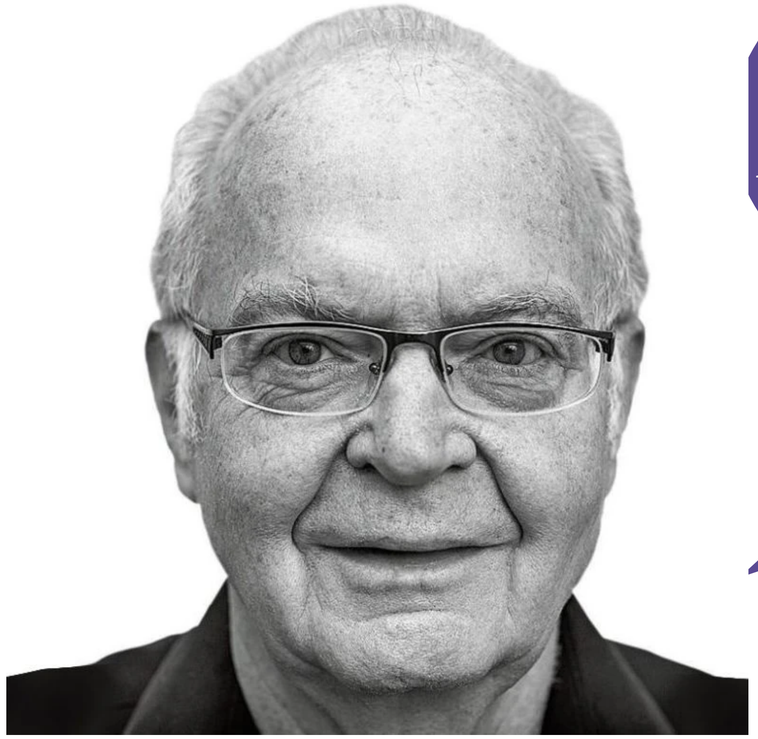
Mathematics and AI (Fall 2025): Jarod Alper
Lecture 2



Science is what we understand
well enough to explain to a computer,
Art is all the rest.

—Donald Knuth (1996)

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)

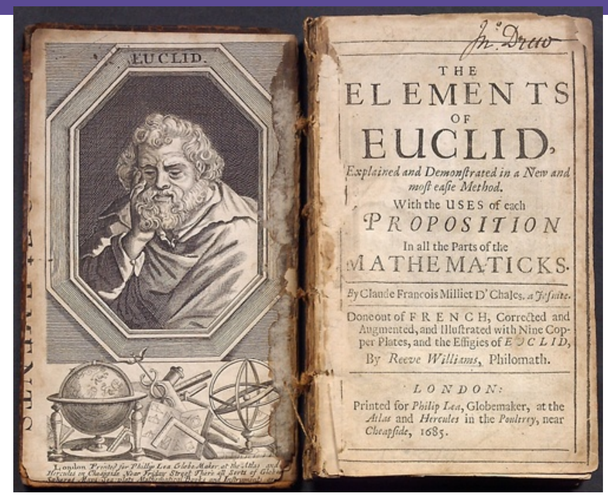
For every commutative noetherian ring C, the ring of polynomials $C[x]$ is noetherian.

mathlib (2019)

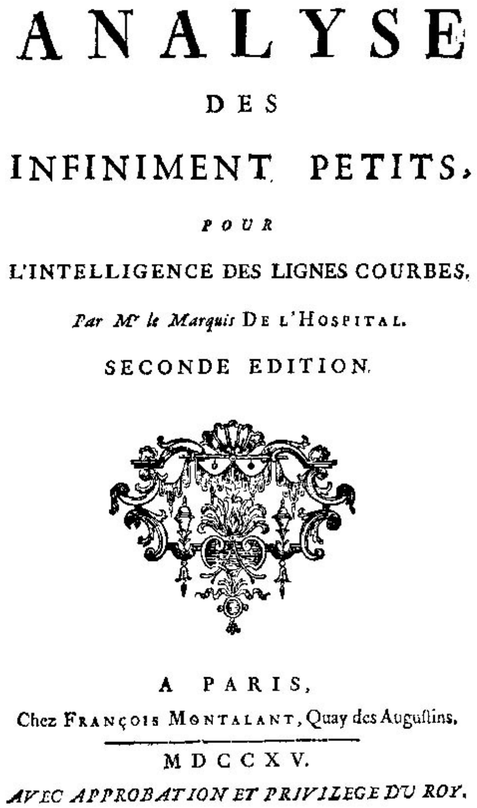
```
protected theorem Polynomial.isNoetherianRing [inst :
IsNoetherianRing R] : IsNoetherianRing R[X].
```

Evolution of Teaching Math

- Prior to 300 BC: informal / oral tradition
- 300 BC: Euclid's *Elements*



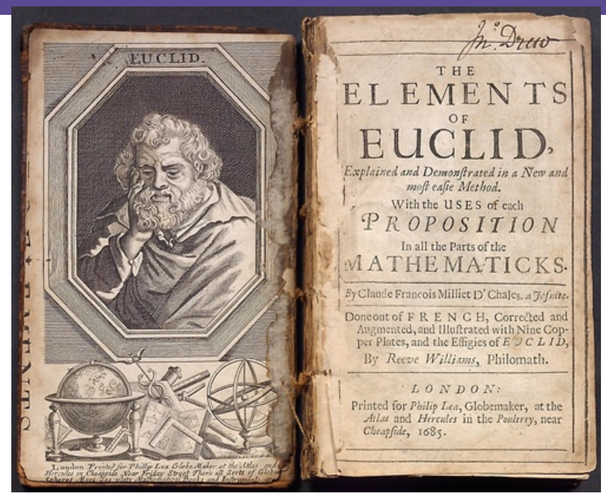
- 1696: l'Hôpital's *Analyse des Infiniment Petits pour l'Intelligence des Lignes Courbes*


ANALYSE
DES
INFINIMENT PETITS,
POUR
L'INTELLIGENCE DES LIGNES COURBES.
Par Mr le Marquis DE L'HOSPITAL.
SECONDE EDITION.

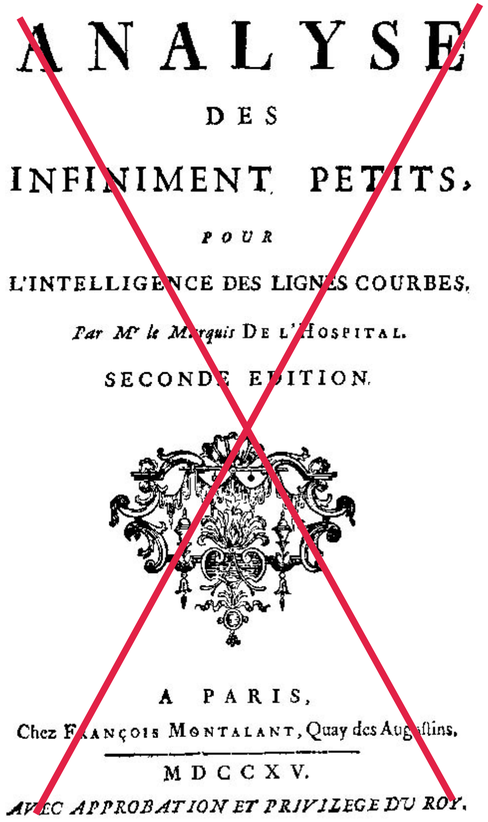
A PARIS,
Chez FRANÇOIS MONTALANT, Quay des Augustins.
MDCCXV.
AVEC APPROBATION ET PRIVILEGE DU ROY.

Evolution of Teaching Math

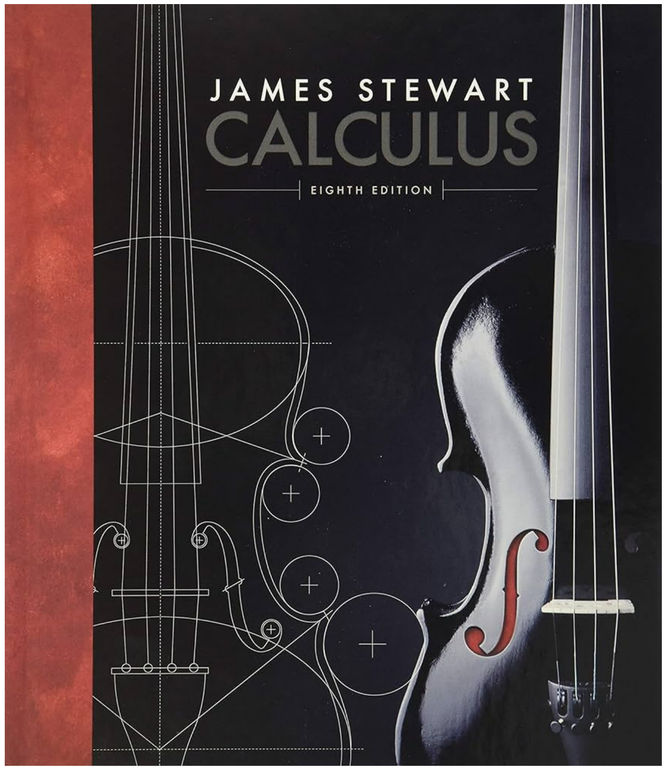
- Prior to 300 BC: informal / oral tradition
- 300 BC: Euclid's *Elements*



- 1696: l'Hôpital's *Analyse des Infiniment Petits pour l'Intelligence des Lignes Courbes*



- Today: Essentially same text



The product of mathematics is clarity and understanding. Not theorems, by themselves.... In short, mathematics only exists in a living community of mathematicians that spreads understanding and breathes life into ideas both old and new.

—Bill Thurston (2010)

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

Gower's 1999: Rough structure and classification

2 Will Mathematics Exist in 2099?

For at least a century, mathematicians have thought about the possibility of automating mathematics. Hilbert famously asked in his tenth problem whether there was an algorithm for solving Diophantine equations, and later extended this to the question of whether there was an algorithm which would find a proof of any mathematical statement that had one. In 1936, Turing, equally famously, formalized the notion of algorithm and soon afterwards demonstrated the insolubility of the halting problem, thus showing that no such algorithm existed. While this result, and later demonstrations that several natural and well known problems in mathematics were also impossible to solve systematically, may have initially seemed somewhat negative, they also had a positive side. The idea that all our creativity and insight might be reduced to something mechanical was, after all, not very appealing. Turing's result therefore came as a relief, since it left mathematicians with something to do.

Gower's 1999: Will mathematics exist in 2099?

techniques may be helpful. Rather than giving several examples of the use of standard methods to solve problems, let me return to the question of automating mathematics and present an imagined dialogue between a mathematician and a computer in two or three decades' time. The idea of the dialogue is that the computer is very helpful to the mathematician, while not doing anything particularly clever. This represents an unthreatening intermediate stage between what we have now, computers that act as slaves doing unbelievably boring calculations for us, and full automation of mathematics. I have written the dialogue in English, but this is supposed to be a translation of a more formal language which has not yet been invented. (I shall discuss this point a little more later.)

Mathematician. Is the following true? Let $\delta > 0$. Then for N sufficiently large, every set $A \subset \{1, 2, \dots, N\}$ of size at least δN contains a subset of the form $\{a, a+d, a+2d\}$?

Computer. Yes. If A is non-empty, choose $a \in A$ and set $d = 0$.

M. All right all right, but what if d is not allowed to be zero?

C. Have you tried induction on N, with some $\delta = \delta(N)$ tending to zero?

M. That idea is no help at all. Give me some examples please.

C. The obvious greedy algorithm gives the set

$$\{1, 2, 4, 5, 10, 11, 13, 14, 28, 29, 31, 32, 37, 38, 40, 41, \dots\}.$$

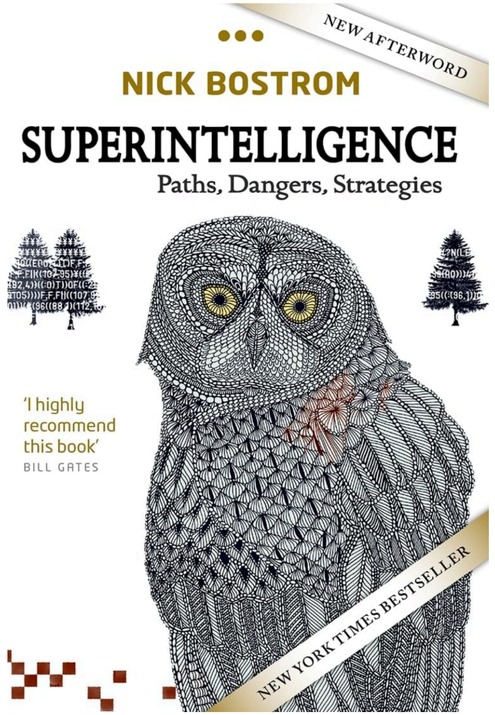
I notice that large parts of the set are translations of other parts. In fact, this set is very like the Cantor set, so this gives a bound of $\delta \geqslant N^{(\log 2/\log 3)-1}$.

Existence risk

Popular literature addressing AI risks

- Nick Bostrom, *Superintellgience: Paths, Dangers, and Strategies*, 2014
- Max Tegmark, *Life 3.0: Being human in the age of Artificial Intelligence*, 2017
- Toby Ord, *The Precipice: Existential risk and the future and humanity*, 2020

A paper clip apocalypse



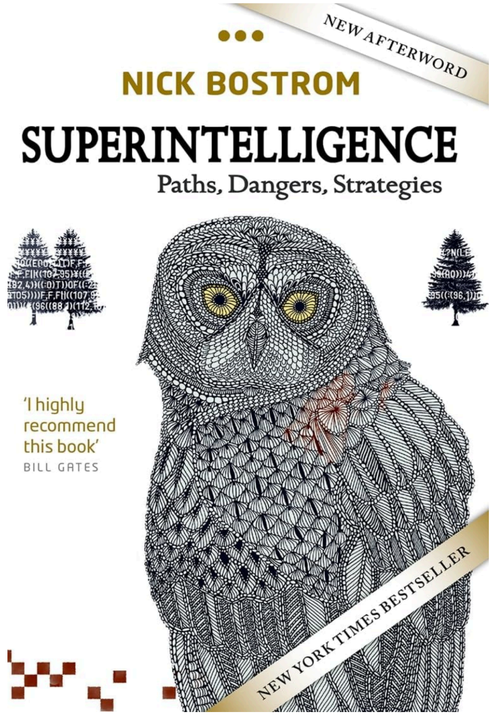
2014

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

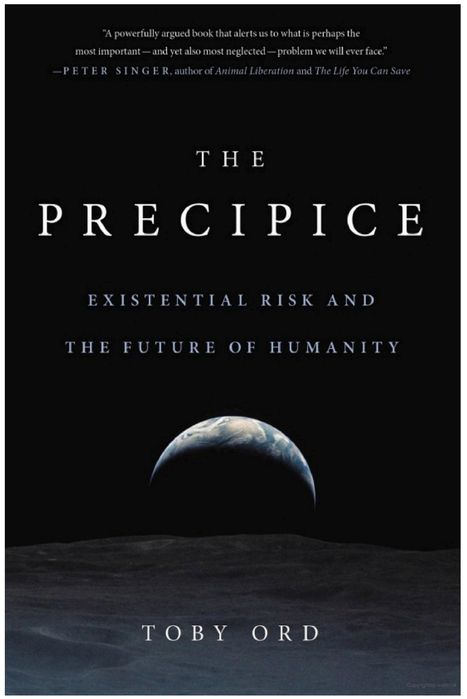


2014

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.

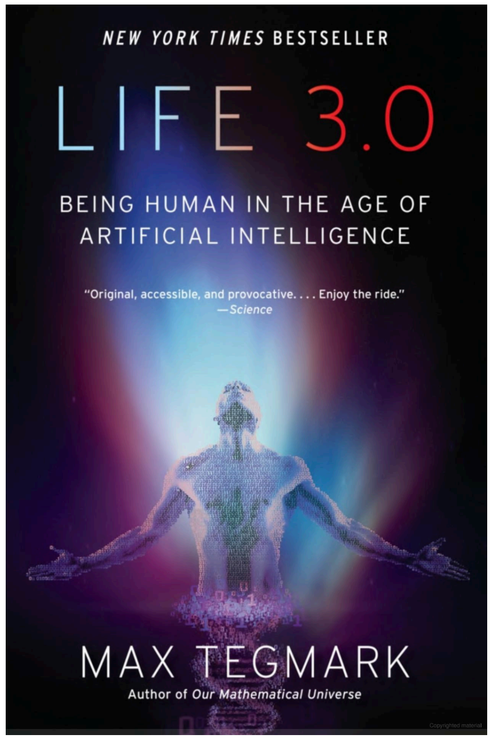
Unless the AI's motivation system is of a special kind, or there are additional elements in its final goal that penalize strategies that have excessively wide-ranging impacts on the world, there is no reason for the AI to cease activity upon achieving its goal. On the contrary: if the AI is a sensible 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?

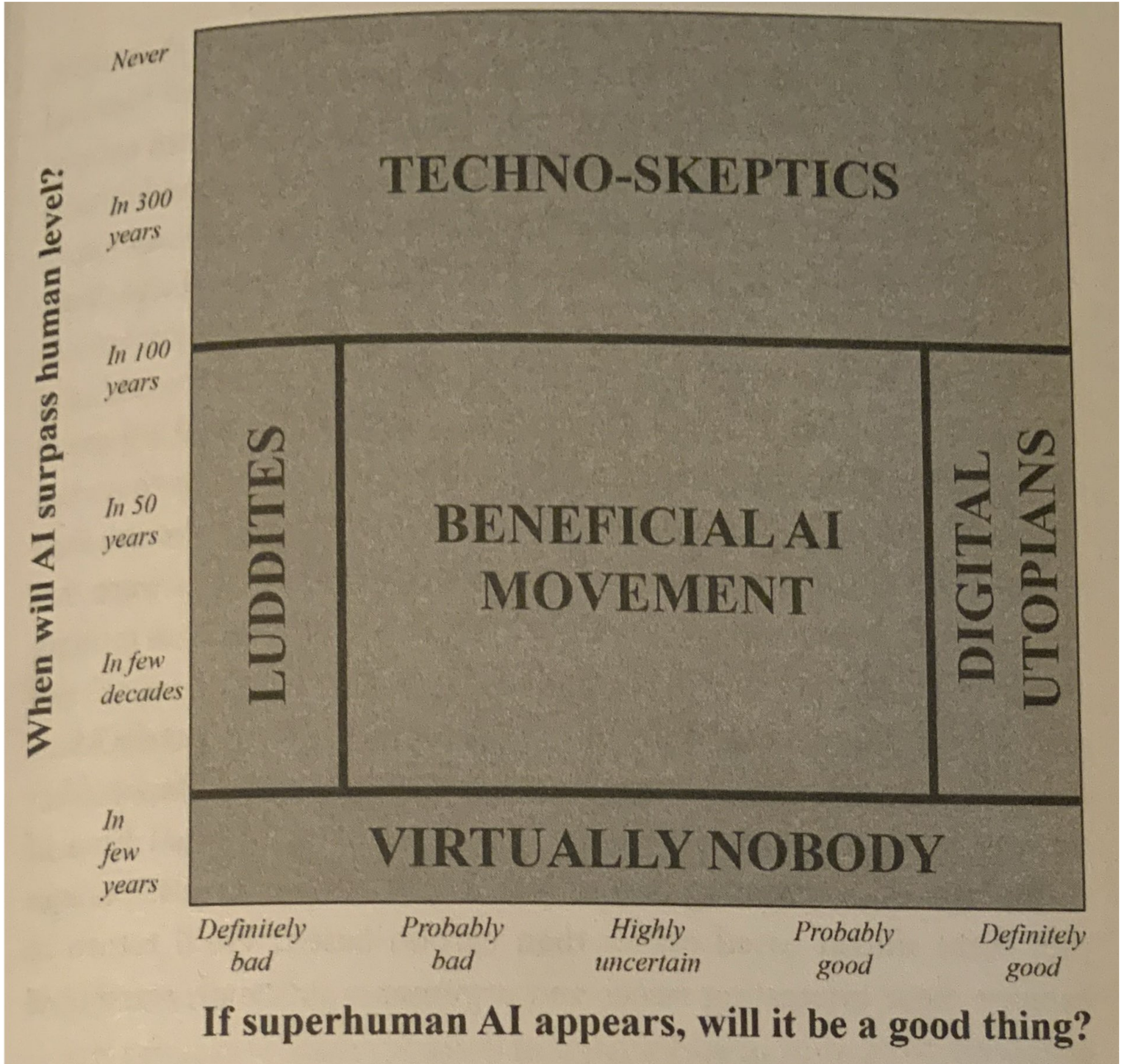


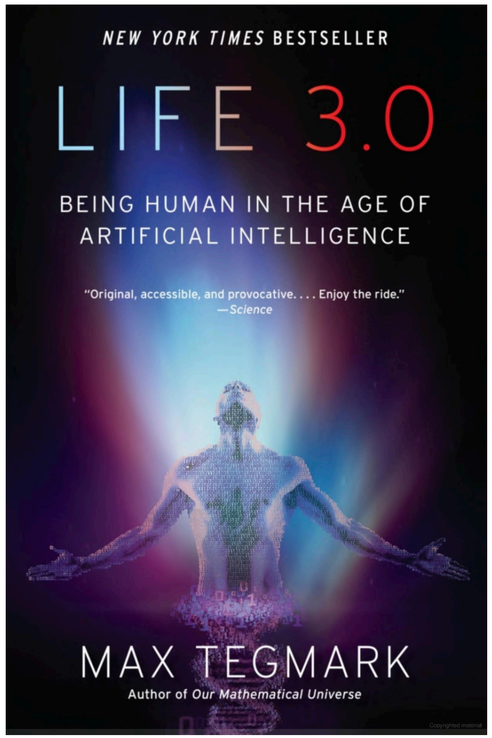
2020

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**

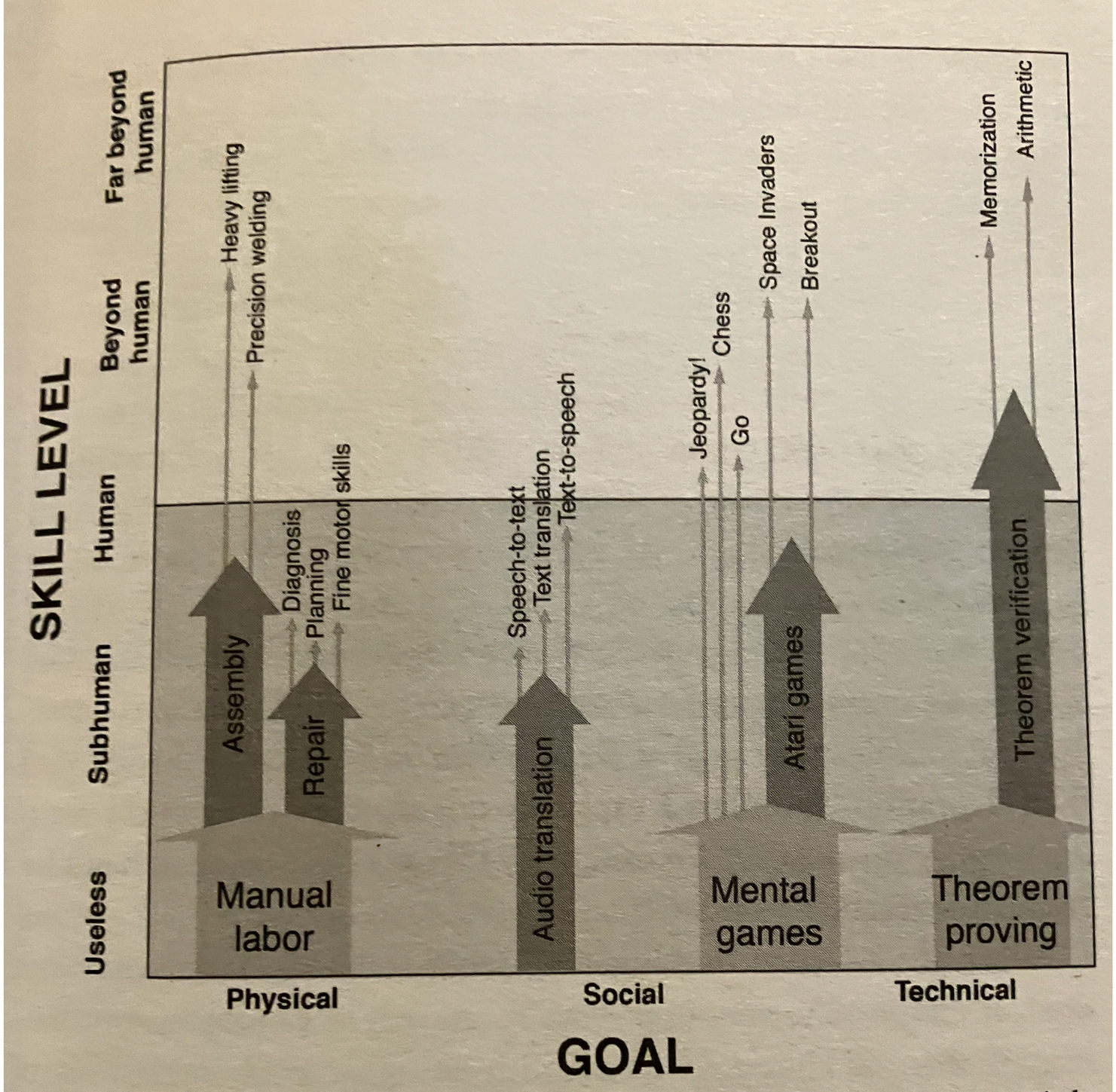

NEW YORK TIMES BESTSELLER
LIFE 3.0
BEING HUMAN IN THE AGE OF ARTIFICIAL INTELLIGENCE
"Original, accessible, and provocative. . . . Enjoy the ride."
—Science
MAX TEGMARK
Author of Our Mathematical Universe

2017


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

2017


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

Gower’s 1999: Will mathematics exist in 2099?

Gower's 2025: Will mathematics exist in 2035?

Scenarios in the next 10 years

—The external disaster (nuclear…): low but non-zero
—The general singularity: hard to say
—The math singularity: hard to say
—The steamroller: low to medium
—The black box: low
—The gradual takeover: (conditionally) high
—The collapse of incentives: medium to high
—The plateau: medium
—The glorified calculator: low

Evolution of AI

- 19th century & before: Many mathematicians and academics philosophized about the existence of computing machines
- 1940s: invention of computing devices and mathematical models of neurons
- 1950: Alan Turing's paper *Computing Machinery and Intelligence*
- 1956: Dartmouth conference organized by John McCarthy and birth of the term 'artificial intelligence'
- 1958: First neural network—Frank Rosenblatt, *The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain*
- 1960s and 70s: AI winter
- 1980s: backpropogation discovered independently by Werbos, Rummelhart, and others. Hopfield & Hinton received 2024 Nobel Prize in Physics for related work. Yann LeCun invents convolution and recurrent neural networks.
- …. many breakthroughs since.

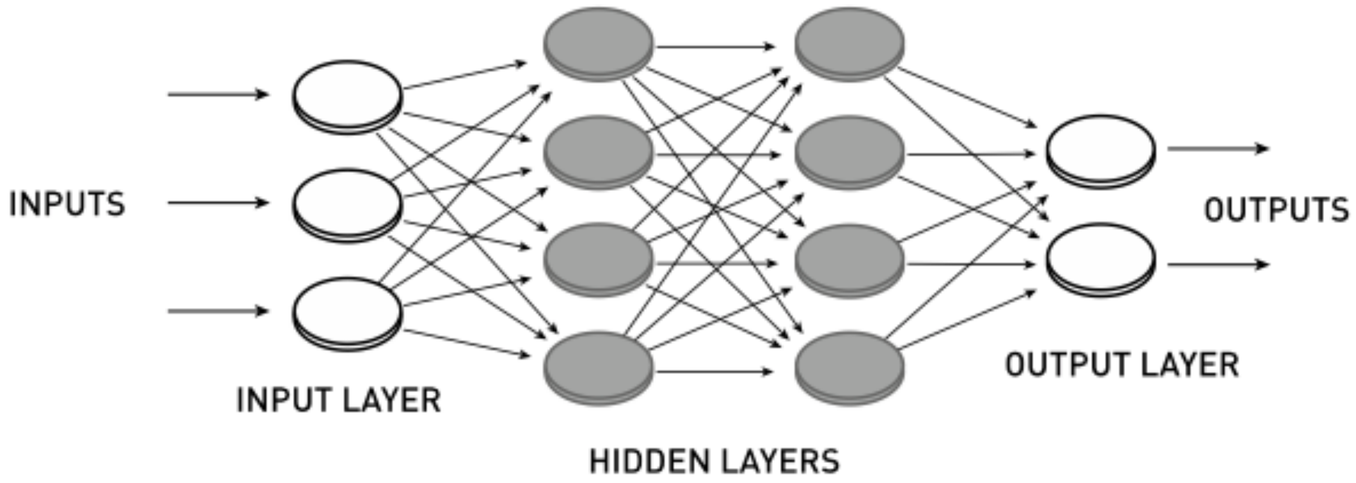
Neural networks

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

affine linear maps

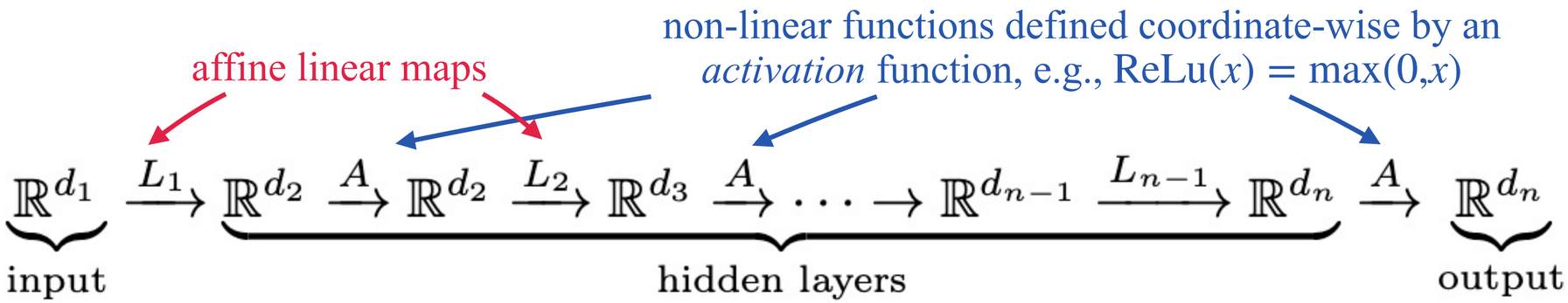
non-linear functions defined coordinate-wise by an *activation* function, e.g., ReLu(x) = max(0,x)

$$\underbrace{\mathbb{R}^{d_1}}_{\text{input}} \xrightarrow{L_1} \underbrace{\mathbb{R}^{d_2} \xrightarrow{A} \mathbb{R}^{d_2} \xrightarrow{L_2} \mathbb{R}^{d_3} \xrightarrow{A} \cdots \to \mathbb{R}^{d_{n-1}} \xrightarrow{L_{n-1}} \mathbb{R}^{d_n}}_{\text{hidden layers}} \xrightarrow{A} \underbrace{\mathbb{R}^{d_n}}_{\text{output}}$$



Neural networks

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?*

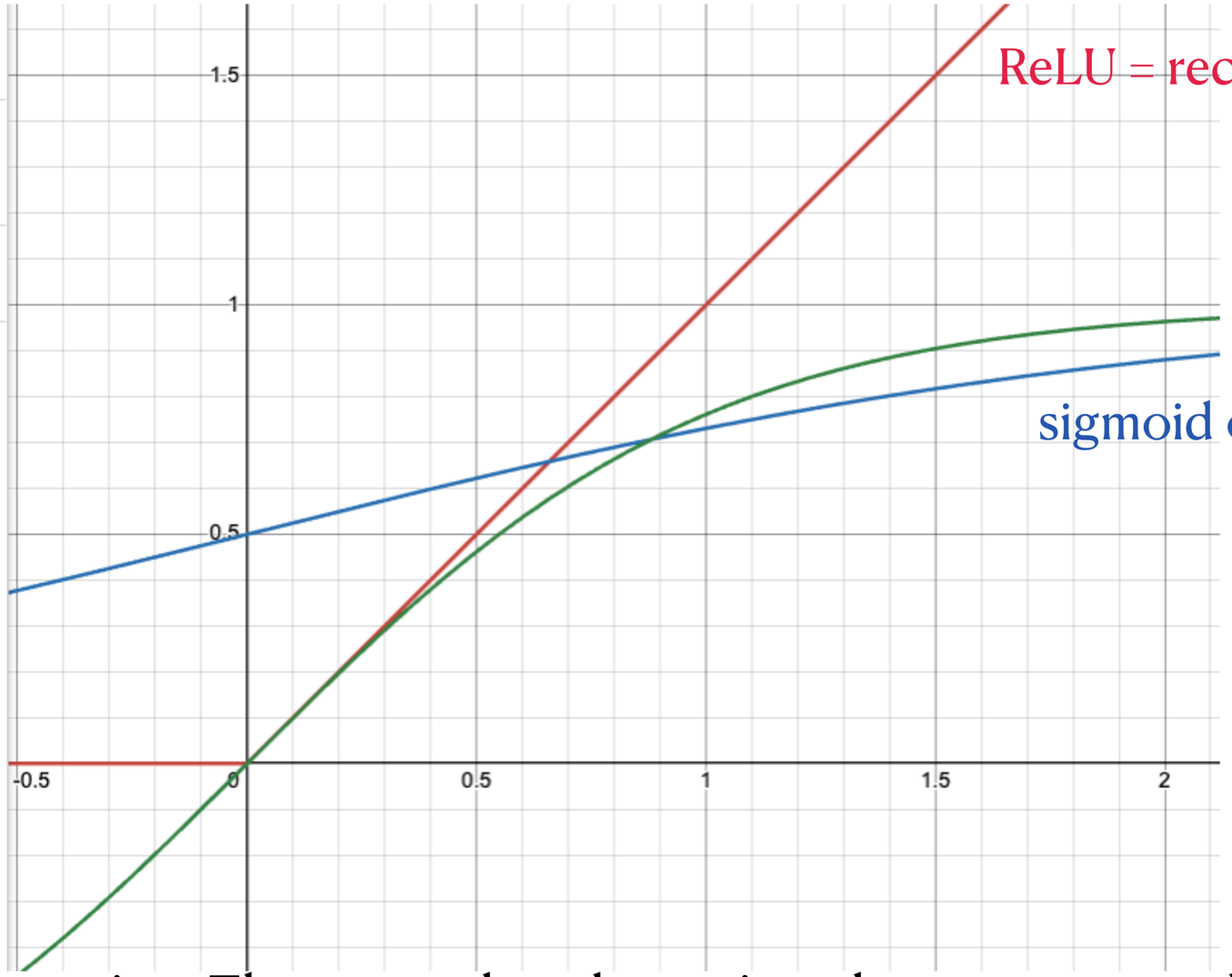
Activation functions

1 $a_1(x) = \max(0,x)$

2 $a_2(x) = \frac{1}{1+e^{-x}}$

3 $a_3(x) = \tanh(x)$

4



ReLU = rectified linear unit

tanh

sigmoid or logistic

These are applied coordinate-wise. There are other alternatives that are not coordinate-wise:

example: $\text{softmax}(x_1, \ldots, x_n)_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$

Stochastic gradient descent

Goal: Find the minimum of a function $f(x_1, \ldots, x_n) \colon \mathbb{R}^n \to \mathbb{R}$

In machine learning, this means finding the model parameters that minimize a cost function f, which measures how wrong the model's predictions are.

A Hiker in the Fog Analogy: Imagine you're a hiker trying to get to the lowest point in a foggy mountain range.

- **Standard Gradient Descent (The "Careful" Hiker):** You stop and check your map and compass at every single step, looking at the slope of the *entire landscape* around you to decide which direction is steepest downhill. This is very accurate but incredibly slow and requires a huge map.
- **Stochastic Gradient Descent (The "Quick" Hiker):** Instead of looking at the whole landscape, you just look at the slope of the ground right under your feet for *one random direction* (a single data point). You take a quick step in that downhill direction.
  - This path will be **noisy and wobbly**. You might even go slightly uphill sometimes!
  - However, you move **much faster**.
  - Over many small steps, the overall trend will guide you to the bottom of the valley.

Backpropogation