# AGI from the perspective of categorical logic and geometry

YKY

April 25, 2024

## 0 Basics

- The starting point of categorical logic is really the **Curry-Howard isomorphism**, without which you won't be able to understand the sequel.
- Curry-Howard correspondence is the idea of using a mathematical **function** $f : A \to B$ to simulate or implement the process of logic deduction, specifically $A \Rightarrow B$.
- From this perspective, a logic **proposition** $A$ corresponds to the **domain** $A$ of a function. That is, a proposition is akin to something like a **space**.
- Objects in that space are so-called **proof objects**, we use $\blacksquare$ to denote them.
- It may take a while to get used to, but in fact we see this idea in use almost every day: in **neural networks**.
- A neural network maps certain vectors to vectors. Each vector is a proof.
- The space near a positional vector (under some error margin) ought to represent the **same** concept. So we might as well think of the neighborhood space as a logical proposition.
- This way of doing things is really very obvious and natural.

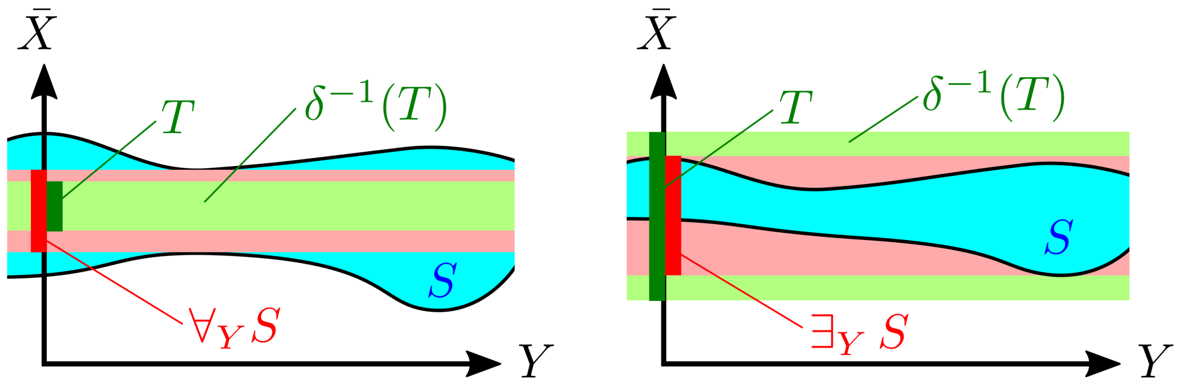
One evidence that this correspondence is on the right track is the truth table of $A \Rightarrow B$, which perfectly matches the cardinalities of the function spaces:

| $A$ | $B$ | $A \Rightarrow B$ | $B^A$ |
|---|---|---|---|
| 0 | 0 | 1 | $0^0 = 1$ |
| 0 | 1 | 1 | $1^0 = 1$ |
| 1 | 0 | 0 | $0^1 = 0$ |
| 1 | 1 | 1 | $1^1 = 1$ |

(1)

The goal of categorical logic is to use categorical tools as much as possible, to **describe** the structure of logic.

- "Propositions = some kind of spaces" generalizes naturally in category theory to "propositions = **objects** in a category".
- We use **products** in category theory to express logic $\wedge$ and $\vee$, and **exponentiations** $B^A$ for $A \Rightarrow B$. The latter are also **morphisms** $A \to B$ in the category.
- $\forall$ and $\exists$ are described as **adjoints** to certain variable-substitution maps. For example in $\forall x.\phi(x, y)$ the quantifier $\forall x$ projects the space of $(x, y)$ down to $(y)$ only, so the resulting expression is no longer about $x$.



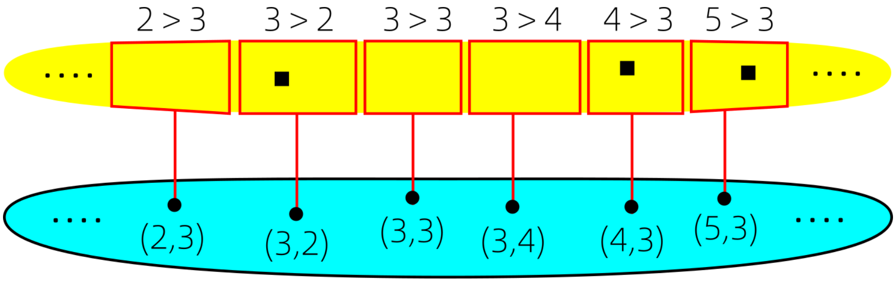
(2)

(This part is a bit complicated, but I have explained it elsewhere. The important thing is to understand the overall concept and not get lost in the details just yet)

①

# 1 Where is GPT?

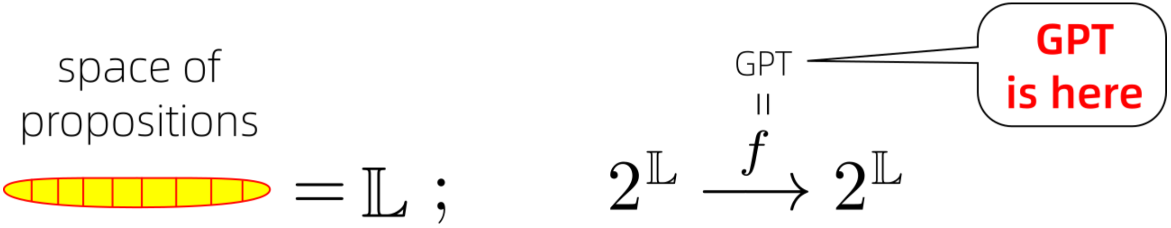
Logic **predicates** are described as **fibrations**, that is, an upper space "indexed" by the base space:



(3)

Here we are considering the binary relation ">", so the indexing space consists of pairs of natural numbers $\mathbb{N} \times \mathbb{N}$. We can form relational propositions $> (a, b)$, and because of Curry-Howard, these propositions are "spaces", ie, yellow squares above. Each square is a proposition, which may or may not have a proof (■).

The union of all the yellow squares above is a **sheaf**, which is the space $\mathbb{L}$ of propositions. **GPT** is a **logic consequence operator** that maps propositions to propositions. But note: GPT is a **set-valued map**. Its domain is not $\mathbb{L}$ but the the power set $2^{\mathbb{L}}$ or $\mathcal{P}(\mathbb{L})$:



(4)

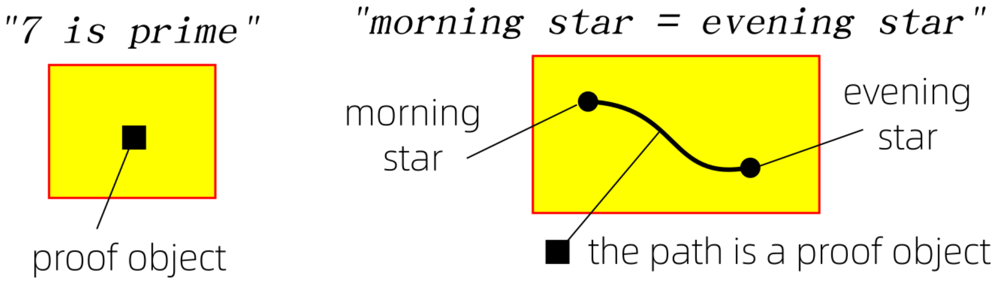
Knowing where GPT fits into the scheme of things, provides some clarity. At least for me, because I'm very familiar with **logic-based AI**, I tend to understand the mathematics from this perspective.

## 2 What is HoTT?

As Curry-Howard suggest that propositions are some kind of “spaces”, can they have topological structure? This is the original idea of **HoTT** (homotopy type theory) proposed by Voevodsky.

From my limited understanding, a proposition either has a proof or not have a proof. If there is a proof, there is no difference between one proof or another. But HoTT posits that they can differ. In the internal space of a proposition (of identity type $a = b$), a **path** is a proof that $a$ and $b$ are regarded as equal.

An example: The ancients regarded Venus as the Morning Star and Evening Star, without knowing that they were actually the same star. This is an example of the difference between intension and extension, which can be handled by **intensional logic**, and can be implemented with **modal logic** and Montague semantics. For details, see this article: https://plato.stanford.edu/entries/logic-intensional/



(5)

For another example, a group can have different **group presentations**, a situation that seems applicable by HoTT.

These spaces that are not path-connected have a **groupoid** structure, with multiple levels such as 1-groupoid, 2-groupoid, ... up to $\infty$-groupoid. I am still learning about these aspects.

As the reader can see, HoTT is concerned with the interior of “truth” (ie, the yellow square), but AGI is mainly concerned with **inference**, which acts on the proposition space (ie, the entire yellow “banana” space).

This is not to say that HoTT is useless for AGI, but its impact is subtle and I cannot judge it yet.

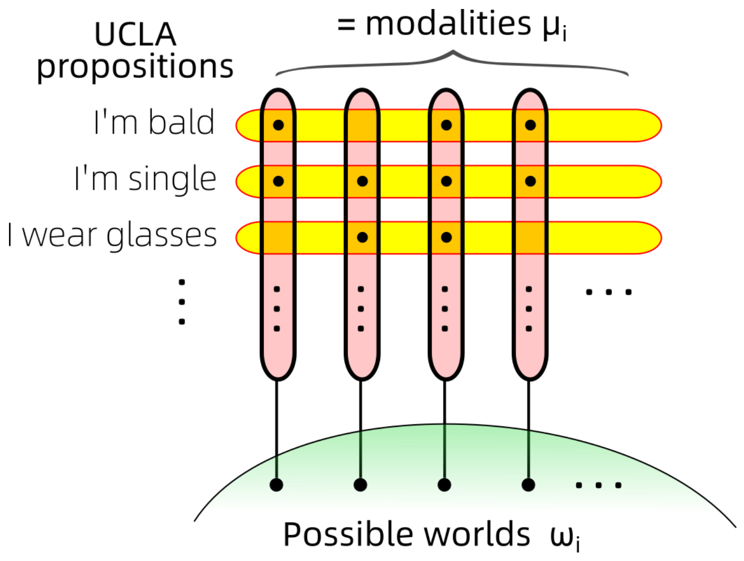
In fact, the **Transformer** has the ability to learn very complex syntactic manipulations, that is, it can implicitly learn the derivation steps of various logics such as modal logic. If so, it seems that we be “bypassed” various special logics without the need to explicitly implement them. But it is also plausible that we by imposing certain logical structural constraints that we can accelerate deep learning. These need to be verified experimentally and are promising research directions.

# 3 Modal logic and possible worlds

Modal logic and possible world semantics are very powerful forms of logic that can resolve many problems in philosophical logic. When I studied symbolic logic AI in the past, I mostly ignored it because it was troublesome to implement a modal logic engine on a computer. As AGI gets closer to reality, I feel the need to take a closer look at modal logics.

The main point of this section is that modal logic can be interpreted by sheaves and toposes. Topos semantics is the "latest" development in logical semantics (though it's actually quite old, haha). It is said that topos semantics can handle almost all the logical semantics we know of (though I don't even know what are the exceptions).

The basic picture is this:



(6)

- The green set underneath are just **indexes** to each possible world, for example $\mathbb{N} = \{1, 2, 3, 4, ....\}$
- Each red "**stalk**" represents a possible world. They form a **fibration** over the base.
- The yellow horizontal bars show the state of truth of each logic **proposition** for each possible world.
- **Modality** = synonym for possible world, a terminology used in John L Bell's book[1]

This basic structure is laid out in the paper: *Topology and Modality: The Topological Interpretation of First-Order Modal Logic* [Awodey & Kishida 2008].

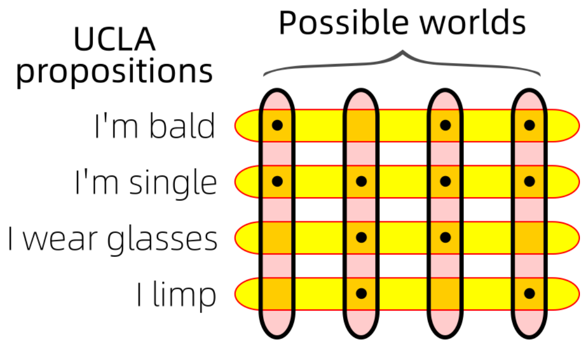
Modal propositional logic can be interpreted by a simple **topological semantics**, while first-order predicate logic can be interpreted by **denotational semantics**. The "product" of these two gives the **sheaf semantics** of first-order modal logic. (Those who have studied programming-language semantics may have heard of denotational semantics.) Let's briefly introduce these two semantics:

## 3.1 Topological semantics of modal logic

This was first proposed by Tarski-McKinsey in 1944.

- A possible world corresponds to an **open set** in a topological space
- A proposition is equivalent to the **set** of possible worlds in which it is true (= set of open sets). These are known as "**UCLA propositions**" after researchers at the University of California
- The modal operators $\Box$ and $\diamond$ correspond to the topological operations of **closure** and **interior** respectively.

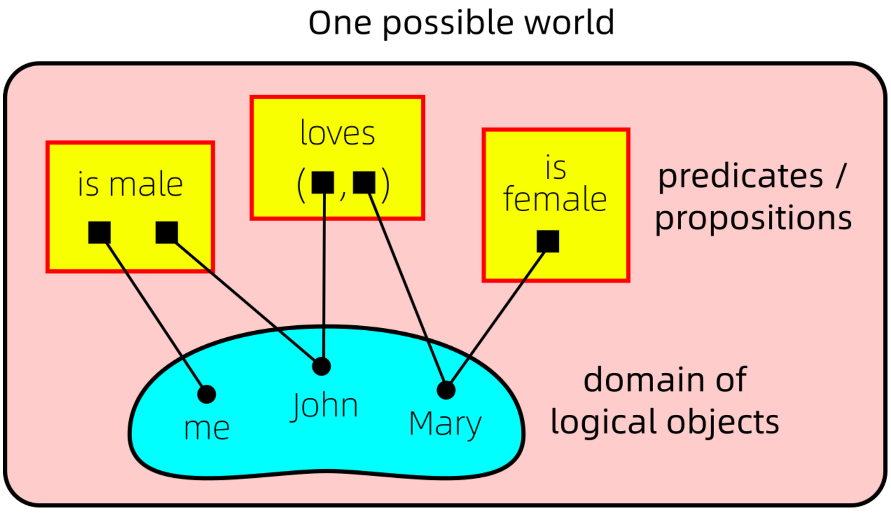
For example, if we only consider the following 4 possible worlds, then "I limp" is not necessarily true, but "I am single" is. The proposition "$\Box$ I am single" has interior = the entire domain, thus the proposition is true.



(7)

## 3.2 Denotational semantics of first-order logic

This part is the standard treatment of model theory of first-order logic that can be found in many textbooks, so I don't want to spend too much time on it. Basically, we use a **structure** $\mathcal{M} = \langle D, R_i, f_j, c_k \rangle$ to interpret logical propositions. $D$ is the **domain** of a logical object, such as { apple, banana, orange, John, Mary }, etc. $R_i$ are **relations** or predicates, such as: "$x$ is a person", "$x$ likes fruit $y$", etc. $f_i$ are **functions** defined in the domain, such as: "$x$'s mother", "$x$'s favorite fruit", etc. $c_i$ are **constants**, such as "$c_1$ = John" etc. The structure $\mathcal{M}$ contains all information needed to **evaluate** the truth or falsity of any proposition defined in that syntax. $\mathcal{M}$ can be regarded as the data of a possible world, where there is only one world.



(8)

(For simplicity, I mixed both unary and binary predicates in the same yellow space)

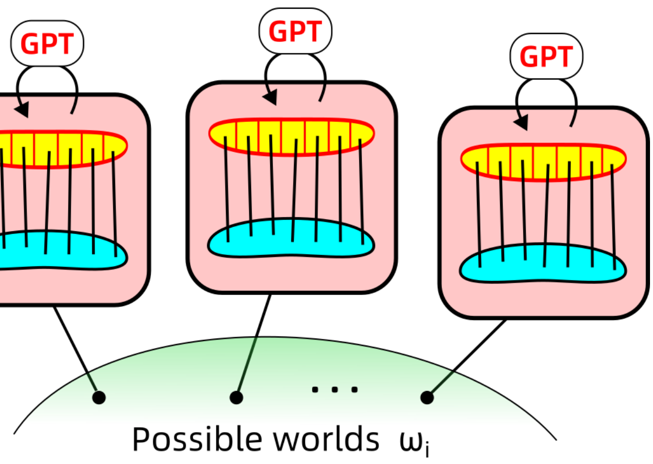
[1] *Toposes and Local Set Theories – An Introduction* [J L Bell 1988]

4

# 4 Sheaf semantics

The so-called combination seems to be a product, or more simply, lining up the fibers in the direction, which is an addition:

Form a sheaf, and sheaf can also be regarded as topos. The overall image is roughly like this:



(9)

- From the perspective of the human brain, its ability to process "possible worlds" is very limited.
- The most classic example is when playing chess, each predicted move is a possible world. In general, people can usually predict only 3-5 moves.
- Each possible world is actually only one proposition different from the current world. It seems that in practice, there is no need to imagine possible worlds as "monsters".
- Possible worlds are generated dynamically when thinking. We cannot quickly train a GPT according to every possible world, so the above GPT's are copies of the same training results.
- If you want to implement modal logic reasoning, you need to expand the function of the "little GPT" above so that it can handle reasoning in multiple possible worlds. The method I think of for the time being is purely to follow the idea of classic logical AI, such as tagging each possible world with a specific proposition, and then calculating
- When the number of possible worlds is small, the concept of topological closure / interior does not seem to be very enlightening. From a computer point of view, continuous space is very "ideal" and is actually very difficult to achieve.

These are quite intuitive. I will look at them in detail when I have time, but now I suddenly feel that this direction may not be too useful...

## 4.1 Fiber products

Note that the categorical product (left) is defined differently from the fiber product (right):

$$\begin{tikzcd}
C \arrow[rrd, bend left, "\pi'_B"] \arrow[rd, dashed, "!"] \arrow[rdd, bend right, "\pi'_A"'] & & \\
& A \times B \arrow[r, "\pi_B"] \arrow[d, "\pi_A"] & B \\
& A &
\end{tikzcd}
\qquad
\begin{tikzcd}
A \times_I B \arrow[r, "\pi_B"] \arrow[d, "\pi_A"] \arrow[rd, dashed, "!"] & B \arrow[d, "g"] \\
A \arrow[r, "f"] & I
\end{tikzcd} \tag{10}$$

Especially note that in the fiber product, $\pi_A \circ f = ! = \pi_B \circ g$ as the diagram commutes. The upshot is that, in the fiber product, the projection from the product along $A$ or $B$ to $I$ gives the same results. For example, we can have the pair $(\text{John}_1, \text{Mary}_1)$ in the fiber product, meaning John from World 1 and Mary from World 1. But we cannot have $(\text{John}_3, \text{Mary}_1)$ in the fiber product because they exist in **different worlds**. Whereas in the categorical (Cartesian) product this is possible.

## 4.2 Sheaf semantics

- In the topological interpretation, the pre-order structure of $\mathbf{K}$ is forgotten, and $\Box$ is interpreted as the **interior** operation. I'm not sure if the two interpretations coincide. But the interior interpretation works because it satisfies **syntactic** axioms of $\Box$.
- In frame-valued logic, the truth value of a proposition is an element of a **Heyting algebra** representing possible worlds, instead of the classical $\{\top, \bot\}$.
- By the method of **UCLA propositions**, Kripke interpretations and frame-valued interpretations can be made to coincide, except that in the Kripke interpretation we have the extra information that each frame element belongs to a world $k$.
-

⑤

# 5 Algebraic geometry

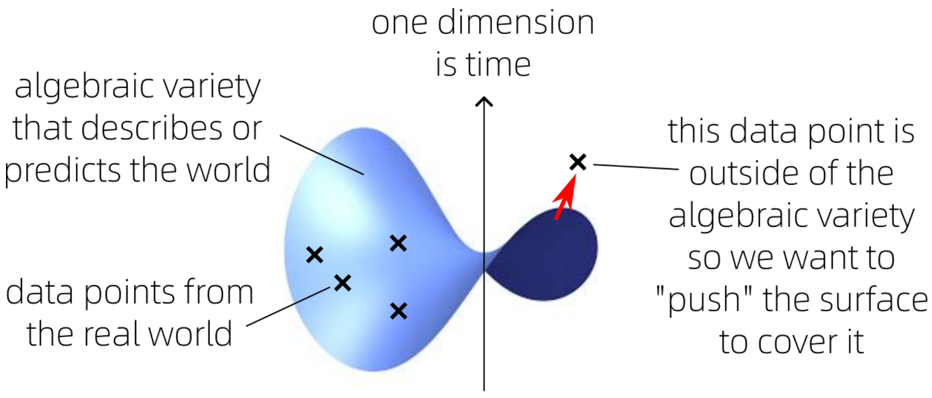
The fundamental duality in algebraic geometry is:

$$\left\{ \begin{array}{c} \text{spaces, or} \\ \text{varieties} \end{array} \right\} \longleftrightarrow \left\{ \begin{array}{c} \text{commutative} \\ \mathbf{k}\text{-algebras} \end{array} \right\} \tag{11}$$

within this correspondence, we have:

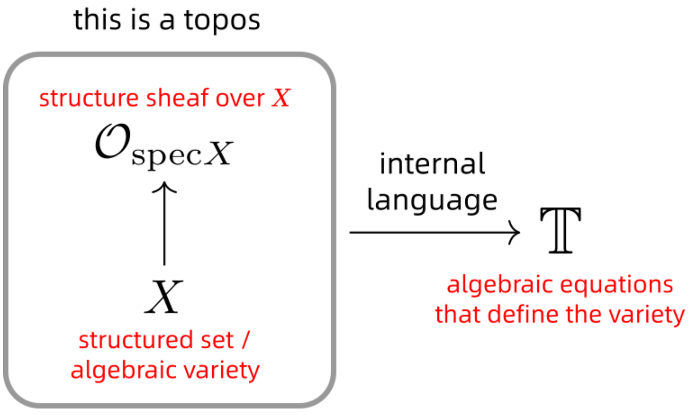
$$\left\{ \begin{array}{c} \text{points in} \\ \text{geometry} \end{array} \right\} \longleftrightarrow \left\{ \text{ prime ideals } \right\} \tag{12}$$

My visualization of the "Grothendieck picture" of AGI is like this:



(13)

$$\begin{tikzcd} \underset{\text{classifying topos}}{\mathcal{E}_{\mathbb{T}}} \arrow[r, shift left, "\text{internal language}"] & \underset{\text{theory}}{\mathbb{T}} \arrow[l, shift left] \end{tikzcd} \tag{14}$$



(15)

In the Grothendieck-algebraic approach, logic formulas correspond to (ideals in) commutative algebras. Crucially, such an algebra allows to define **algebraic equations** whose **zero-sets** define geometric **spaces**. In contrast, in the Curry-Howard-categorical approach, logic formulas correspond to functional mappings or morphisms, one can form limits via categorical means (for example, the unification algorithm in logic can be implemented as **equalizers** categorically), but there is no corresponding notion of algebraic equations as far as I can see. For example, a logic formula such as:

$$\mathrm{father}(X, Y) \wedge \mathrm{father}(Y, Z) \rightarrow \text{grand-father}(X, Z) \tag{16}$$

creates a new proposition (its conclusion), which is a type-theoretic space. As we have more and more logic rules, we build up more and more of such types (spaces). In contrast, in the Grothendieck picture, one starts with the entire affine space $\mathbb{A}^n$ or projective space $\mathbb{P}^n$, while each introduction of a logic rule cuts the space smaller (by intersection).

Problem 1: We don't know how to "lift" Boolean algebra to predicate logic to complete the Grothendieck picture.

Problem 2: The time dimension $t$ is part of the geometric space.

It is fascinating to compare the Grothendieck approach to algebraic logic vs the Curry-Howard tradition, and Einstein's general relativity vs quantum mechanics. In both cases, one side is developed by a Jew with relatively few collaborators, and the other side is developed by collaborations of mostly white men.

# 6 What's the use of all these to AGI?

We want to **accelerate** AGI algorithms, which can be basically divided into **inference** and **learning**, of which learning is the computationally higher-complexity part, in other words the **bottleneck**.

In order to accelerate, there is basically only one principled approach: namely via **inductive bias**, based on the "No Free Lunch" theorem.

- Products
- Exponentiation
- Predicates as fibration
- Quantifiers as adjunctions
- Possible worlds as sheaves
-