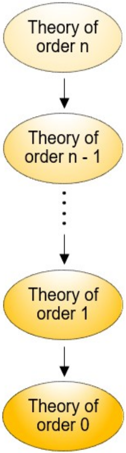
- Topological vector space = vector space relative to the category of topological spaces;
- Lie group = group relative to the category of differentiable manifolds.

The power of relativity techniques is that one can do topos theory relative to an arbitrary base topos *as if* that base topos were the ordinary topos of sets, but the results obtained, once externalized (that is, reinterpreted from the point of view of the 'ground level') are typically much more complex.

Chains of relative theories

Learning thus corresponds to a sequence of relative toposes, each defined over the previous one, with the ground one corresponding to 'raw data'. In fact, each topos in the sequence spatially embodies a certain amount of information, which can also be presented in logical form: while ordinary, ground-base, classifying toposes can only classify first-order theories, by means of relative toposes one can classify theories of arbitrarily high order.

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.



Syntactic learning

In order to practically implement the above ideas, one should, first of all, empower any learning system with large formal vocabularies that will serve for expressing the concepts that the system will learn from data.

The idea is to enforce the learning to take place at the abstract level of syntax, rather than at that of a particular semantics, as it currently happens.

While the vocabulary should be given at the outset, we could let the system discover any (non-already embedded) logical rules expressible in that vocabulary by itself, by using the usual techniques. Also, we should let it suggest enrichments of the vocabulary on the basis of invariances empirically discovered in the data, thereby achieving a form of 'emergence of concepts'.

In this way, we could obtain systems capable of inferring all sorts of 'syntactic rules' from data (e.g. the grammar rules of a language from a great amount of samples of texts, or the rules of a game from a big collection of matches, etc.): these rules could, of course, be of different nature and complexity (e.g., propositional, first-order, higher-order, etc.).

Making AI systems 'reason'

All of this goes in the direction of constructing logical languages for AI agents, allowing them, in a sense, to 'reason'. After all, how can we expect a system to learn in a robust way if we do not give it the possibility of reasoning linguistically? Think about how our learning, as human beings, would be impaired without the possibility of expressing, testing and communicating our ideas by using languages.

Think also about the process of learning of a natural language. Little babies have no other possibility to learn a language than to merely rely on data (note, however, that their brain has already a lot of *a priori structures* which are used to organize and categorize the knowledge that they gather from the environment). Still, their knowledge of the language remains fragile until they are brought into contact, notably at school, with grammar, which represents syntax in this context.

The role of grammar is crucial in making *explicit* a lot of the *implicit* which had been accumulated in the previous 'bottop-up' learning process, and in bringing the understanding to a higher level, notably in terms of explainability.

Uncovering hidden structures and rules

Syntax represents the 'skeleton' of structures, whose concrete manifestations we access through data. Trying to 'lift' from a particular concrete setting to the level of syntax represents an abstract form of understanding which enjoys much more relisience and adaptability than the usual forms of statistically-based learning, which do not go to the roots of the reasons for the existence of the given relations.

The deep reasons for the regularities that we may observe in concrete contexts actually live at the syntactic level (think, for instance, of motives in Algebraic Geometry, or to definability and preservation theorems in Logic).

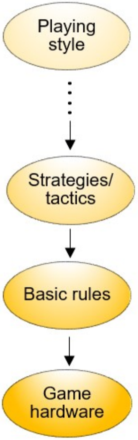
Philosophically speaking, we should teach learning systems to lift from the phenomenological level (of concrete manifestations of topos invariants) to the ontological one (of toposes themselves, regarded as classifying toposes of logical theories).

In particular, we should aim for an integration of a 'bottop-up' approach to artificial learning, such as the one which is dominant today, with a 'top-down' one based on logic and topos invariants.

The logic of games

Games can be very conveniently modelized in terms of chains of relative toposes, with the ground topos formalizing 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).

Each morphism in the chain encodes the way in which the rules at that level are concretely realized in terms of lower-level ones; 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 beheviour of the "pieces" of the game) are "arithmetically encoded" in the game's 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).

Grothendieck topologies

Definition

A Grothendieck topology on a small category $\mathscr{C}$ is a function J which assigns to each object c of $\mathscr{C}$ a collection $J(c)$ of sieves on c in such a way that

(i) (maximality axiom) the maximal sieve $M_c = \{f \mid cod(f) = c\}$ is in $J(c)$;

(ii) (stability axiom) if $S \in J(c)$, then $f^*(S) \in J(d)$ for any arrow $f : d \to c$;

(iii) (transitivity axiom) if $S \in J(c)$ and R is any sieve on c such that $f^*(R) \in J(d)$ for all $f : d \to c$ in S, then $R \in J(c)$.

The sieves S which belong to $J(c)$ for some object c of $\mathscr{C}$ are said to be J-covering.

A site is a pair $(\mathscr{C}, J)$ consisting of a category $\mathscr{C}$ and a Grothendieck topology J on $\mathscr{C}$.

Grothendieck toposes

One can define sheaves on an arbitrary site in a formally analogous way to how one defines sheaves on a topological space. This leads to the following

Definition

- A Grothendieck topos is a category (equivalent to the category) $\mathbf{Sh}(\mathscr{C}, J)$ of sheaves on a (small-generated) site $(\mathscr{C}, J)$.
- A geometric morphism of toposes $f : \mathscr{E} \to \mathscr{F}$ is a pair of adjoint functors whose left adjoint (called the inverse image functor) $f^* : \mathscr{F} \to \mathscr{E}$ preserves finite limits.
- Let f and $g : \mathscr{E} \to \mathscr{F}$ be geometric morphisms. A geometric transformation $\alpha : f \to g$ is defined to be a natural transformation $a : f^* \to g^*$.
- Given two toposes $\mathscr{E}$ and $\mathscr{F}$, geometric morphisms from $\mathscr{E}$ to $\mathscr{F}$ and geometric transformations between them form a category, denoted by $\mathbf{Geom}(\mathscr{E}, \mathscr{F})$.
- A point of a topos $\mathscr{E}$ is a geometric morphism $\mathbf{Set} \to \mathscr{E}$.

Examples of geometric morphisms

- A continuous function $f : X \to Y$ between topological spaces gives rise to a geometric morphism $\mathbf{Sh}(f) : \mathbf{Sh}(X) \to \mathbf{Sh}(Y)$. The direct image $\mathbf{Sh}(f)_*$ sends a sheaf $F \in Ob(\mathbf{Sh}(X))$ to the sheaf $\mathbf{Sh}(f)_*(F)$ defined by $\mathbf{Sh}(f)_*(F)(V) = F(f^{-1}(V))$ for any open subset V of Y. The inverse image $\mathbf{Sh}(f)^*$ acts on étale bundles over Y by sending an étale bundle $p : E \to Y$ to the étale bundle over X obtained by pulling back p along $f : X \to Y$.
- Every Grothendieck topos $\mathscr{E}$ has a unique geometric morphism $\mathscr{E} \to \mathbf{Set}$. The direct image is the global sections functor $\Gamma : \mathscr{E} \to \mathbf{Set}$, sending an object $e \in \mathscr{E}$ to the set $\mathrm{Hom}_{\mathscr{E}}(1_{\mathscr{E}}, e)$, while the inverse image functor $\Delta : \mathbf{Set} \to \mathscr{E}$ sends a set S to the coproduct $\coprod_{s \in S} 1_{\mathscr{E}}$.
- For any site $(\mathscr{C}, J)$, the pair of functors formed by the inclusion $\mathbf{Sh}(\mathscr{C}, J) \hookrightarrow [\mathscr{C}^{\mathrm{op}}, \mathbf{Set}]$ and the associated sheaf functor $a : [\mathscr{C}^{\mathrm{op}}, \mathbf{Set}] \to \mathbf{Sh}(\mathscr{C}, J)$ yields a geometric morphism $i : \mathbf{Sh}(\mathscr{C}, J) \to [\mathscr{C}^{\mathrm{op}}, \mathbf{Set}]$.

Geometric morphisms to $\mathbf{Sh}(\mathscr{C}, J)$ I

Given a cartesian category (i.e. a category with all finite limits) $\mathscr{C}$, a functor $F : \mathscr{C} \to \mathscr{E}$ is said to be cartesian if it preserves finite limits. We shall denote by $\mathbf{Cart}(\mathscr{C}, \mathscr{E})$ the category of cartesian functors $\mathscr{C} \to \mathscr{E}$ and natural transformations between them.

Definition

Let $\mathscr{E}$ be a Grothendieck topos.

- A family $\{f_i : a_i \to a \mid i \in I\}$ of arrows in $\mathscr{E}$ with common codomain is said to be epimorphic if for any pair of arrows $g, h : a \to b$ with domain a, $g = h$ if and only if $g \circ f_i = h \circ f_i$ for all $i \in I$.
- If $(\mathscr{C}, J)$ is a site, a functor $F : \mathscr{C} \to \mathscr{E}$ is said to be J-continuous if it sends J-covering sieves to epimorphic families.

The full subcategory of $\mathbf{Cart}(\mathscr{C}, \mathscr{E})$ on the J-continuous flat functors will be denoted by $\mathbf{Cart}_J(\mathscr{C}, \mathscr{E})$.

Geometric morphisms to $\mathbf{Sh}(\mathscr{C}, J)$ II

Theorem

For any cartesian site $(\mathscr{C}, J)$ and Grothendieck topos $\mathscr{E}$, we have an equivalence of categories

$$\mathbf{Geom}(\mathscr{E}, \mathbf{Sh}(\mathscr{C}, J)) \simeq \mathbf{Cart}_J(\mathscr{C}, \mathscr{E})$$

natural in $\mathscr{E}$.

This equivalence sends a geometric morphism $f : \mathscr{E} \to \mathbf{Sh}(\mathscr{C}, J)$ to the functor given by the composite $f^* \circ l$ of $f^* : [\mathscr{C}^{\text{op}}, \mathbf{Set}] \to \mathscr{E}$ with the canonical functor $l : \mathscr{C} \to \mathbf{Sh}(\mathscr{C}, J)$.

Fact

This theorem generalizes to the case of an arbitrary site $(\mathscr{C}, J)$, replacing the notion of cartesian functor by that of flat functor.

Classifying toposes

In the seventies, thanks to the work of a number of categorical logicians, notably including M. Makkai and G. Reyes, it was discovered that:

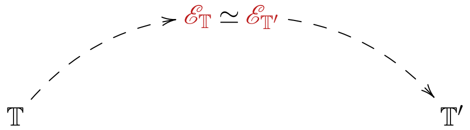
- With any mathematical theory $\mathbb{T}$ (of a very general form - technically speaking, a geometric theory) one can canonically associate a topos $\mathscr{E}_{\mathbb{T}}$, called its classifying topos, which represents its 'semantical core'.
- Two given mathematical theories have the same classifying topos (up to equivalence) if and only if they have the same 'semantical core', that is, if and only if they are indistinguishable from a semantic viewpoint. Two such theories are said to be Morita-equivalent.
- Conversely, any topos is the classifying topos of some theory (in fact, of infinitely many theories).
- A topos can thus be seen as a canonical representative for equivalence classes of theories modulo Morita-equivalence.

Toposes as 'bridges'

- The notion of Morita-equivalence formalizes in many situations the feeling of 'looking at the same thing in different ways', or 'constructing a mathematical object through different methods', which explains its ubiquity in Mathematics.
- In fact, many important dualities and equivalences in Mathematics can be naturally interpreted as arising from Morita-equivalences.
- Any two theories which are bi-interpretable in each other are Morita-equivalent but, very importantly, the converse does not hold.
- Moreover, the notion of Morita-equivalence captures the dynamics inherent to the very concept of mathematical theory; indeed, a mathematical theory alone gives rise to an infinite number of Morita-equivalences.
- Topos theory itself is a primary source of Morita-equivalences. Indeed, different representations of the same topos can be interpreted as Morita-equivalences between different mathematical theories.

Toposes as 'bridges'

- The existence of different theories with the same classifying topos translates, at the technical level, into the existence of different representations for the same topos.
- Topos-theoretic invariants, that is properties of (or construction on) toposes which are invariant with respect to their different representations, can thus be used to transfer information from one theory to another:



- Transfers of information take place by expressing a given invariant in terms of the different representations of the topos.

Toposes as 'bridges'

- Different properties (resp. constructions) arising in the context of theories classified by the same topos come thus to be seen as different *manifestations* of a *unique* property (resp. construction) lying at the topos-theoretic level.

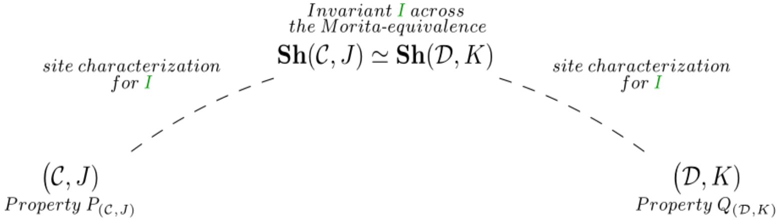
- Every invariant behaves in this context as a 'pair of glasses' enabling to discern some information hidden in the given Morita equivalence; different invariants allow to englighten and transfer different information.

- This methodology is technically feasible because the relationship between a topos and its representations is very natural, enabling us to transfer invariants across different representations (and hence, between different theories) in an effective (though generally non-trivial, and even, in some cases, very complex) way.

The 'bridge-building' technique

- Decks of 'bridges': Morita-equivalences (or more generally morphisms or other kinds of relations between toposes)
- Arches of 'bridges': Characterizations for topos-theoretic invariants in terms of the two different representations

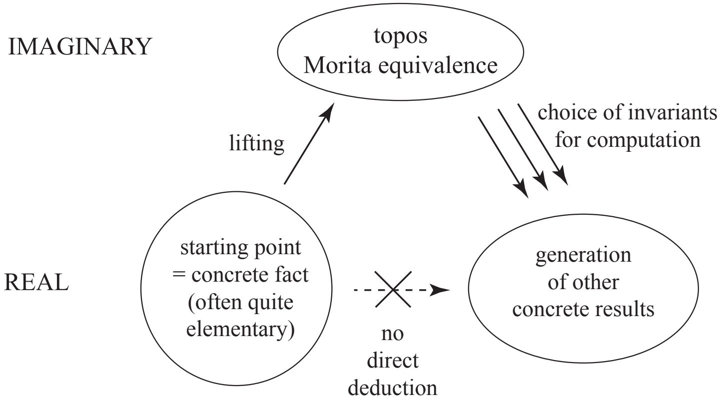
A typical 'bridge' between different site representations for the same topos looks as follows:



This 'bridge' yields a logical equivalence between the 'concrete' properties $P_{(\mathscr{C},J)}$ and $Q_{(\mathscr{D},K)}$, interpreted in this context as manifestations of a unique property I lying at the level of the topos.

A leap into the 'imaginary'

We can schematically represent the way of obtaining concrete results by applying the 'bridge' technique in the form of an ascent followed by a descent between two levels, the 'real' one of concrete mathematics and the 'imaginary' one of toposes:



The duality between 'real' and 'imaginary'

- The passage from a site (or a theory) to the associated topos can be regarded as a sort of 'completion' by the addition of 'imaginaries' (in the model-theoretic sense), which materializes the potential contained in the site (or theory).

- The duality between the (relatively) unstructured world of presentations of theories and the maximally strucured world of toposes is of great relevance as, on the one hand, the 'simplicity' and concreteness of theories or sites makes it easy to manipulate them, while, on the other hand, computations are much easier in the 'imaginary' world of toposes thanks to their very rich internal structure and the fact that invariants live at this level.

A mathematical morphogenesis

- The essential ambiguity given by the fact that any topos is associated in general with an infinite number of theories or different sites allows to study the relations between different theories, and hence the theories themselves, by using toposes as 'bridges' between these different presentations.
- Every topos-theoretic invariant generates a veritable mathematical morphogenesis resulting from its expression in terms of different representations of toposes, which gives rise in general to connections between properties or notions that are completely different and apparently unrelated from each other.
- The mathematical exploration is therefore in a sense 'reversed' since it is guided by the Morita-equivalences and by topos-theoretic invariants, from which one proceeds to extract concrete information on the theories that one wishes to study.