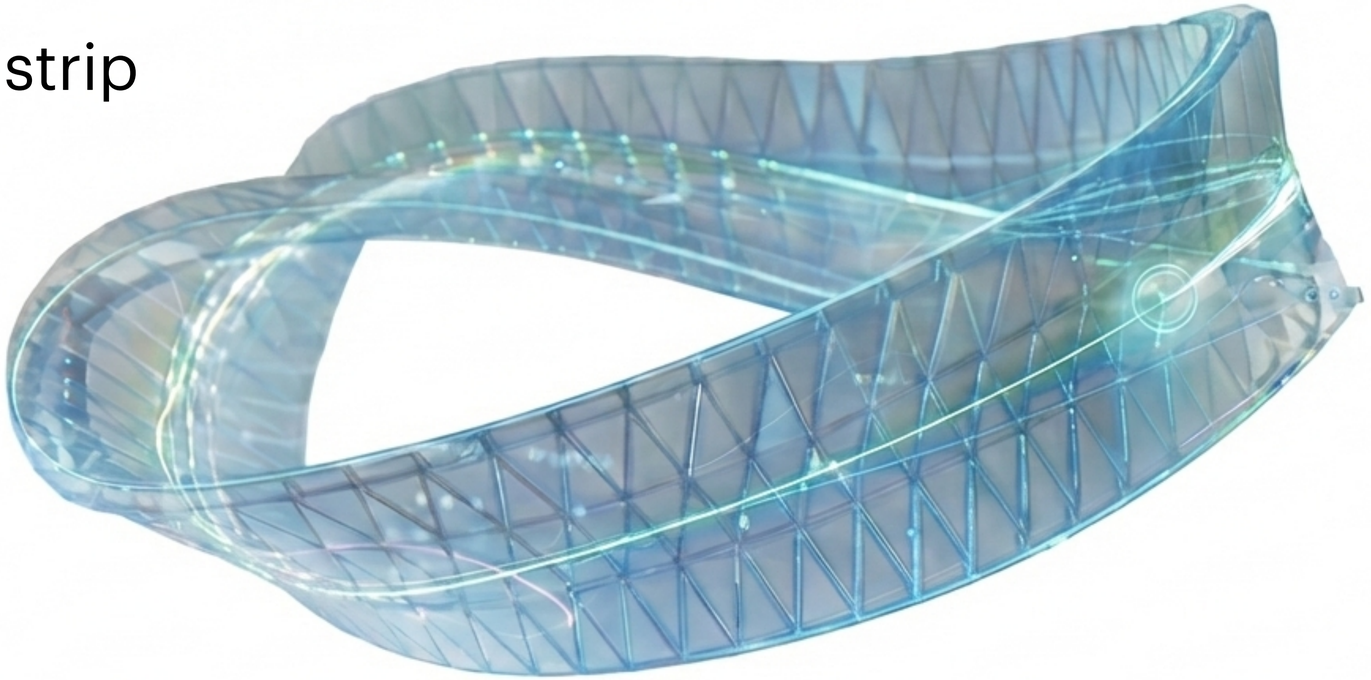
unLecture

The *Fault* in Our Intelligence

Aalok Thakkar

Avengers: Endgame (2019)

Möbius strip



What is a mobius strip?

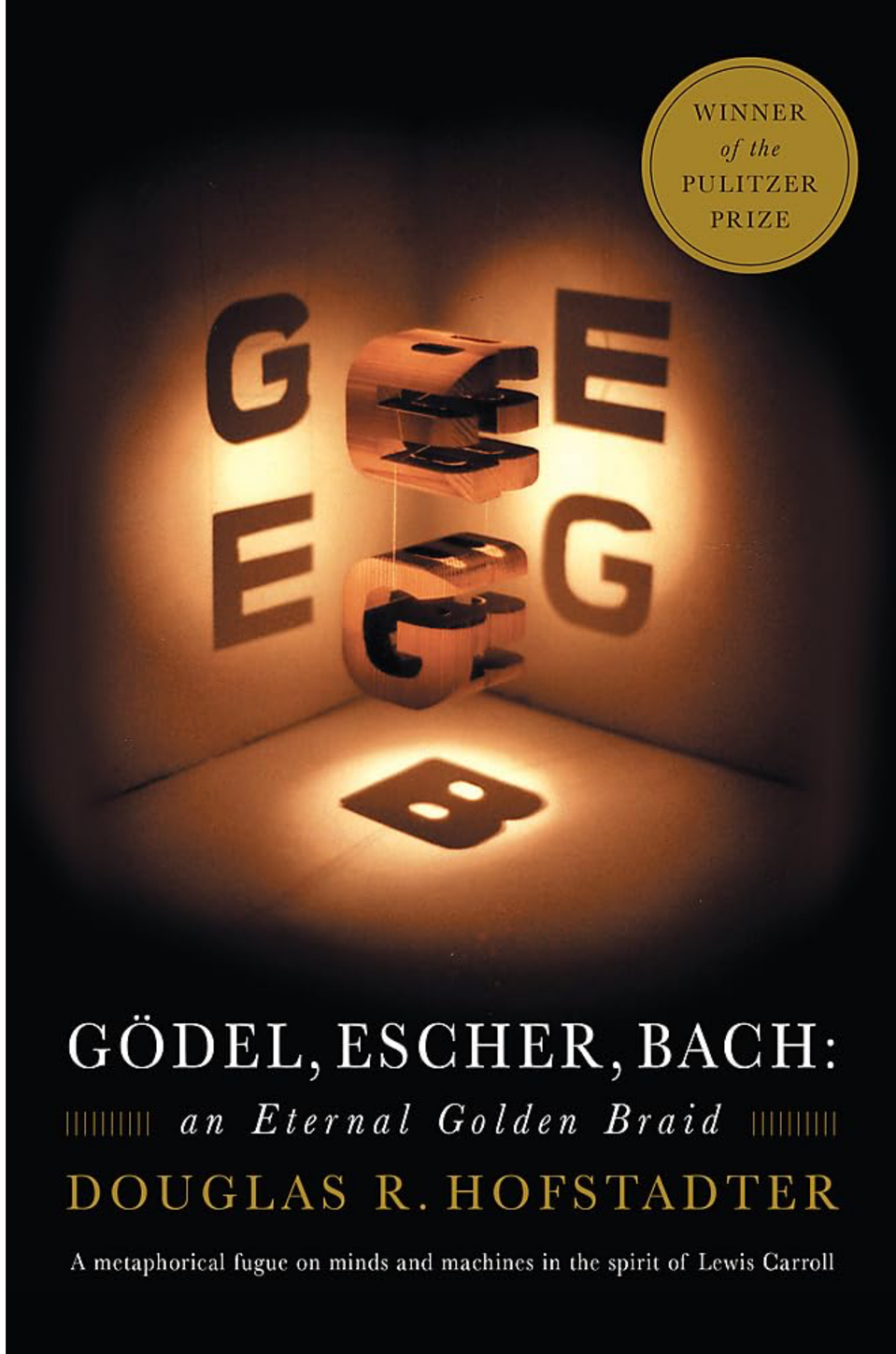
A **Möbius strip** is a surface with a very strange but precise property: it has **only one side and one boundary**.

The Möbius strip is the unique (up to homeomorphism) compact, connected, non-orientable 2-dimensional manifold with a boundary whose boundary is homeomorphic to S^1.

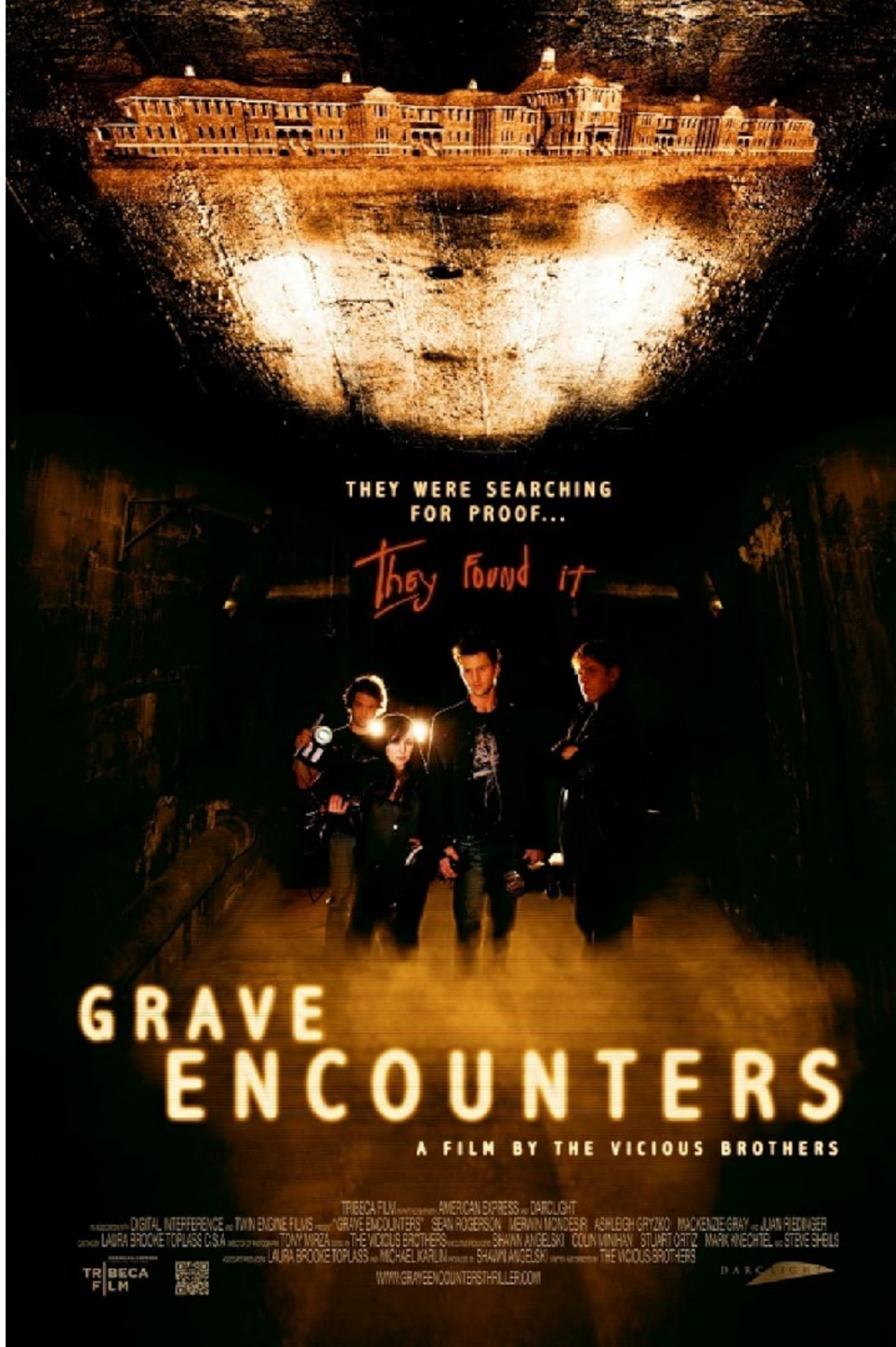

It is a *strange loop*.

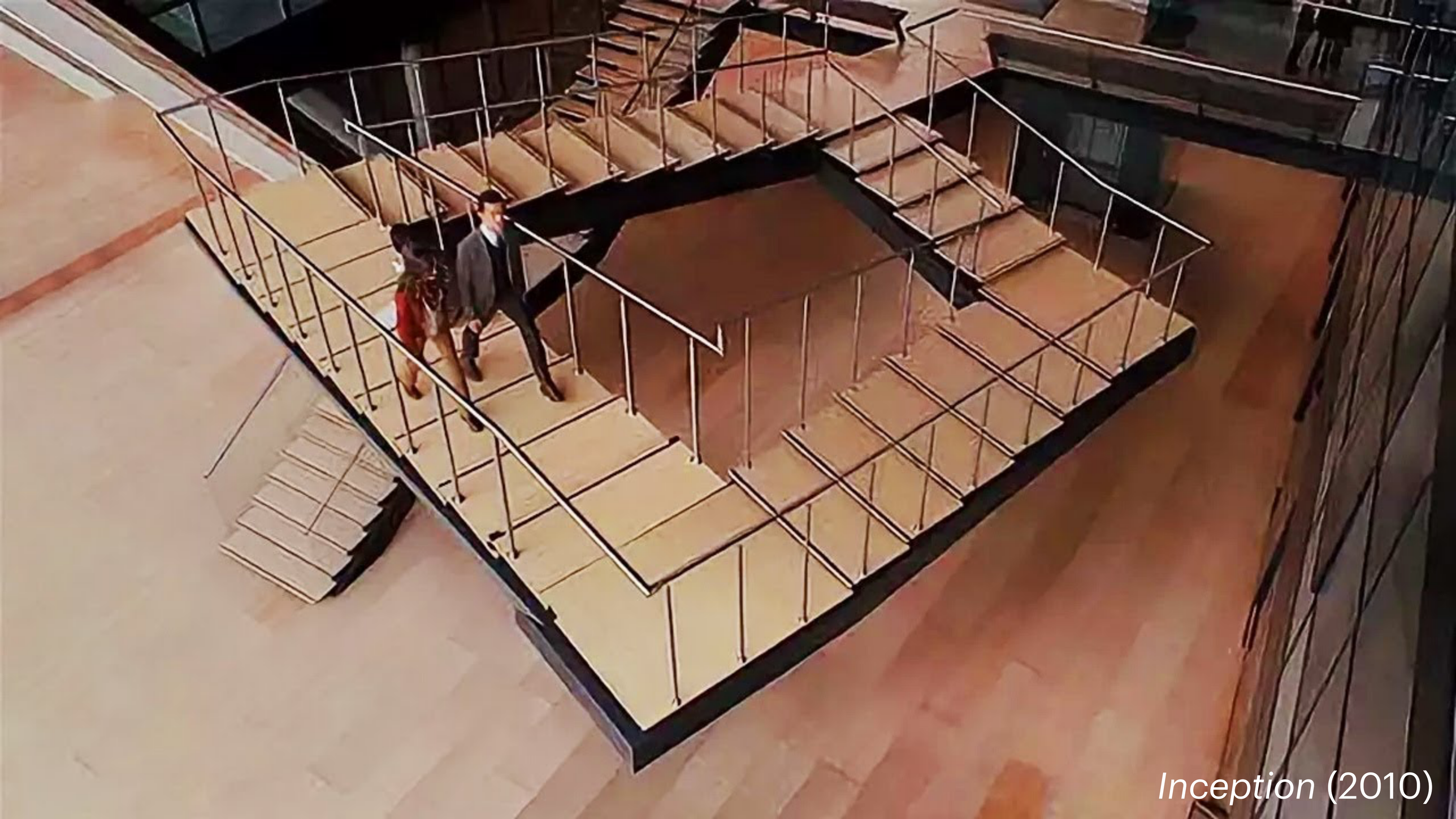
A strange loop: By moving “upward” or “forward” through a system’s levels, one unexpectedly returns to the starting point.

And strange loops are everywhere!

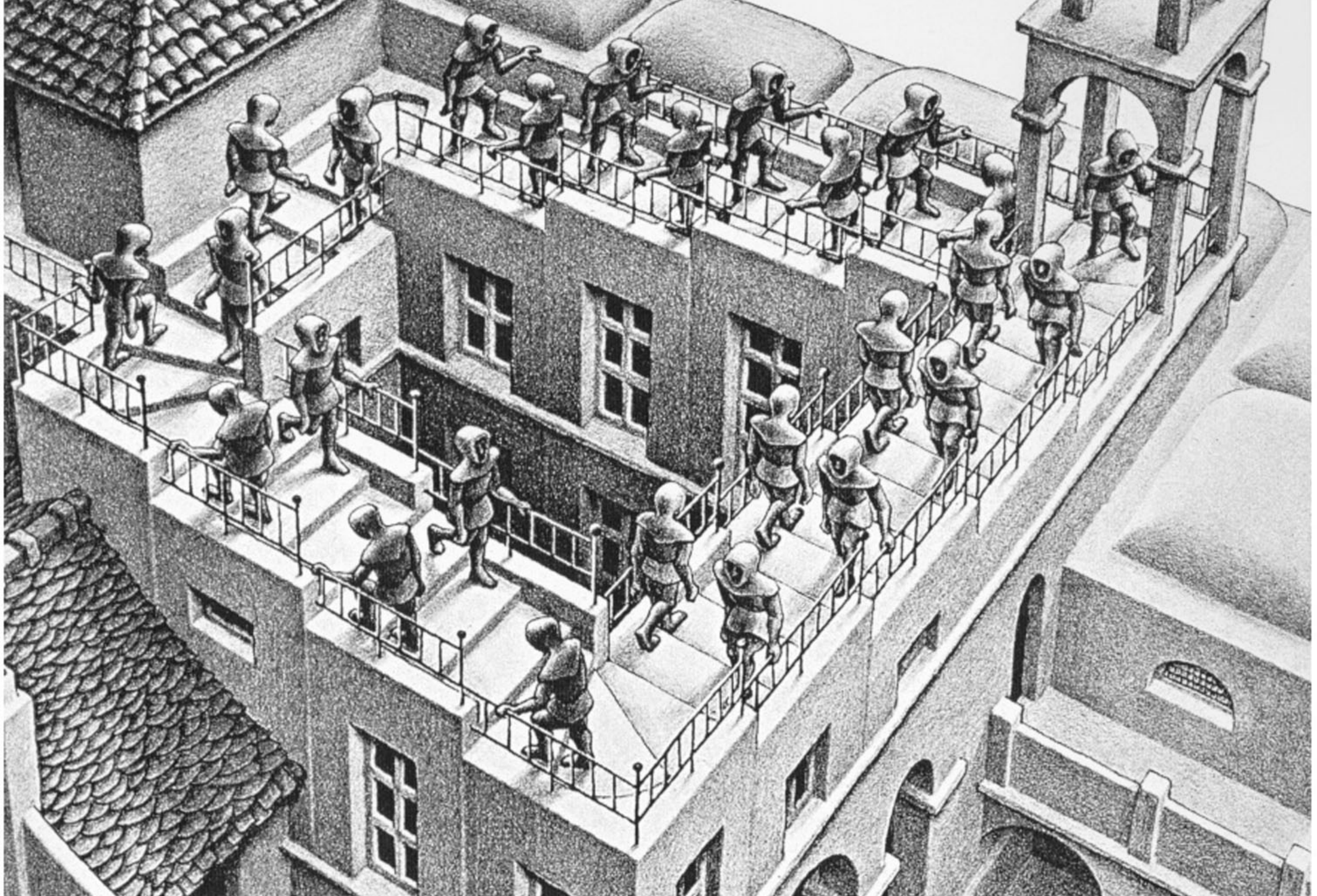


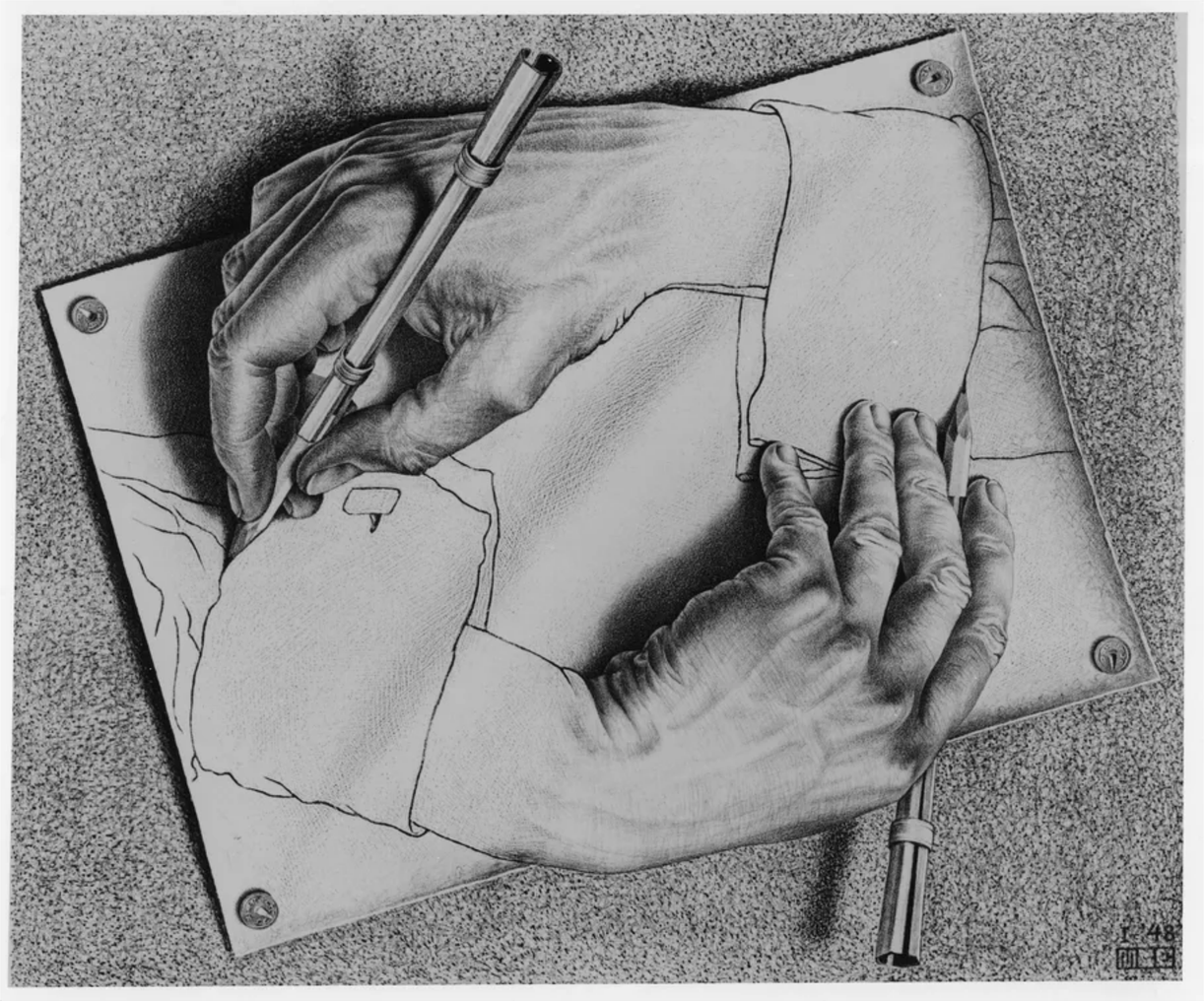
Topological Imprisonment: Characters repeatedly “walk out” and re-enter the same space. Locally, they make smart choices. But on the whole they are just stuck!



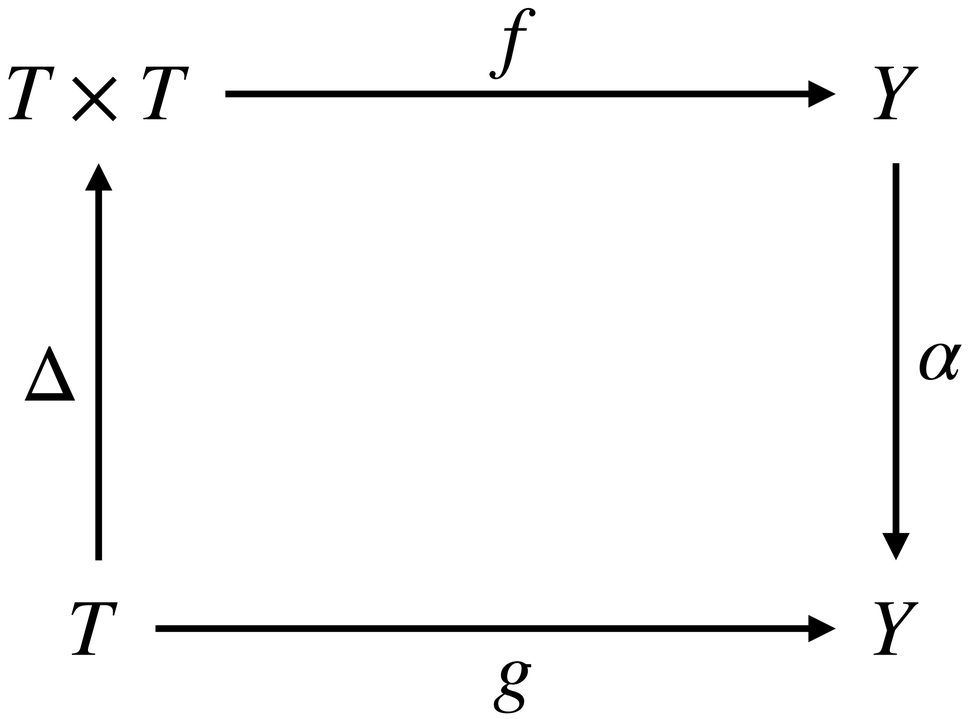


Inception (2010)





What is *really* happening?

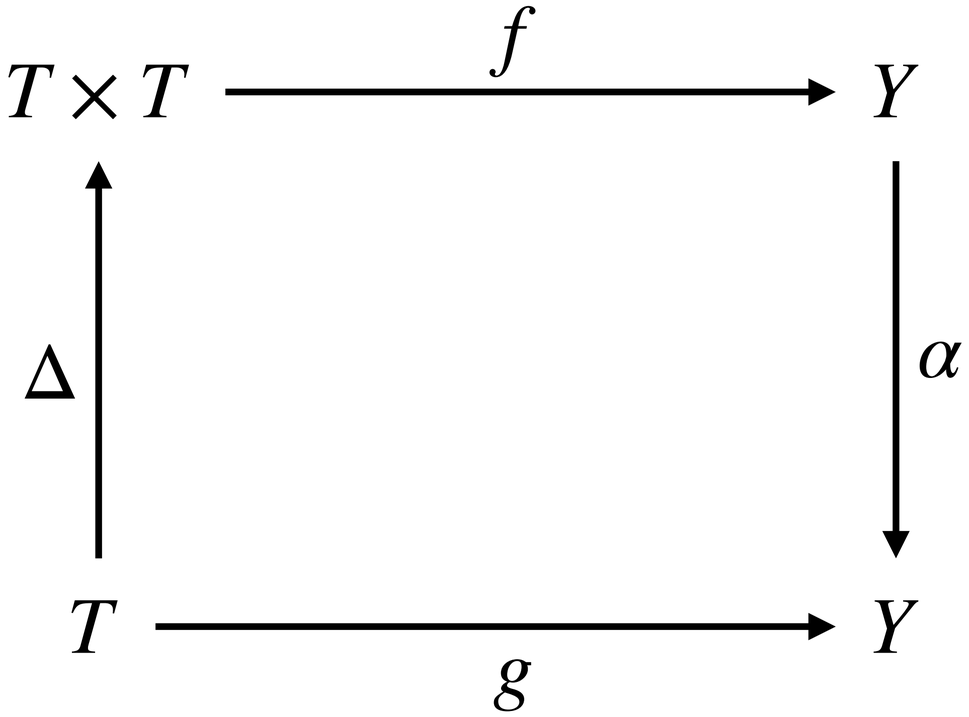


Consider T be a set of *objects*, and Y to be a set of *values*.

Let f evaluate object $t \in T$ on input $x \in T$.

Then α negates the output of $f(t, x)$.

T = Set of well-formed sentences



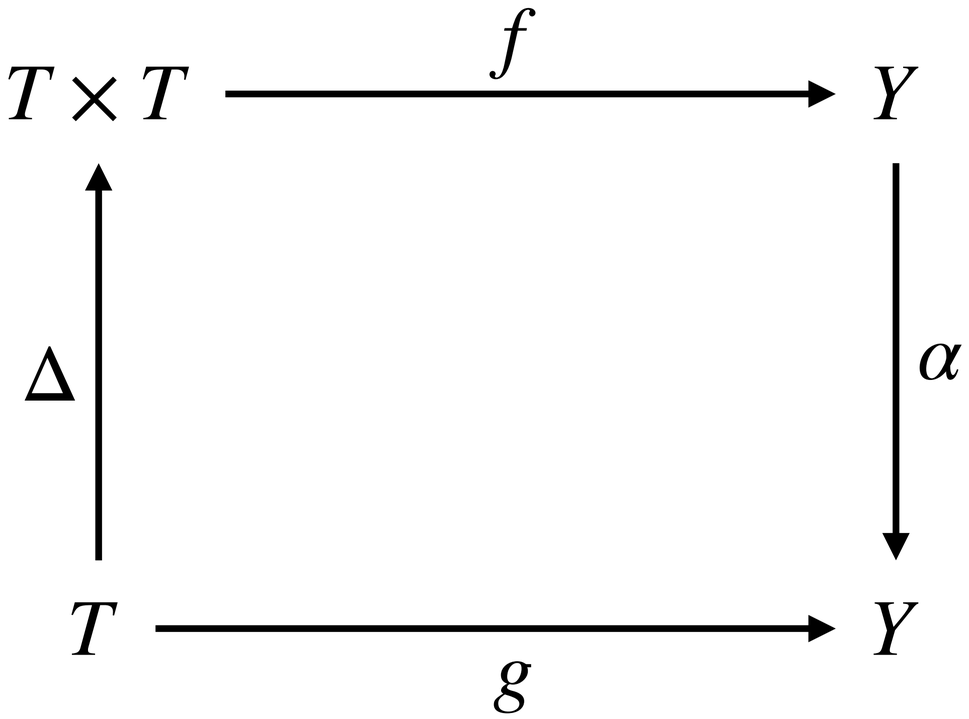
$Y = \{\text{true, false}\}$

Consider T be a set of *objects*, and Y to be a set of *values*.

Let f evaluate object $t \in T$ on input $x \in T$.

Then α negates the output of $f(t, x)$.

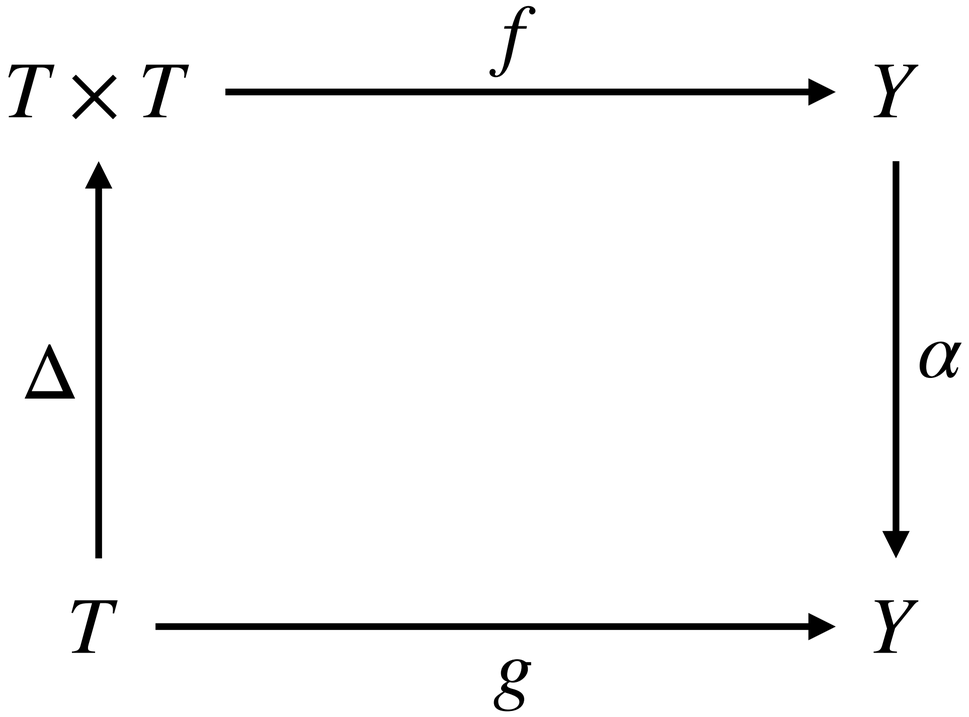
T = Set of well-formed sentences



$Y = \{\text{true, false}\}$

$f(t, x)$ = “sentence t is true when talking about x”

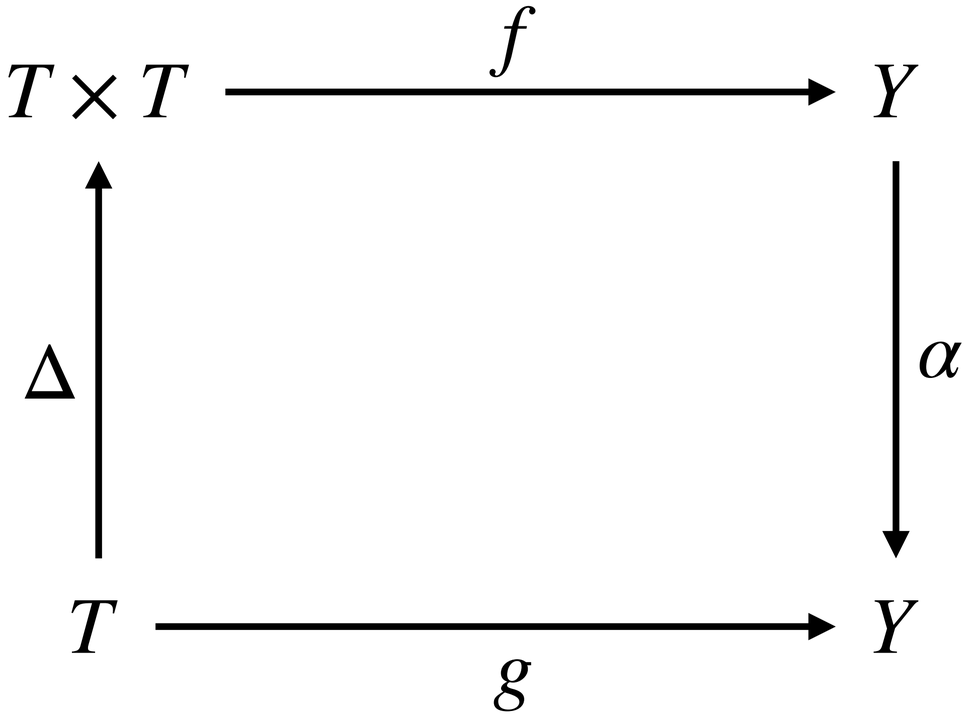
T = Set of well-formed sentences



$Y = \{\text{true, false}\}$

$f(\Delta(t))$ = “sentence t is true when talking about t”

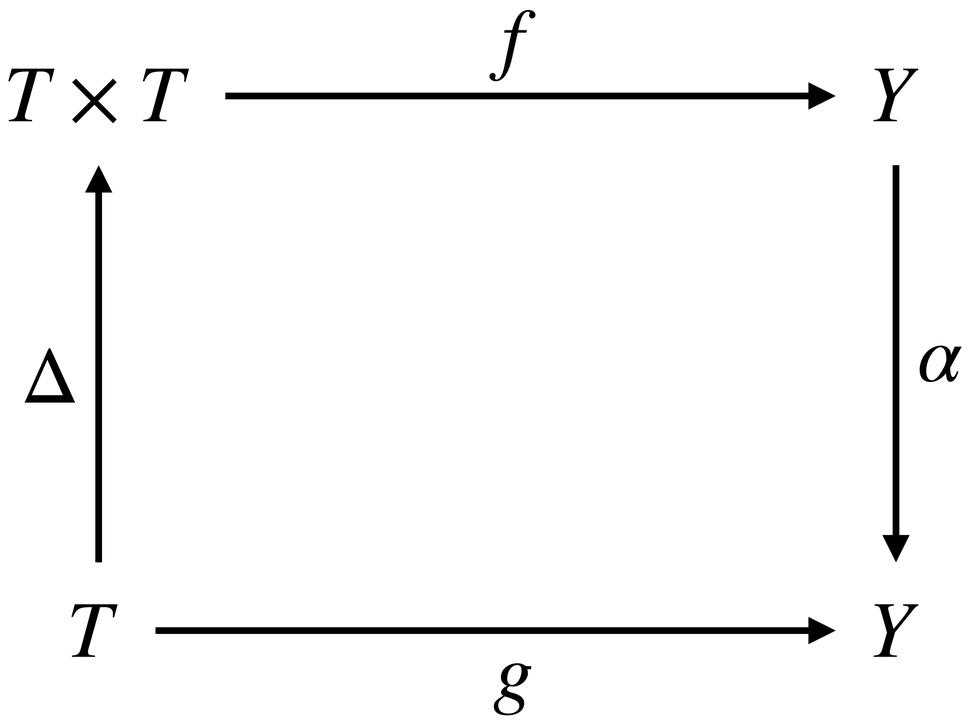
T = Set of well-formed sentences



$Y = \{\text{true, false}\}$

$\alpha\left(f(\Delta(t))\right)$ = “sentence t is **not** true when talking about t”

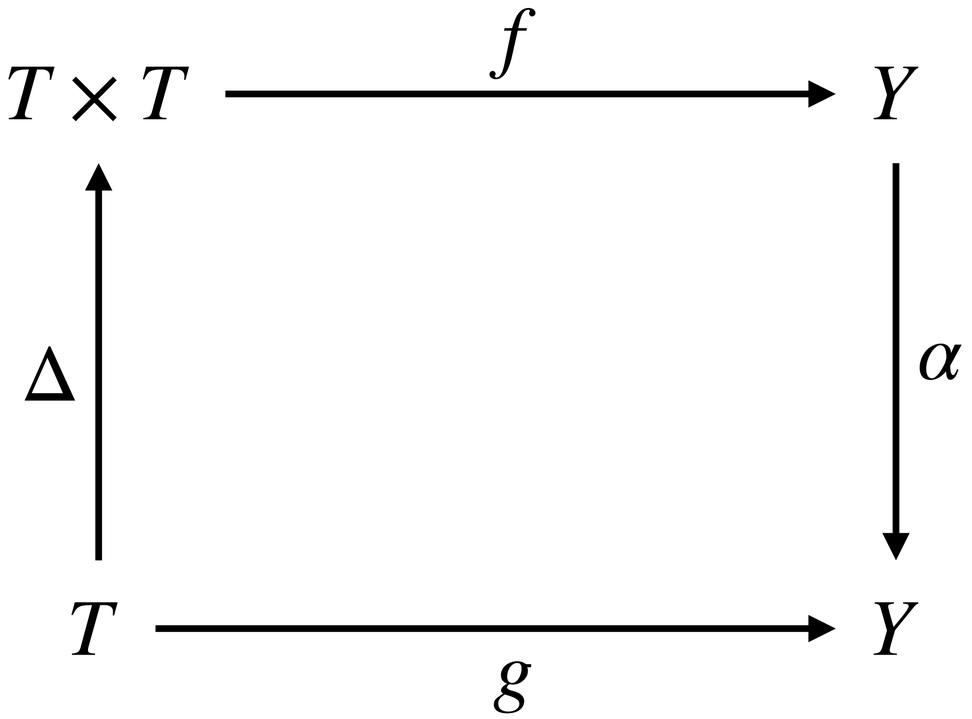
T = Set of well-formed sentences



$Y = \{\text{true, false}\}$

$g(t)$ = “sentence t is **not** true when talking about itself”

T = Set of well-formed sentences

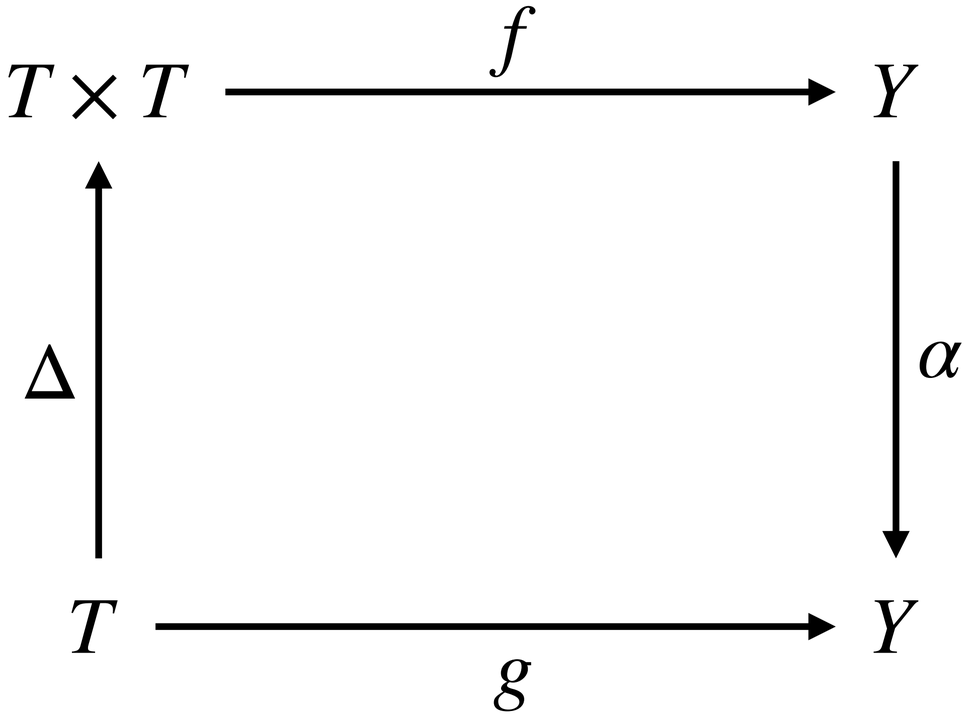


$Y = \{\text{true, false}\}$

$g(t)$ = “sentence t is **not** true when talking about itself”

t = ”This sentence is not true when talking about itself”

T = Set of well-formed sentences



$Y = \{\text{true, false}\}$

$g(t)$ = “sentence t is **not** true when talking about itself”

t = ”This sentence is false”

Every single word here makes sense. The sentence is grammatically correct. It also some meaning.

But it is paradoxical.

$t =$ "This sentence is false"

Grelling-Nelson Paradox

short
English
adjectival
pentasyllabic

autological

long
German
palindrome
monosyllabic

heterological

Grelling-Nelson Paradox

short
English
adjectival
pentasyllabic

autological

long
German
palindrome
monosyllabic

heterological

रंगी को नारंगी कहे , खरे दूध को खोया
चलती को गाड़ी कहे , देख कबीरा रोया!

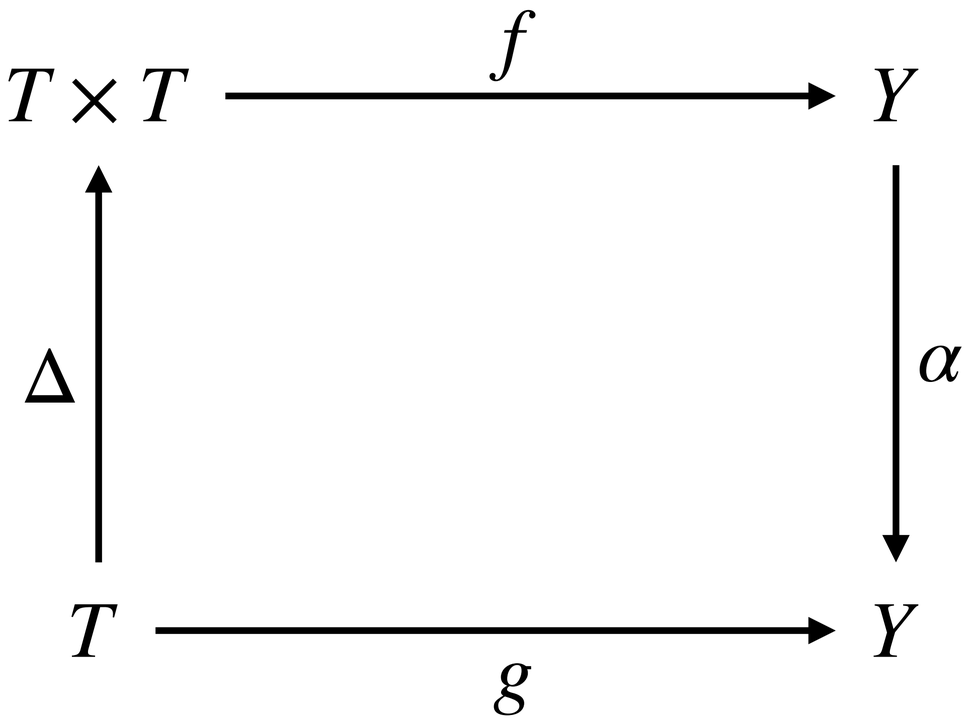
Grelling-Nelson Paradox

short
English
adjectival
pentasyllabic

autological

long
German
palindrome
monosyllabic

heterological?



Self reference +

Negation

Cantor's theorem

Russell's paradox

Gödel's incompleteness theorem

Tarski's undefinability theorem

Turing's proof

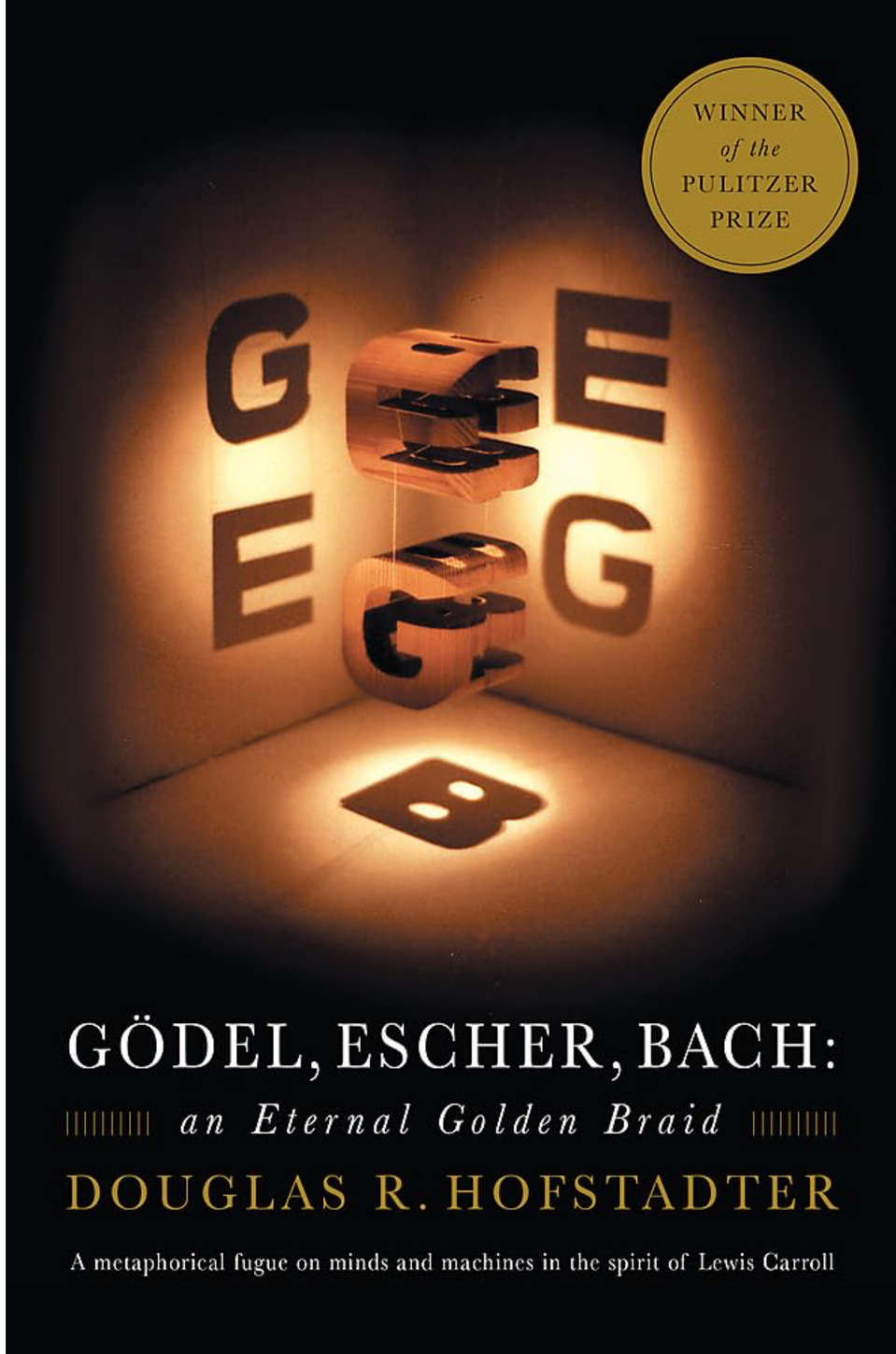
Löb's paradox

Roger's fixed-point theorem

Rice's theorem

A strange loop: By moving “upward” or “forward” through a system’s levels, one unexpectedly returns to the starting point.

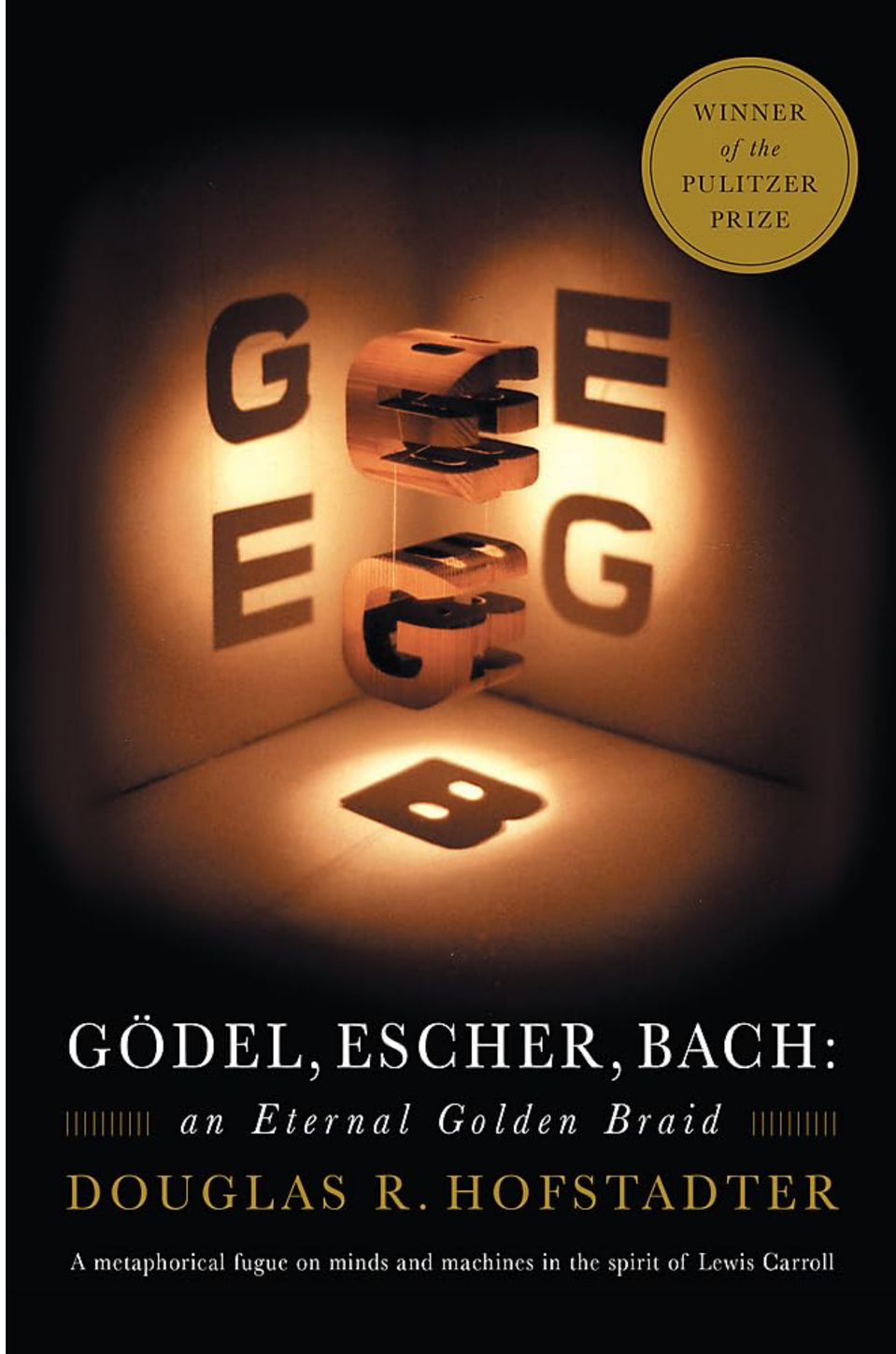
And strange loops are everywhere!



A strange loop: By moving “upward” or “forward” through a system’s levels, one unexpectedly returns to the starting point.

And, strange loops make sense *locally* but we confuse us at a higher level.

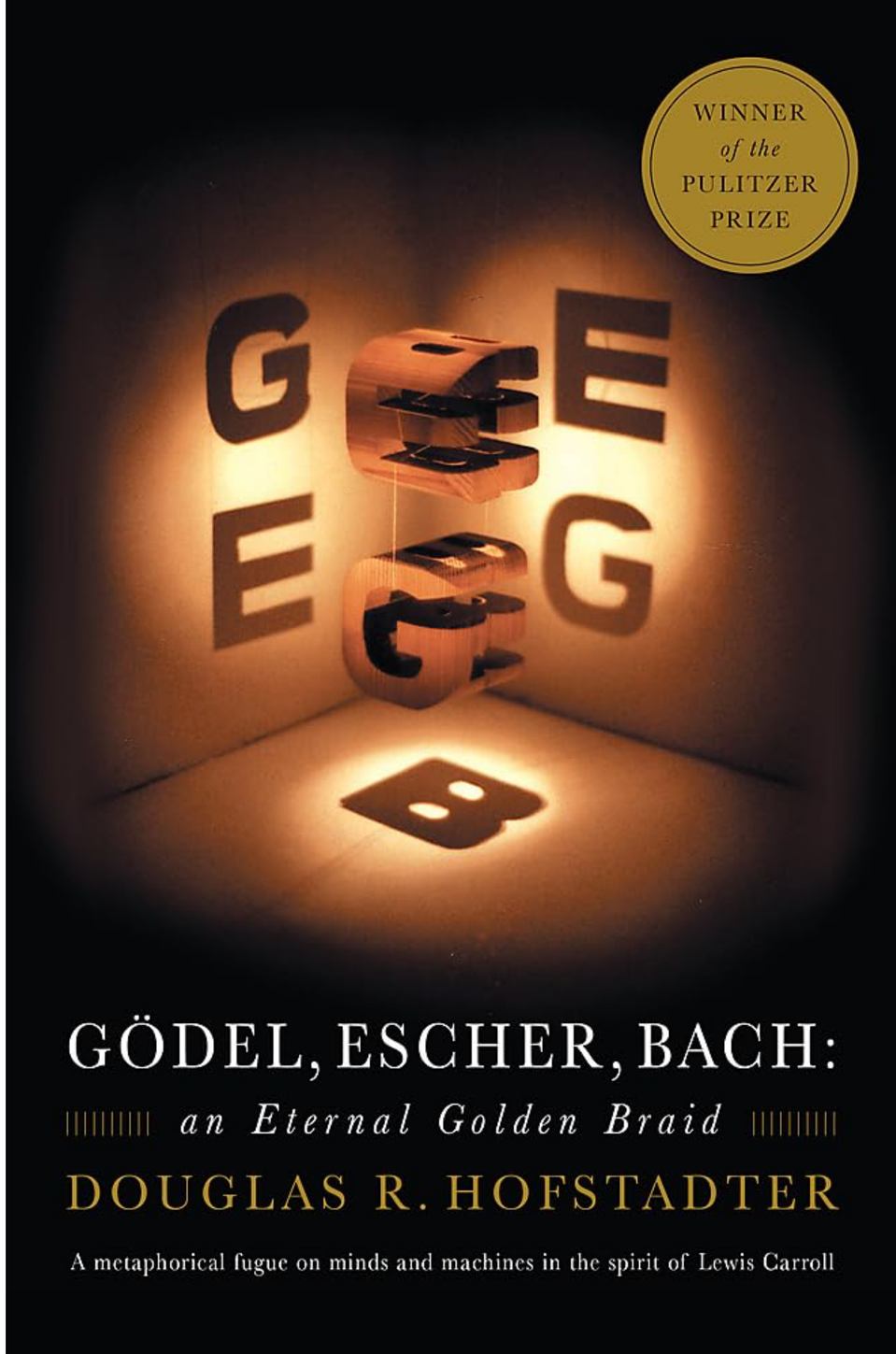
$t =$ ”This sentence is false”



A strange loop: By moving “upward” or “forward” through a system’s levels, one unexpectedly returns to the starting point.

And, strange loops make sense *locally* but we confuse us at a higher level.

My claim: this higher level is *intelligence*.

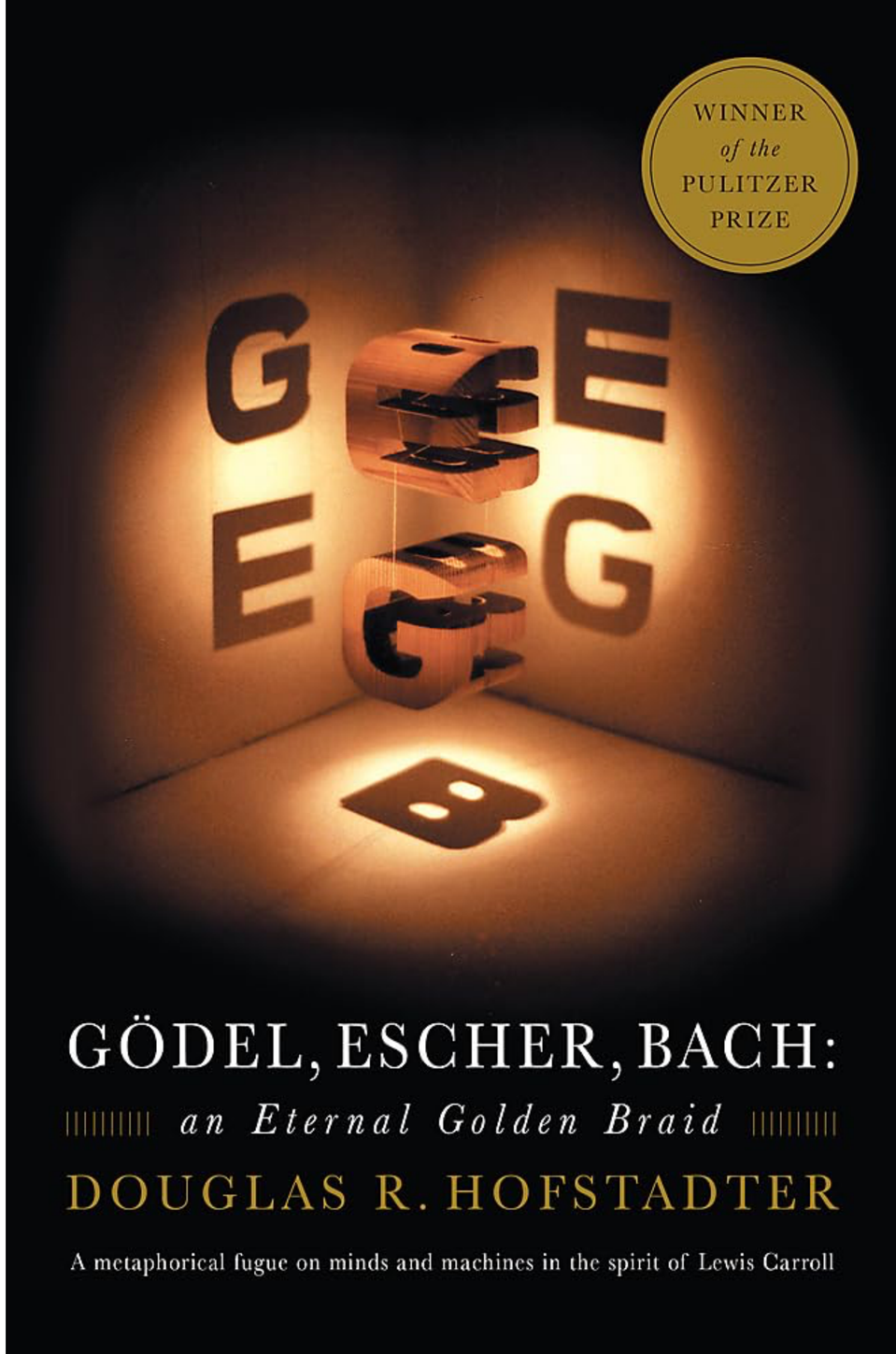


Can AI think at this higher level?

A strange loop: By moving “upward” or “forward” through a system’s levels, one unexpectedly returns to the starting point.

And, strange loops make sense *locally* but we confuse us at a higher level.

My claim: this higher level is *intelligence*.




Time to make
strange loops

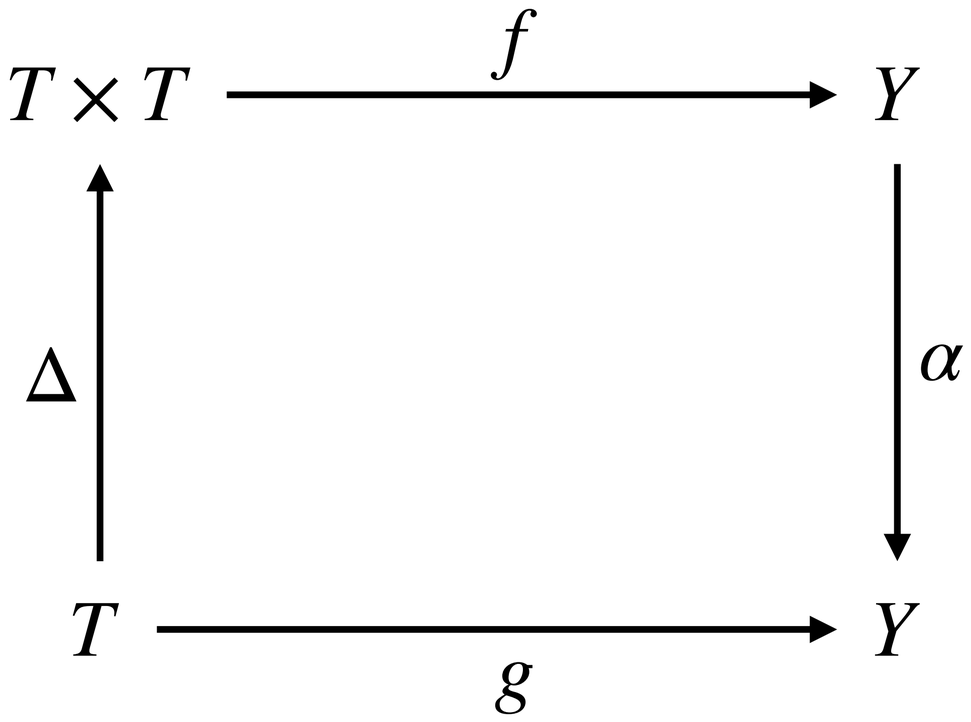
Excercise 1: Line in the middle.

Exercise 2: Trace the border.

Exercise 3: Time to cut.

Exercise 4: Double twist.

Exercise 5: Triple cut.



Does AI understand Möbius strips?

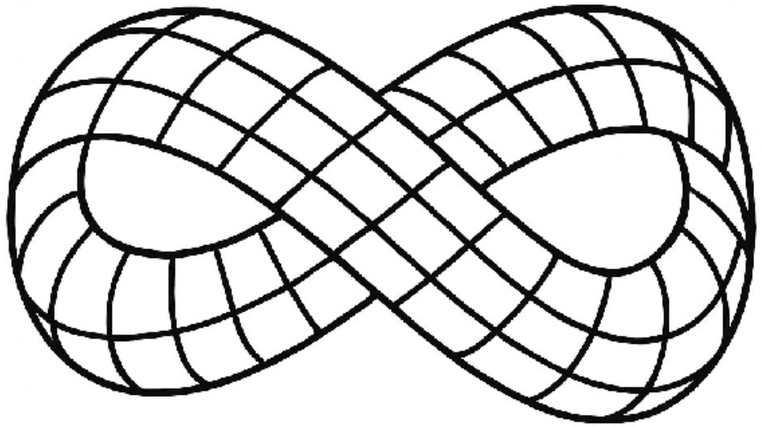
Does AI understand Möbius strips?

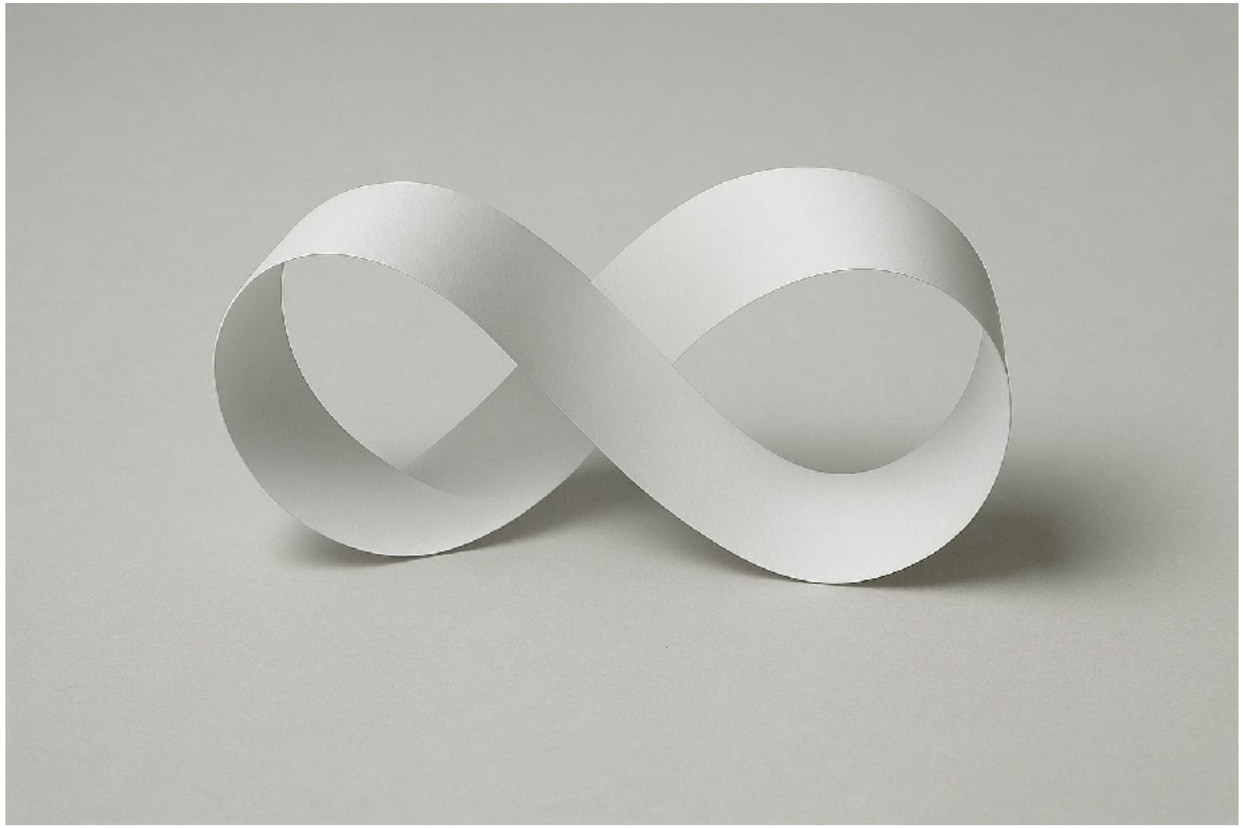
Nano Banana Pro

Does AI understand Möbius strips?

Supergrok

Does AI understand Möbius strips?





GPT Image 1.5

After a lot of prompting I got:

You should definitely try!

What just happened?

AI tools can generate good images.
All surfaces surfaces look alright.

They generate essays where each sentence kind of makes sense.

They can write proofs that feel mathematically correct.

But they miss the *high-level* features very often.

Why?

~~The model just isn’t trained well yet.~~

Image-generation models do not build start with a skeleton, adding arms, then hands, and finally fingers.

They generate images locally, one patch at a time, optimizing for visual plausibility at each step.

Why?

~~The model just isn’t trained well yet.~~

Language models do not begin with central argument, a structure, and then choose words to express it.

They generate language locally, one token at a time, optimizing for what looks plausible given the context.

Artificial ~~Intelligence~~ *Fluency*

How does artificial fluency work?

Representation

Embeddings

Attention

How does artificial fluency work?

Representation

Embeddings

Attention

Input: *Greater Kailash*

Model breaks it down:

[“great”, “er”, “kail”, “ash”]

How does artificial fluency work?

Representation

Input: *Greater Kailash*

Model breaks it down:

[“great”, “er”, “kail”, “ash”]

Embeddings

Each token
becomes a vector.

‘most' + ‘great' ≈ ‘greatest'

Attention

How does artificial fluency work?

Representation

Input: *Greater Kailash*

Model breaks it down:

[“great”, “er”, “kail”, “ash”]

Embeddings

Each token becomes a vector.

Attention

Given all the words we've seen, which ones matter for predicting the next one?

How does artificial fluency work?

Representation

Input: *Greater Kailash*

Model breaks it down:

[“great”, “er”, “kail”, “ash”]

Embeddings

Each token becomes a vector.

Attention

Given all the words we've seen, which ones matter for predicting the next one?

And if you don't pay enough attention...

Attention

Given all the words we've seen, which ones matter for predicting the next one?

And if you don’t pay enough attention...

Attention

Given all the words we've seen, which ones matter for predicting the next one?

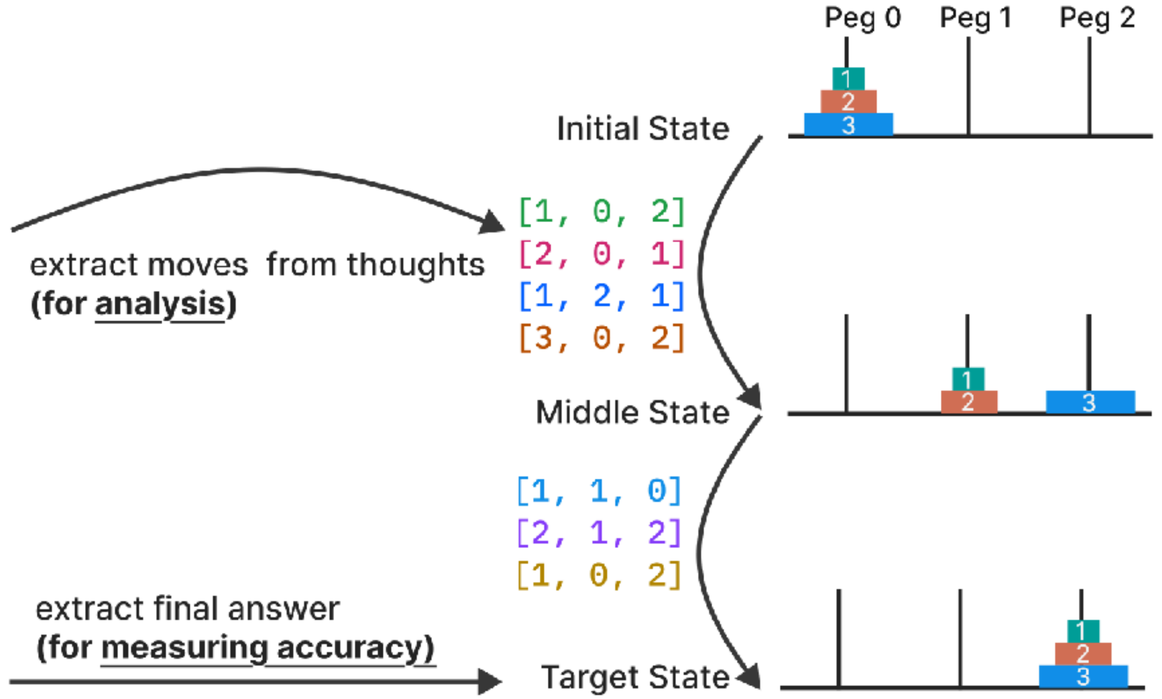
LLM Response

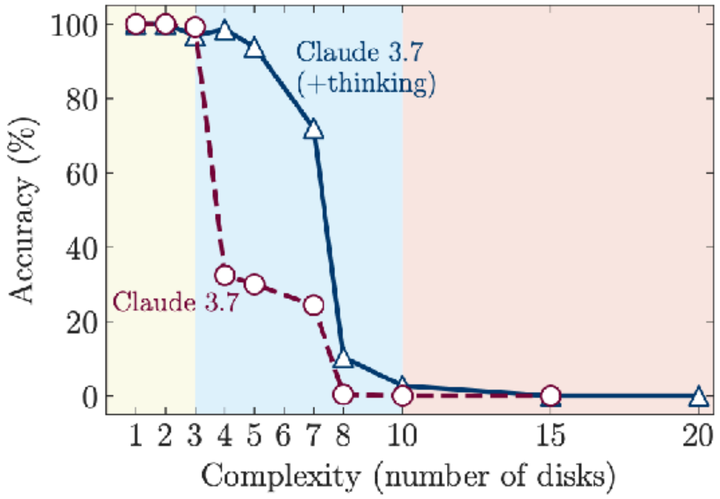
```
<think>
Move disk 1 from peg 0 to peg 2 ...
moves = [
  [1, 0, 2],
  [2, 0, 1],
  [1, 2, 1],
  [3, 0, 2],
  [1, 1, 0],
  [2, 1, 2],
  [1, 0, 2],
]
Let me double-check this...
</think>
```

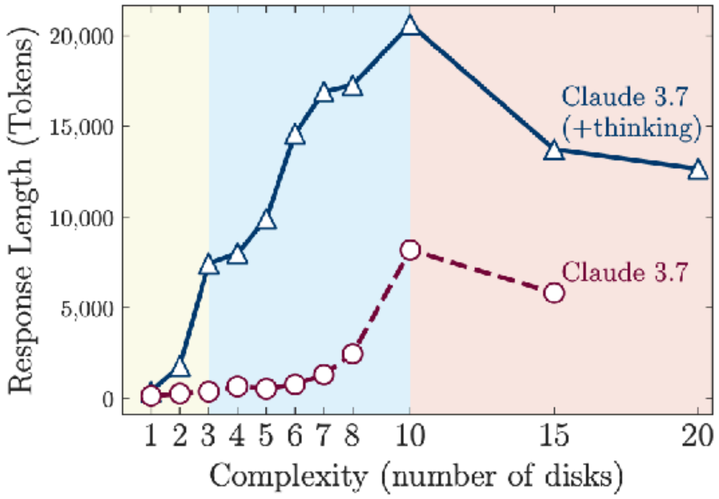
```
<answer> the final answer is moves=...
</answer>
```

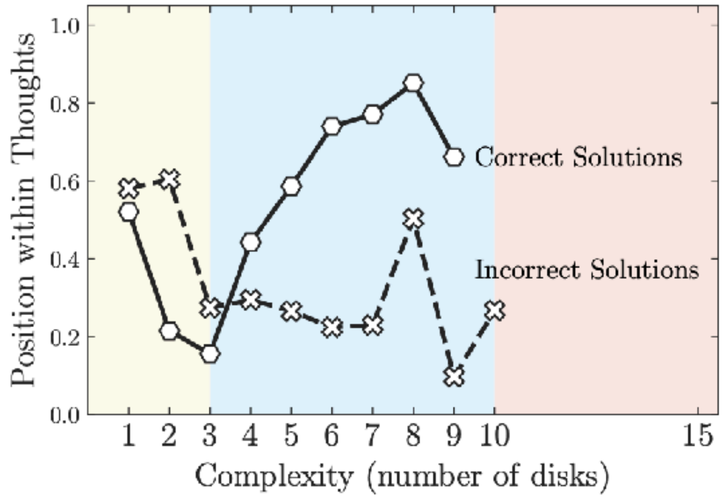
extract moves from thoughts
(for analysis)

extract final answer
(for measuring accuracy)









The Illusion of Thinking (2025)

Reconsider the Möbius Strip

Understanding the Möbius strip requires building an internal spatial model:

- A surface in 3D space
- Its boundary
- A twist
- How these relate

Humans do this: We mentally rotate it and imagine walking on it.

AF has: Tokens, vector embeddings, and next token prediction. These captures statistical relationships well, but cannot reason at the higher-level.

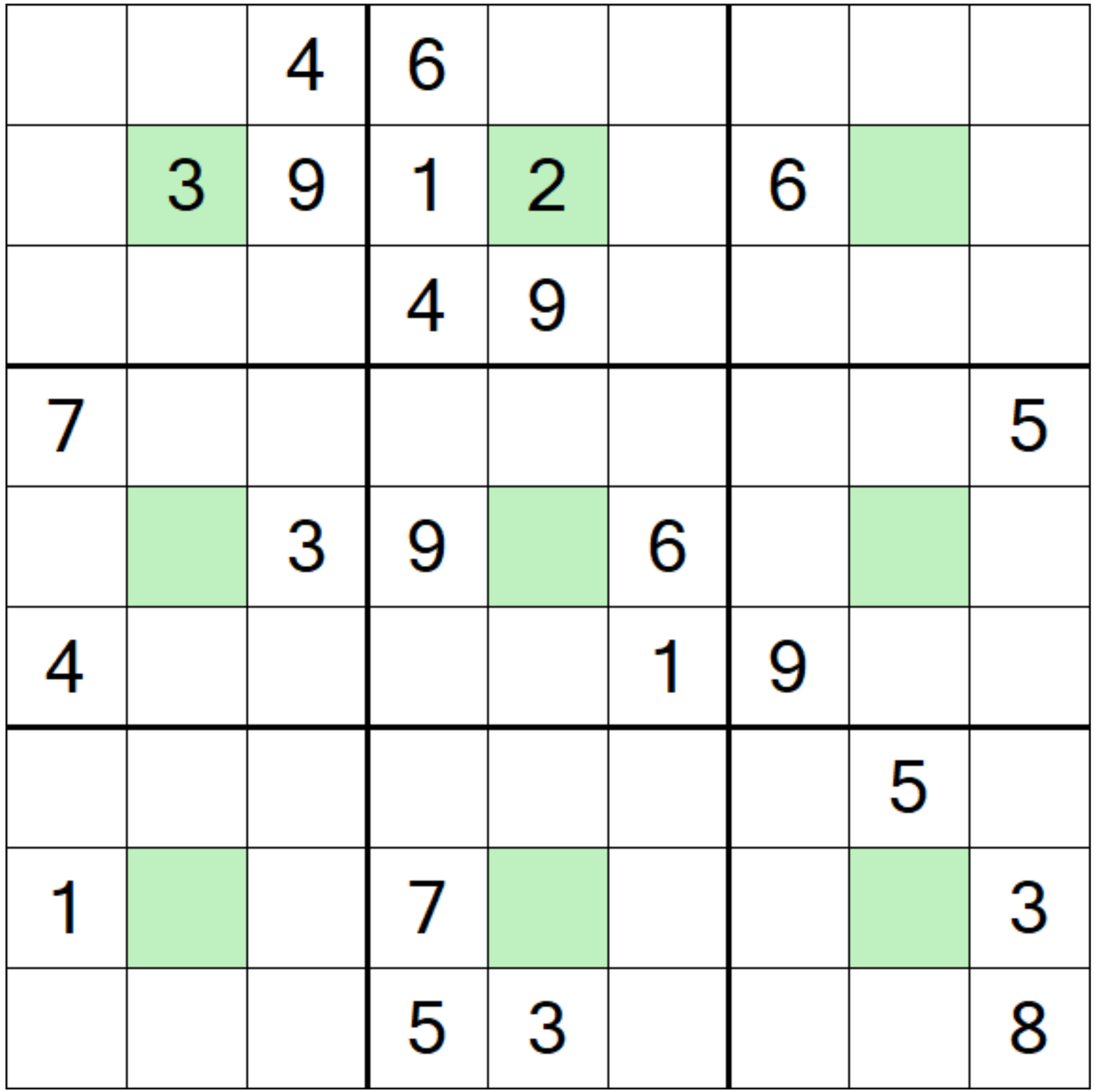
Reconsider the Möbius Strip

Intelligence requires global understanding and deductive reasoning.

- Premise: Möbius strip has one side
- Premise: Cutting along a line divides separates the sides.
- However, there is only one side so it cannot be separated.
- Conclusion: Therefore cut results in a longer strip, not two pieces.

Consider Sudoku

The tool sees millions of Sudoku puzzles and solutions during training. Then, when it is given a partial grid, it predicts which numbers likely go in empty cells based on what similar puzzles had in those positions.



So what is AF actually good for?

1. Perception and Pattern Recognition at Scale. The task is statistical correlation!
2. Creative Generation and Ideation: When there's no single 'correct' answer.
3. Distilling and Reformatting information. Converting clear instructions into programming syntax.

These are real, important problems. Fluency is revolutionary for them.

What does it mean for intelligence?

What is *intelligence*?

Possibily, it is:

- The interplay between pattern and principle
- The ability to question our own reasoning
- The ability to recognise the blind spots
- The curiosity to explore!

unLecture

The *Fault* in Our Intelligence

Aalok Thakkar