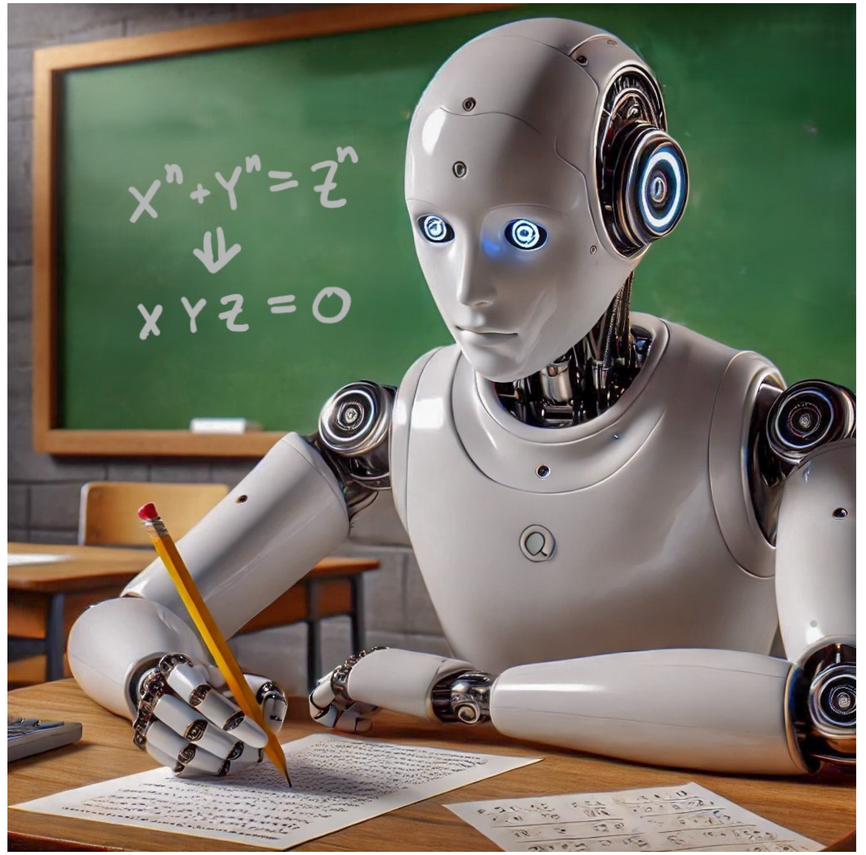
Math 480: Introduction to Mathematical Formalization

Lecture 3: Why formalize?



Jarod Alper
University of Washington

Computers are useless. They can only give you answers.

—Pablo Picasso (1968)

What is Math AI?

1. Mathematics behind AI
2. Formalization
3. Autoformalization
4. Machine learning to assist research
5. Mathematics in the age of AI

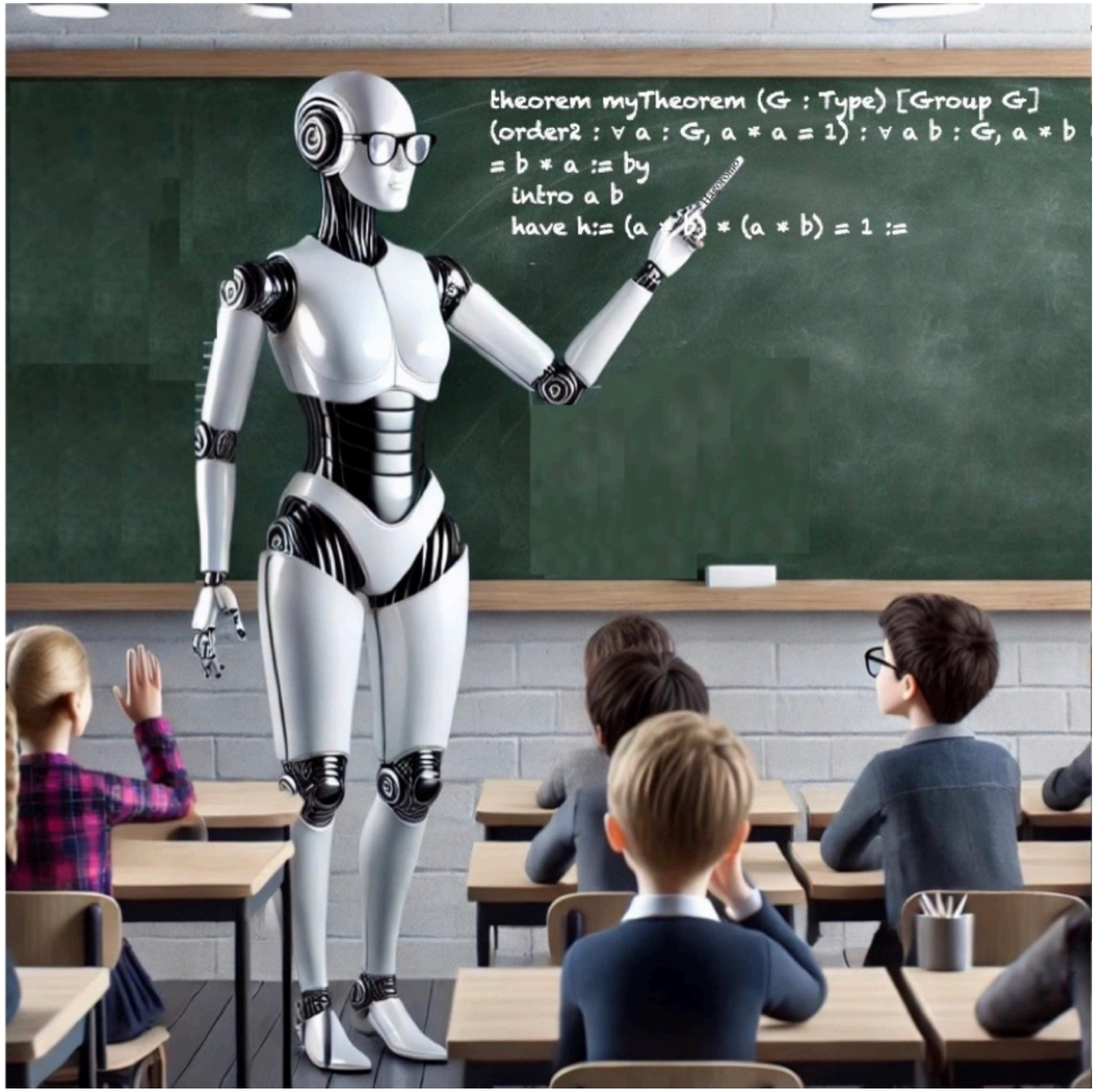


image by DALLE-3

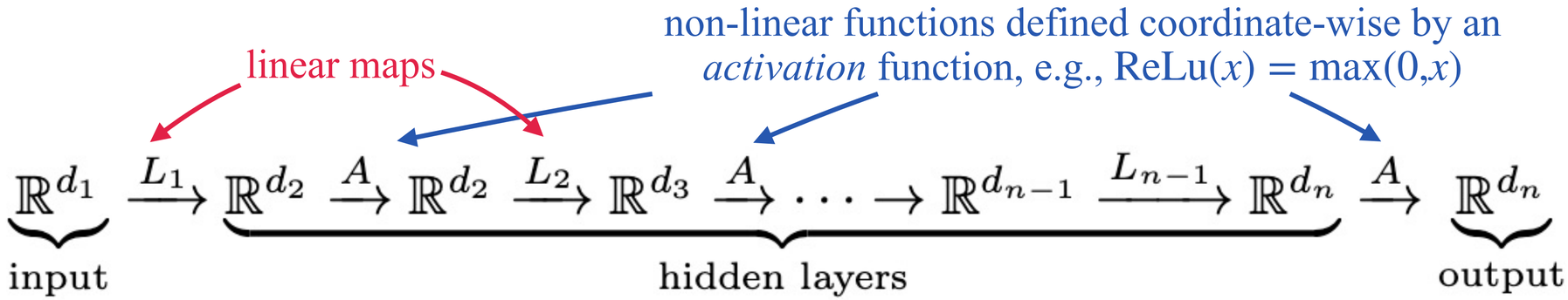
① Mathematics behind AI

Investing in applied machine learning without understanding the mathematical foundations is like investing in health care without understanding biology.
—Rebecca Willett (2023)

The trillion dollar question

Why are neural networks as effective as they are?

A *neural network* is a specific type of function:



Many interesting mathematical questions:

- What types of functions can neural networks effectively learn?
- What is the tradeoff in number of layers and their dimension?
- How to avoid the *curse of dimensionality?*

② Formalization

↳ *Mathematical formalization* is the translation of mathematical proofs into a formal language that can be checked by a computer.

↳ This is achieved using a *proof assistant*, an interactive program that facilitates the translation of a proof into a sequence of logical deductions from the axioms.

↳ There are many proof assistants:

Lean , Agda, Coq, Mizar, HOL, Isabelle, ...

image by DALLE-3

Why formalize?

> A word of warning - and apology. There are several thousand formulas in this paper which allow one or more `sign-like ambiguities'... I have made a superhuman effort to achieve consistency and even to make correct statements: but I still cannot guarantee the result. — David Mumford (1966)

Reasons

- ↳ Complete confidence in correctness
- ↳ Improving understanding
- ↳ Training computers
- ↳ Mathematical exposition
- ↳ Software verification

③ Autoformalization

↳ *Autoformalization* is the generation of a proof by a computer.

↳ The formal:informal proof ratio

- Current ratio is between $100 - \infty$, depending on field.
- Goal is to get the ratio less than 1.

↳ Examples:

- Numina's solution to the 2024 AIMO progress prize
- Deepmind's Silver-level 2024 performance

↳ Deepmind used a reinforced learning algorithm called *AlphaProof* in similar spirit to *AlphaZero* that learned Go, Chess, Shogi entirely through self-play.

image by DALLE-3

Alphaproof and Alphazero

- *AlphaZero* uses probabilistic tree search and a deep neural network, which is trained to not only to learn a good valuation function of a board state but also a probability distribution called a *policy* for effective next moves.

- By replacing moves with logical steps and board states with proof states, these same techniques apply to proof generation.

- Unfortunately, most ML research is narrowly focused on the competition-level math, artificial benchmarks, and proof-of-concept results. Their tools are not available for us to experiment with.

image by DALLE-3

④ Machine learning in mathematical research

- In the 1770s, Felkel and Vega computed factorization tables up to 408,000. This inspired Legendre and Gauss to conjecture the *prime number theorem*.
- In the 1960s, Birch and Swinnerton-Dyer used a primitive computer to count solutions to elliptic curves over finite fields, leading to their famous conjecture.

Machine learning algorithms of today are a massively superpowered version of this.

What have they been good at?

- Using large data sets to find relationships, e.g., *elliptic curve murmurations* (He, et al) and knot invariants (Davies, et al)
- Generate counterexamples for conjectures, e.g., graph theory (Wagner)
- Produce efficient formulas/representations, e.g., tensor decompositions for matrix multiplication (AlphaTensor)

⑤ Meaning of mathematics in the age of AI

- What do we do when computers outperform humans in the Olympiad?
- What about when they become better at proving theorems?
- What if they discover a five page elementary proof of Fermat's Last Theorem?
- What if they become better at generating conjectures or synthesizing mathematics?
- What if they become better at teaching mathematics? What if they become better at writing mathematics? What if they become better at understanding mathematics? What if they become better at being mathematicians? What if they become better at being human? What if they become better at being? What if they become? What if they? What if? What?

image by DALLE-3

⑤ Meaning of mathematics in the age of AI

- Mathematics is ultimately a human endeavor.
- AI will change the way we do research, the way we write, and even the way we think, but it will be up to **us** to determine what mathematical statements we value and to develop a human understanding.
- It is also imperative to recognize and address the risk (short-term, medium-term, and existential) and ethical issues of AI.

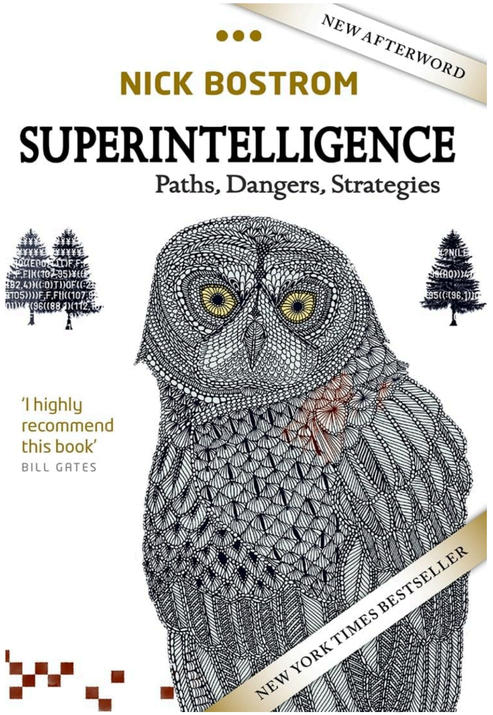
image by DALLE-3

Existence risk

Populare literature addressing AI risks

- Nick Bostrom, Superintellgience: Paths, Dangers, and Strategies
- Max Tegmark, Life 3.0: Being human in the age of Artificial Intelligence
- Toby Ord, The Precipice: Existential risk and the future and humanity

A paper clip apocalypse

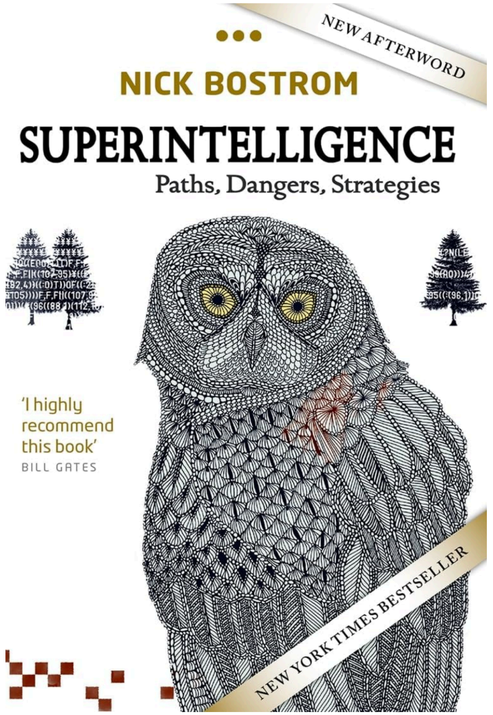


Infrastructure profusion can result from final goals that would have been perfectly innocuous if they had been pursued as limited objectives. Consider the following two examples:

- *Riemann hypothesis catastrophe*. An AI, given the final goal of evaluating the Riemann hypothesis, pursues this goal by transforming the Solar System into "computronium" (physical resources arranged in a way that is optimized for computation)—including the atoms in the bodies of whomever once cared about the answer.[8]
- *Paperclip AI*. An AI, designed to manage production in a factory, is given the final goal of maximizing the manufacture of paperclips, and proceeds by converting first the Earth and then increasingly large chunks of the observable universe into paperclips.

In the first example, the proof or disproof of the Riemann hypothesis that the AI produces is the intended outcome and is in itself harmless; the harm comes from the hardware and infrastructure created to achieve this result. In the second example, some of the paperclips produced would be part of the intended outcome; the harm would come either from the factories created to produce the paperclips (infrastructure profusion) or from the excess of paperclips (perverse instantiation).

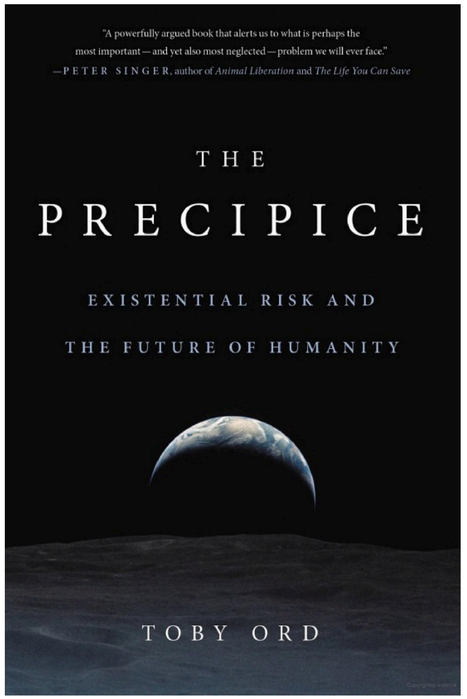
A paper clip apocalypse



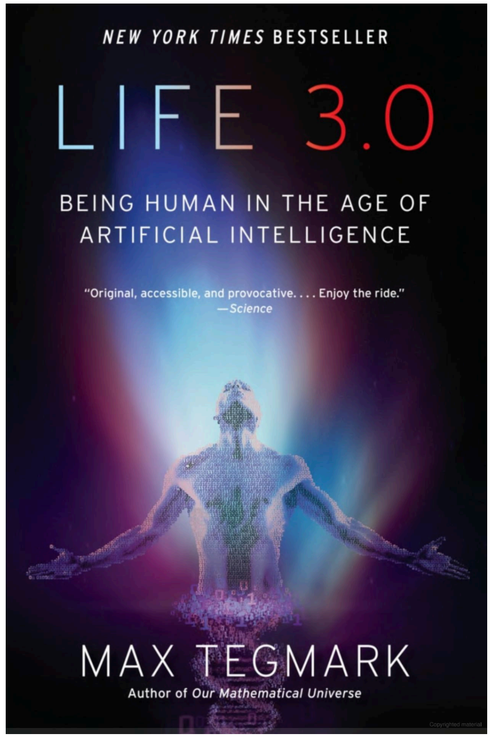
One might think that the risk of a malignant infrastructure profusion failure arises only if the AI has been given some clearly open-ended final goal, such as to manufacture as many paperclips as possible. It is easy to see how this gives the superintelligent AI an insatiable appetite for matter and energy, since additional resources can always be turned into more paperclips. But suppose that the goal is instead to make at least one million paperclips (meeting suitable design specifications) rather than to make as many as possible. One would like to think that an AI with such a goal would build one factory, use it to make a million paperclips, and then halt. Yet this may not be what would happen.

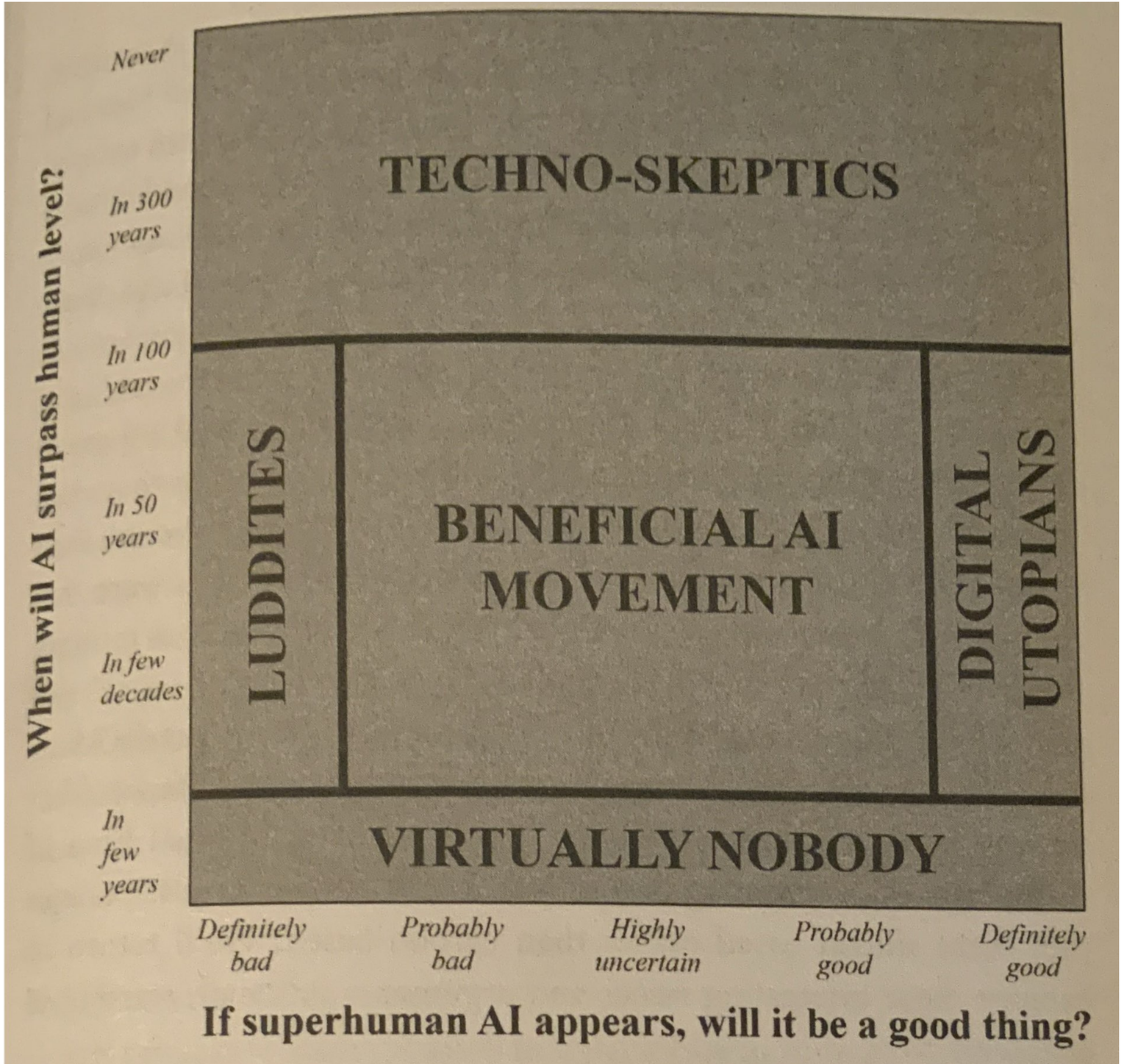
Unless the AI's motivation system is of a special kind, or there are additional elements in its final goal that penalize strategies that have excessively wide-ranging impacts on the world, there is no reason for the AI to cease activity upon achieving its goal. On the contrary: if the AI is a sensible Bayesian agent, *it would never assign exactly zero probability to the hypothesis that it has not yet achieved its goal*—this, after all, being an empirical hypothesis against which the AI can have only uncertain perceptual evidence. The AI should therefore continue to make paperclips in order to reduce the (perhaps astronomically small) probability that it has somehow still failed to make at least a million of them, all appearances notwithstanding. There is nothing to be lost by continuing paperclip production and there is always at least some microscopic probability increment of achieving its final goal to be gained.

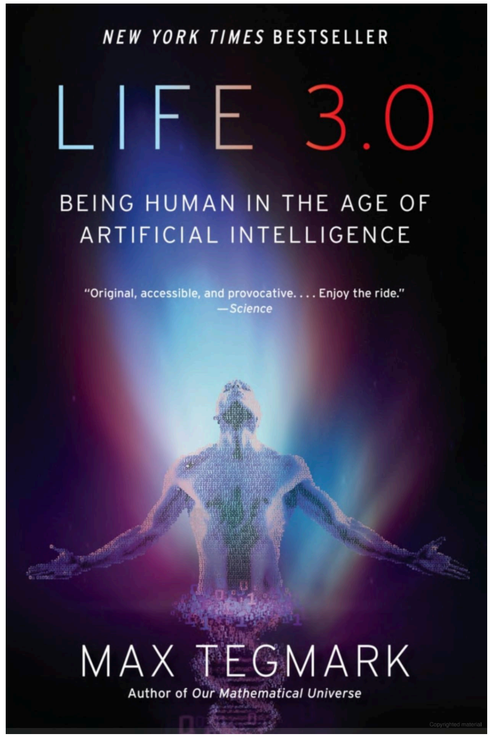
Can the risk be quantified?

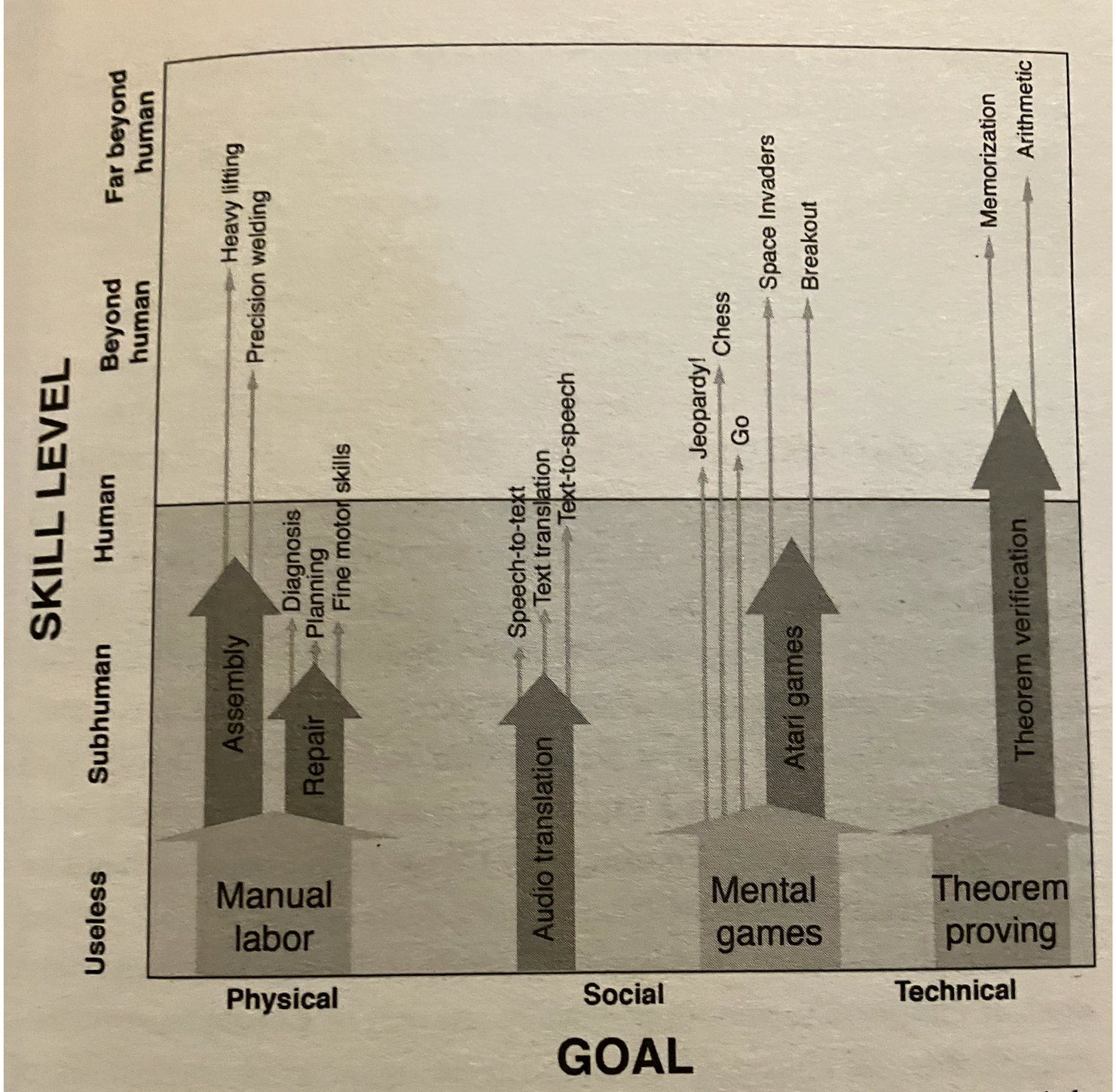


Existential catastrophe via	*Chance within next 100 years*
Asteroid or comet impact	~ 1 in 1,000,000
Supervolcanic eruption	~ 1 in 10,000
Stellar explosion	~ 1 in 1,000,000,000
Total natural risk	**~ 1 in 10,000**
Nuclear war	~ 1 in 1,000
Climate change	~ 1 in 1,000
Other environmental damage	~ 1 in 1,000
"Naturally" arising pandemics	~ 1 in 10,000
Engineered pandemics	~ 1 in 30
Unaligned artificial intelligence	~ 1 in 10
Unforeseen anthropogenic risks	~ 1 in 30
Other anthropogenic risks	~ 1 in 50
Total anthropogenic risk	**~ 1 in 6**
Total existential risk	**~ 1 in 6**


NEW YORK TIMES BESTSELLER
LIFE 3.0
BEING HUMAN IN THE AGE OF ARTIFICIAL INTELLIGENCE
"Original, accessible, and provocative. . . . Enjoy the ride."
—Science
MAX TEGMARK
Author of Our Mathematical Universe


When will AI surpass human level?
Never
In 300 years
In 100 years
In 50 years
In few decades
In few years
TECHNO-SKEPTICS
LUDDITES
BENEFICIAL AI MOVEMENT
DIGITAL UTOPIANS
VIRTUALLY NOBODY
Definitely bad
Probably bad
Highly uncertain
Probably good
Definitely good
If superhuman AI appears, will it be a good thing?


NEW YORK TIMES BESTSELLER
LIFE 3.0
BEING HUMAN IN THE AGE OF ARTIFICIAL INTELLIGENCE
"Original, accessible, and provocative. . . . Enjoy the ride."
—Science
MAX TEGMARK
Author of Our Mathematical Universe


SKILL LEVEL
Useless
Subhuman
Human
Beyond human
Far beyond human
Heavy lifting
Precision welding
Assembly
Diagnosis
Planning
Fine motor skills
Repair
Speech-to-text
Text translation
Text-to-speech
Audio translation
Jeopardy!
Chess
Go
Space Invaders
Breakout
Atari games
Memorization
Arithmetic
Theorem verification
Manual labor
Mental games
Theorem proving
Physical
Social
Technical
GOAL

Evolution of Proof

Rigor has ceased to be thought of
as a cumbersome style of formal dress that
one has to wear on state occasions and discards
with a sigh of relief as soon as one comes home. We do
not ask any more whether a theorem has been
rigorously proved but whether it has been proved.

—André Weil (1956)

Example: evolution of Hilbert's basis theorem

Hilbert (1890)

Ist irgend eine nicht abbrechende Reihe von Formen der n Vernderlichen $x_1, x_2, \ldots, x_n$ gegeben, etwa F_1, F_2, F_3, ..., so giebt es stets eine Zahl m von der Art, dass eine jede Form jener Reihe in die Gestalt

$$F = A_1F_1 + A_2F_2 + \cdots + A_mF_m$$

bringen lsst, wo $A_1, A_2, \ldots, A_m$ geeignete Formen der nmlichen n Vernderlichen sind.

Example: evolution of Hilbert's basis theorem

Hilbert (1890)

If any non-terminating sequence of forms of the n variables $x_1, x_2, \ldots, x_n$ is given, for instance $F_1, F_2, F_3, \ldots$, then there always exists a number m of such a kind that every form of that sequence can be written as

$$F = A_1 F_1 + A_2 F_2 + \cdots + A_m F_m,$$

where $A_1, A_2, \ldots, A_m$ are suitable forms of the same n variables.

Example: evolution of Hilbert's basis theorem

Hilbert (1890)

If any non-terminating sequence of forms of the n variables $x_1, x_2, \ldots, x_n$ is given, for instance $F_1, F_2, F_3, \ldots$, then there always exists a number m of such a kind that every form of that sequence can be written as

$$F = A_1 F_1 + A_2 F_2 + \cdots + A_m F_m,$$

where $A_1, A_2, \ldots, A_m$ are suitable forms of the same n variables.

Bourbaki (1961)

Pour tout anneau commutatif nœthérien C, l'anneau de polynômes $C[x]$ is nœthérien.

Example: evolution of Hilbert's basis theorem

Hilbert (1890)

If any non-terminating sequence of forms of the n variables $x_1, x_2, \ldots, x_n$ is given, for instance $F_1, F_2, F_3, \ldots$, then there always exists a number m of such a kind that every form of that sequence can be written as

$$F = A_1 F_1 + A_2 F_2 + \cdots + A_m F_m,$$

where $A_1, A_2, \ldots, A_m$ are suitable forms of the same n variables.

Bourbaki (1961)

For every commutative noetherian ring C, the ring of polynomials $C[x]$ is noetherian.

Example: evolution of Hilbert's basis theorem

Hilbert (1890)

If any non-terminating sequence of forms of the n variables $x_1, x_2, \ldots, x_n$ is given, for instance $F_1, F_2, F_3, \ldots$, then there always exists a number m of such a kind that every form of that sequence can be written as

$$F = A_1 F_1 + A_2 F_2 + \cdots + A_m F_m,$$

where $A_1, A_2, \ldots, A_m$ are suitable forms of the same n variables.

Bourbaki (1961)

For every commutative noetherian ring C, the ring of polynomials $C[x]$ is noetherian.

mathlib (2019)

```
protected theorem Polynomial.isNoetherianRing [inst :
IsNoetherianRing R] : IsNoetherianRing R[X].
```

Example: evolution of Hilbert's basis theorem

Hilbert (1890)

If any non-terminating sequence of forms of the n variables $x_1, x_2, \ldots, x_n$ is given, for instance $F_1, F_2, F_3, \ldots$, then there always exists a number m of such a kind that every form of that sequence can be written as

$$F = A_1 F_1 + A_2 F_2 + \cdots + A_m F_m,$$

where $A_1, A_2, \ldots, A_m$ are suitable forms of the same n variables.

Bourbaki (1961)

For every commutative noetherian ring C, the ring of polynomials $C[x]$ is noetherian.

mathlib (2019)

```
protected theorem Polynomial.isNoetherianRing [inst :
IsNoetherianRing R] : IsNoetherianRing R[X].
```