

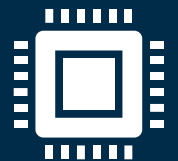
Artificial Intelligence Safety

Introduction

Annachiara Ruospo, *Politecnico di Torino, IT*



Politecnico
di Torino



AI Safety Course

- It will provide a comprehensive introduction to the dependability and safety of AI systems
- It will explore the foundational principles of AI, including deep learning and the hardware architectures that support modern algorithms
- It covers the evolving landscape of AI standardization and industry regulations, such as the EU AI Act
- It discusses state-of-the-art solutions to assess, detect, and mitigate hardware-induced faults in AI systems
- It concludes with a discussion of future trends and challenges in the AI safety field.

The world we imagine, the world we are building

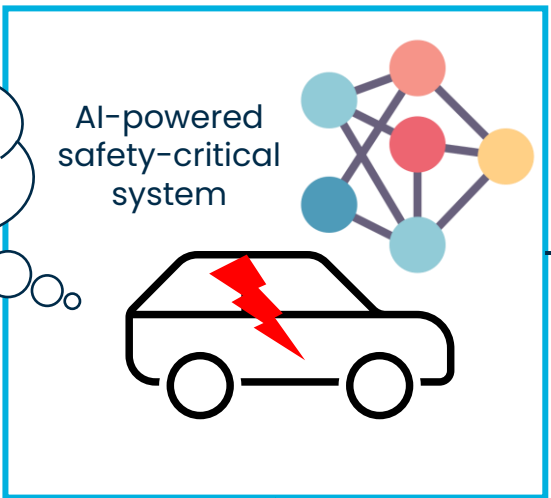
IS THIS SAFE?

Open-world scenario

Real-world applications where the environment is not static, and new categories can emerge after the initial training phase.

Example

Mainly trained based on the **closed-world** assumption



Make predictions

Known Classes

- Low Image resolution
 - Unknown instances
 - System failure
-

The AI model should be automatically updated to include these new classes.

It should incrementally learn new classes over time.

It should automatically detect failures, safety issues, and identify new classes.

Is this safe?

Q1: Is the AI system behaving correctly?

Q2: Is it able to interrupt its activity without causing serious damages?

Q3: Is the AI system able to detect anomalies and unusual situations?

- if a sensor on an autonomous vehicle is malfunctioning, it should alert the driver to take control or pull over.

Q4: Is the AI system acting according to its *intended* functionality (AI alignment)?

- An AI system is considered aligned if it advances its intended objectives.

Q5: Is the AI system able to withstand random-hardware faults and continue to behave correctly (reliability)?

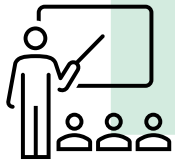
....

AI Safety

Interdisciplinary field that aims at

- building safe and trustworthy AI systems
- establishing confidence in their behavior and robustness
- facilitating their successful adoption in society.

Why a class on AI Safety?



To increase awareness on this area in your research

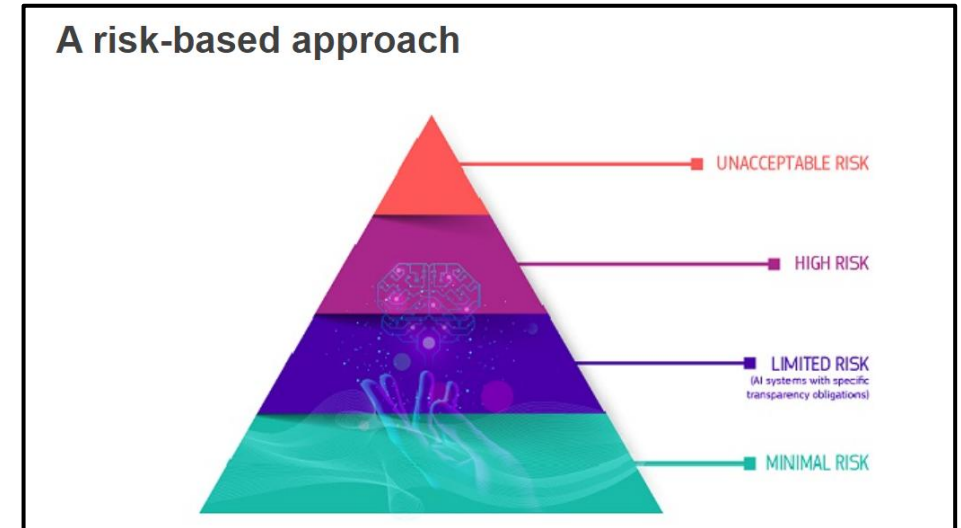
To show the safety and reliability issues that may exist in using AI technologies.

Not only academic and industrial interest

Worldwide efforts by governments and organizations to standardize the adoption of AI-based systems.

Examples are:

- International standards
 - ISO IEC TR 5469 (Artificial intelligence – Functional safety and AI systems)
 - ...
- U.S. National AI Initiative
 - Rules to use AI in the public and private sectors
- EU AI Act
 - Risk-based approach
 - Requirements for high-risk systems.



Goals of the AI Safety Course

- To understand potential risks and challenges associated with the development and deployment of AI systems in safety-critical systems
- To understand the vulnerabilities and fault tolerant proprieties of AI technologies
- To study existing safety and reliability assessments approaches
- To study existing safety measures for detecting and mitigating the occurrence of random hardware faults
- To encourage students to incorporate principles of AI safety into their own research, development, and decision-making processes
- To foster your curiosity
- ...

A bit of history – fathers of AI

Artificial intelligence has its roots in philosophical ideas of the philosopher and mathematician **Gottfried Leibniz** who lived in the 17th and 18th centuries:

- *Characteristica universalis*:
 - a language for explicating rational thinking, so precisely, a machine could replicate it
 - the idea of a universal language that could be used to represent all human knowledge in a symbolic form
 - precursor to modern computing and information processing.

Leibniz did not use the term “computation” in the modern sense, as the concept of a digital computer did not exist during his time.



Fathers of AI – 19th century

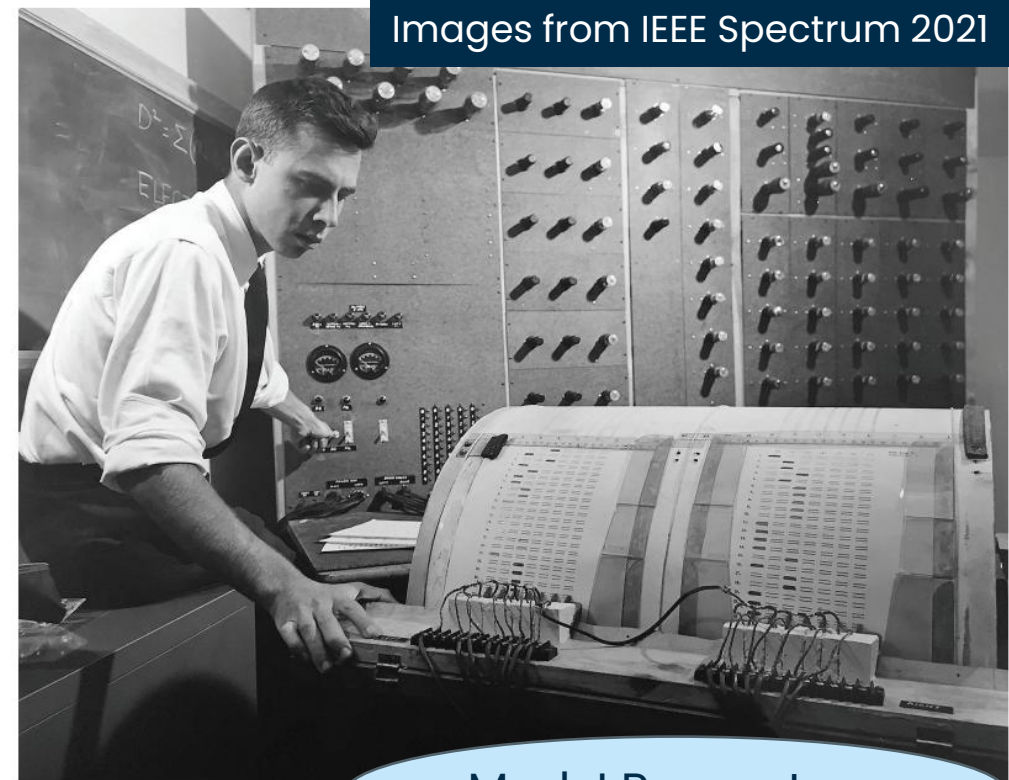
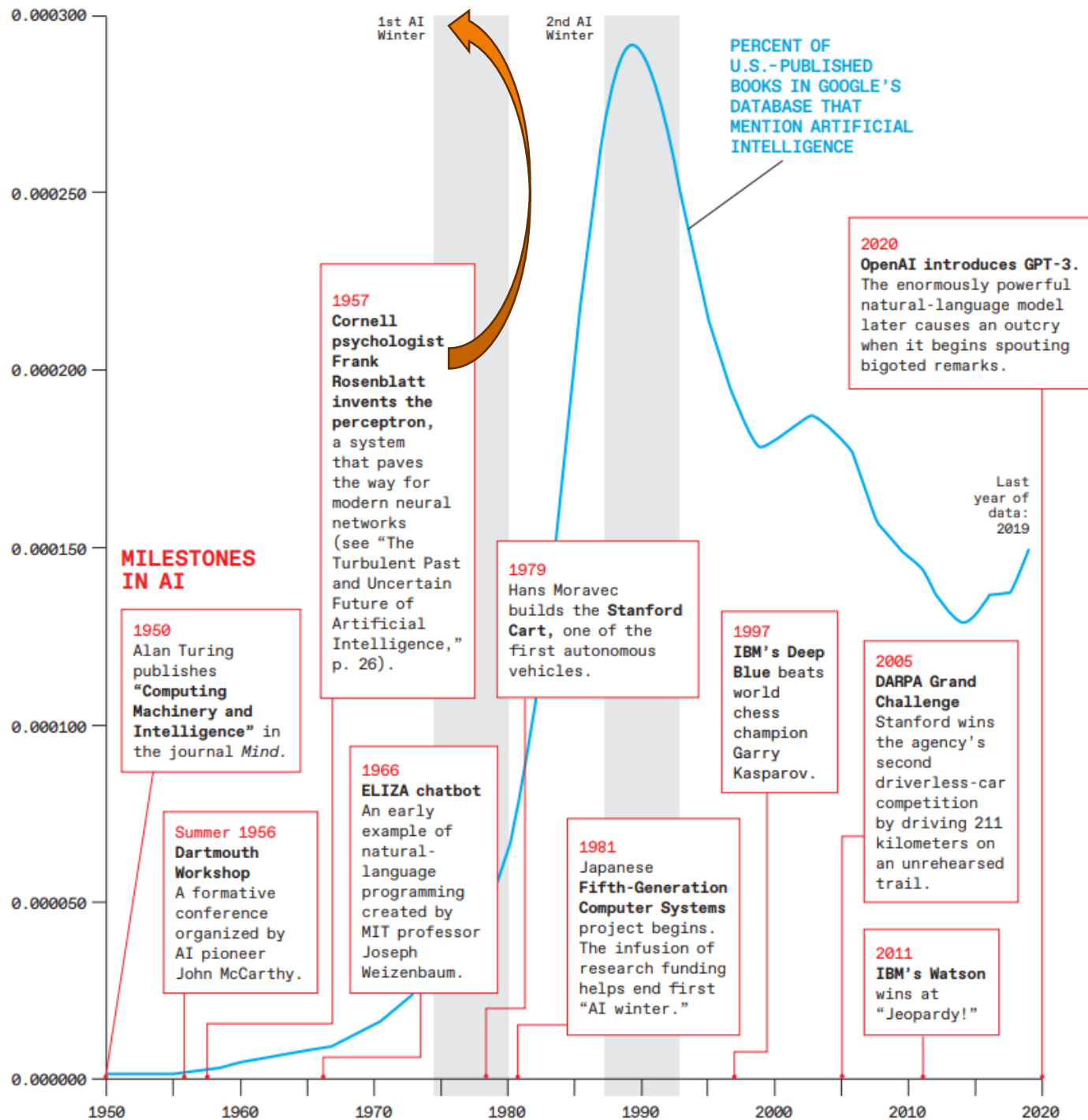
To impersonate thinking in a machine with logical rules.

- “*System of Logic*” by John Stuart Mill published in **1843**, where for the first time in history, logic is explored in terms of a manifestation of a mental process
- “*Laws of Thought*” by George Boole, published in **1854**. In his book, Boole systematically presented logic as a system of formal rules which turned out to be a major milestone in the reshaping of logic as a formal science
- **Artificial intelligence started out with logic.**

Fathers of AI – 20th century

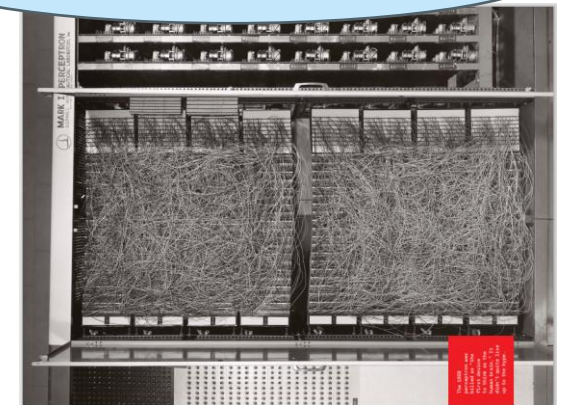
- Walter Pitts and Warren McCulloch published in **1943** a paper titled “*A Logical Calculus of Ideas Immanent in Nervous Activity*”
 - *They shared a lifelong interest in Leibniz, and they wanted to bring his ideas to **create a machine which could implement logical reasoning***
 - Their paper introduced the idea of the artificial neural network
- Alan Turing, the father of computing, marked the first step of the birth of artificial intelligence with his **1950** paper by introducing the Turing test to determine whether a computer can be regarded intelligent.

A.M. Turing, Computing machinery and intelligence. *Mind* 59(236), 433–460 (1950)



Mark I Perceptron

Frank Rosenblatt invented the perceptron, the first prototype of artificial neural network.

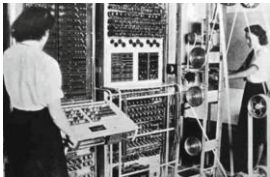


Reason behind the unsuccess of AI in 1950s

Hardware

Computers at that time were not able to deal with the complexity required by neural networks

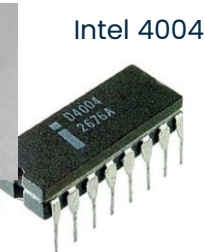
Colossus Mark II
World War II



Apollo Guidance Computer
MIT Instrumentation Laboratory

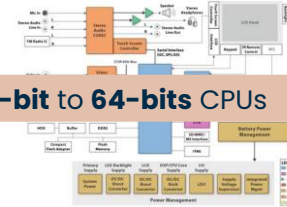


TMS1000 – Texas
Instrument



Intel 4004

Media player – Texas Instrument



From **4-bit** to **64-bit** CPUs

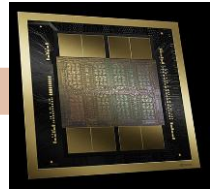
NVIDIA A100



TPU – Google



NVIDIA
Blackwell



1930

...

1960

1970

1980

1990

2000

2010

2024

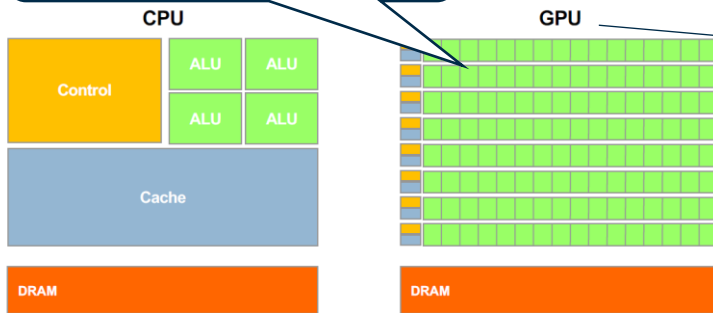
- The first single-chip microprocessor architecture was proposed by Texas Instrument in 1971 and it was a **4-bit CPU**. Microprocessor designs soon evolved from **4-bit to 8-bit CPUs**
- For 20 years the market was dominated by this kind of complexity
- In the last decades we have witnessed an increased complexity: with CPU sizes ranging from 16-, 32-, and 64-bit to 128-bit designs
- GPUs and AI-oriented architectures.

(Main) reason behind the success of AI today

Hardware

Modern hardware devices can deal with the complexity required by neural networks and AI

Same silicon area, more arithmetic operations

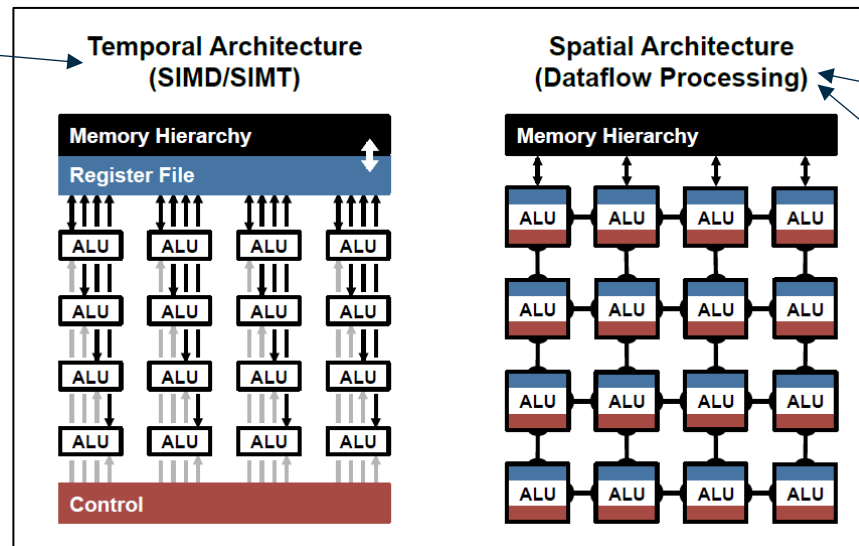


GPUs

Technology created specifically for graphics

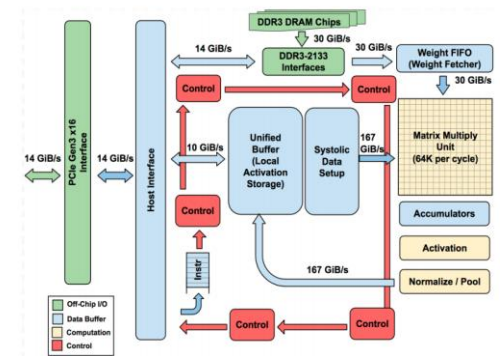
- GPUs have become critically important for a wide range of high-performance computing applications beyond graphics
- Specialized for **parallel intensive computation**.

Parallel Compute Paradigms



[V. Sze 2017 - ISCA Tutorial 2019]

Tensor Processing Units (TPUs) Domain-specific ASICs



NeuFlow accelerator Xilinx Virtex 6 FPGA platform

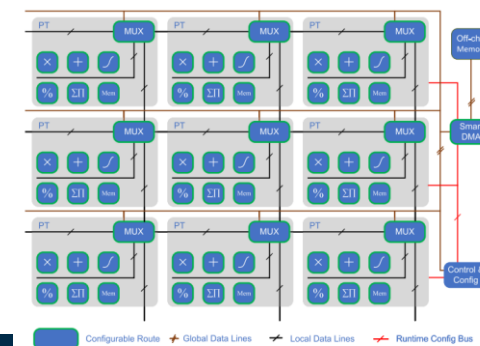


FIGURE 14. 2-D grid of Processing Tiles (PTs) in NeuFlow architecture, adopted from [84].

A bit of history – AI safety

- In **2011**, Roman Yampolskiy introduced the term "AI safety engineering", arguing that "the frequency and seriousness of risky events will steadily increase as AIs become more capable"
- In **2014**, the **Future of Life Institute** was founded by MIT cosmologist Max Tegmark, Skype co-founder Jaan Tallinn, DeepMind research scientist Viktoriya Krakovna, Tufts University postdoctoral scholar Meia Chita-Tegmark, and UCSC physicist Anthony Aguirre
 - ensuring AI remains safe, ethical and beneficial.

A bit of history – AI safety

- In **2015**, dozens of AI experts signed an **open letter** on artificial intelligence calling for research on the societal impacts of AI and outlining concrete directions
 - To date, the letter has been signed by over 11k people including Yann LeCun, Shane Legg, Yoshua Bengio, and Stuart Russell.



Title
Research Priorities for Robust and Beneficial
Artificial Intelligence: An Open Letter

<https://futureoflife.org/open-letter/ai-open-letter/>

Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter (2015)

...

Our AI systems must do what we want them to do

...

Because of the great potential of AI, it is important to research how to maximize its benefits while avoiding potential pitfalls

...

The progress in AI research makes it timely to focus research not only on making AI more capable, but also on maximizing the societal benefit of AI.

...

<https://futureoflife.org/open-letter/ai-open-letter/>

A bit of history – AI Safety

- In **2015**, a group of academics led by Professor Stuart Russell founded the Center for Human-Compatible AI at the University of California Berkeley
- The **AI safety summit** took place in November 2023, and focused on the risks of misuse and loss of control associated with frontier AI models
- Stanford Center for AI Safety
 - <https://aisafety.stanford.edu/index.html>

