UiO : University of Oslo

institutt for informatikk

IN3050/IN4050 - Introduction to Artificial Intelligence and Machine Learning

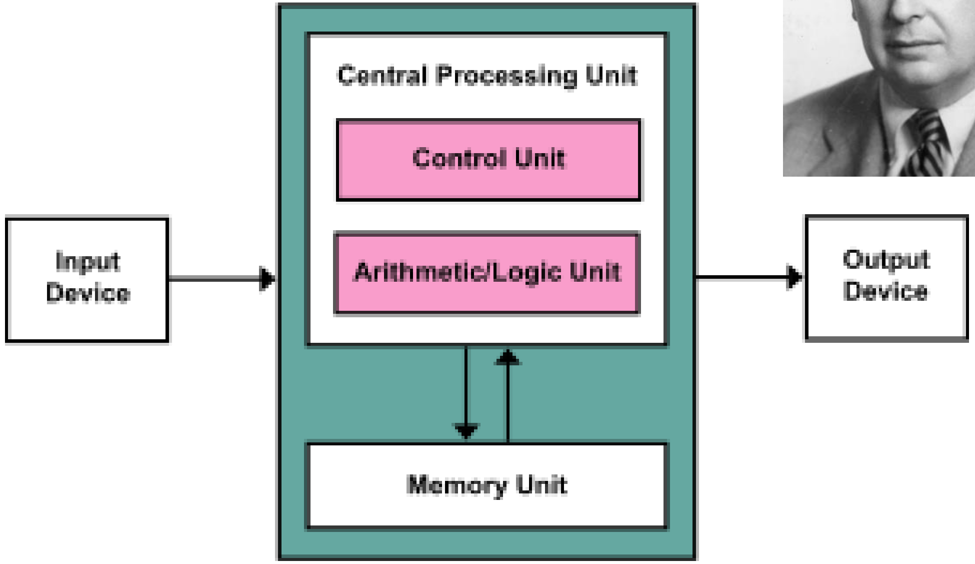
Lecture 13

Philosophy, Ethics, and the History of AI and Machine Learning

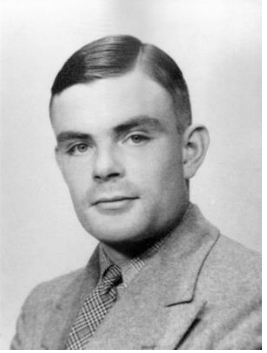
Pooya Zakeri

A quick history of AI (Philosophy of AI before AI itself)

- Invention of the computer
  - **Konrad Zuse (1941)**
  - **Clossus (1943)**
  - **ENIAC (1945)**
  - **John von Neumann architecture**
- The name of AI
- Invented by **John McCarthy** (Dartmouth Conference, 1956), Early AI Goals
  - aimed to model human cognition as rule-based computation over symbolic representations
  - **Alan Turing**: Machine could be intelligent (1950)

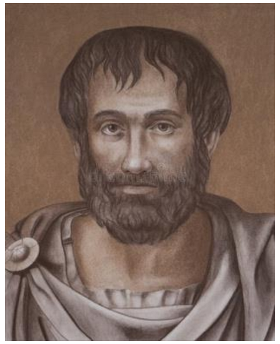


John von Neumann architecture



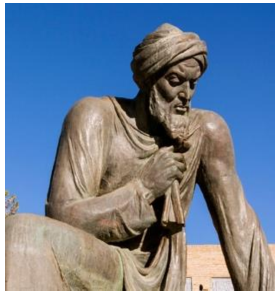
Quick history of AI (before Modernity)

Aristotle (384–322 BCE), *Logic*



- Formal Logic and Deduction: Aristotle's system of deductive reasoning is a direct precursor to the formal logic used in AI algorithms, particularly in rule-based systems where machines derive conclusions based on logical premises.
- Knowledge Representation: Aristotle's work on categorization and classification is conceptually linked to knowledge representation in AI, where data is structured in a way that allows machines to draw inferences.

Al-Khwarizmi (780 – 850 CE), *Algorithms*



- Algorithm and Algebra

Al-Kindi (801-837 CE), *Pattern recognition*

- Al-Kindi's work in cryptography involved analyzing patterns in language to decode encrypted messages.

Quick history of AI (before Modernity)

Avicenna (980-1037 CE), *Scientific reasoning*

- While he relied on deductive reasoning in philosophy, he used a different approach in medicine. Avicenna contributed inventively to the development of inductive logic and the development of a scientific method of open inquiry.

Ramon Llull (1232-1316 CE), *Combinatorial logic*

- He developed the Ars Magna (Great Art), a method for combining concepts mechanically to reach logical truths.

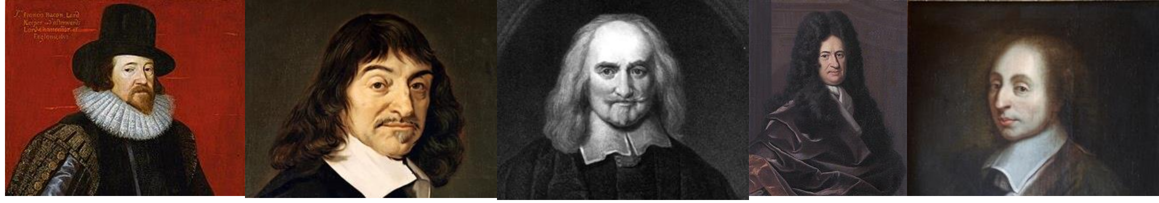
William of Occam (1287-1347 CE), *Simplicity and generalization*

- Occam’s razor—“Entities should not be multiplied beyond necessity,” or in simpler terms, “The simplest explanation is usually the best.

Jean Buridan (1300–1358 CE), *Autonomy and decision-making*

- Causality, logic, decision-making, and epistemology
- Predictive Determinism in ML
- Self-Motion and Autonomous Agency
- …

16th-17th Centuries



- Francis Bacon, *The birth of the scientific method*
  - He argued for the possibility of scientific knowledge based only upon inductive reasoning and careful observation of events in nature.
- René Descartes, *The mechanistic model of life—but with a dualist twist*
  - All things that are artificial are also natural,
  - E.g., An animal is a machine, but a very complex one. Human beings are partly machine, But also extra thinking (not machine) (Mind-Body problem)
  - Automata → Animals → Human Beings
- Thomas Hobbes, *Thinking as computation*
  - Proposed that thinking is a type of computation. In "Leviathan" (1651), he wrote that reasoning is nothing more than reckoning (i.e., adding and subtracting thoughts).
- Gottfried Wilhelm Leibniz, *The mechanization of reasoning*
  - Mechanization of Reasoning: Leibniz believed that human thought could be reduced to logical calculation. He envisioned a mechanical device that could perform calculations to solve problems of logic, similar to a modern computer or AI performing automated reasoning.
  - Early Symbolic Logic: Leibniz's ideas about creating a symbolic language to represent all human knowledge anticipate later developments in symbolic AI and formal logic, which are foundational to computational theories.
- Blaise Pascal, *The mechanization of calculation*
  - Created one of the first mechanical calculators called the Pascaline.

18th -19th Centuries

- The Industrial Revolution (late 18th century onwards)
  - Introduced the idea that human labor and thought could be mechanized or made systematic.
  - Encouraged a mechanistic worldview, seeing humans and society as systems that could be optimized.
  - Laid the foundation for later automation, computing machinery, and algorithmic thinking (e.g., Charles Babbage's Analytical Engine in the 19th century builds directly on this mindset).
  - Philosophically, it raised the question: If machines can replace physical labor, could they one day replace mental labor too?

Julien Offray de La Mettrie, *Materialism (Man a Machine)*

- In his 1748 book L'Homme Machine, he argued that the human mind is entirely a product of physical processes, no soul, no immaterial essence.
- A foundational philosophical statement of mechanistic materialism, the idea that thinking could in principle be replicated by a machine if we understood its workings.
- La Mettrie's materialism develops earlier mechanistic ideas—especially Hobbes's view of reasoning as calculation—and sets the stage for later debates (e.g., Turing on machine thought) and modern computational models.

Ada Lovelace:

A poet, the first person to ever write and publish a full set of instructions for a computing device, the world's first computer programmer.

Philosophy

φίλος (philos) ‘loving, friend of’ and σοφία (sophia) 'wisdom'. "*love of wisdom*.”

- Stoics
  - Philosophy as an exercise to train the mind and thereby achieve eudaimonia and flourish in life.
- Immanuel Kant
  - The task of philosophy is united by four questions:
    - *What can I know?*
    - *What may I hope?*
    - *What should I do?*
    - *What is the human being?*
- Edmund Husserl
  - Philosophy as a "rigorous science" investigating essence.
- Heidegger
  - Study of Beling

Other views:

- Many definitions of philosophy emphasize its intimate relation to science
- Philosophy is conceptual analysis, which involves finding the necessary and sufficient conditions for the application of concepts
- Thinking about thinking
- the meaning of life
- linguistic therapy (Ludwig Wittgenstein)
- concept creation (Gilles Deleuze)

Main Branches in Philosophy

- Metaphysics & ontology
  - Metaphysics explores the most fundamental aspects of reality, including existence, objects and their attributes, the relationships between wholes and parts, space and time, events, and causation.
- Epistemology (Theory of knowledge)
  - An effort to define the nature of knowledge, identify the kinds of knowledge that can be attained, and explore the methods of acquiring knowledge.
- Logic
  - Logic is the study of correct reasoning. It aims to understand how to distinguish good from bad arguments
- Ethics
  - Ethics studies what constitutes right conduct. It focuses on what we ought to do and what it would be best to do.

Other major branches of Philosophy

- Aesthetics
- Political philosophy
- Phenomenology
- Contemporary Philosophy
- Philosophy of science
- Philosophy of religion
- Philosophy of X
    - Philosophy of biology
    - Philosophy of economics
    - Philosophy of art
    - ….
    - ….
    - Philosophy of Technology
    - Philosophy of language
    - Philosophy of mind
    - Philosophy of AI/Machine Learning

What is AI?

- Philosophical Inquiry: Defining intelligence, cognition, and behavior in AI systems
- Scope: Difference between simulating vs. truly replicating intelligence

One foundational question for AI philosophy is: What is AI? Answering this requires defining concepts like intelligence and cognition. Is AI a tool to mimic human intelligence or a potential form of synthetic cognition? Distinguishing between simulating intelligence and achieving a true understanding is a central theme in AI philosophy.

Artificial Intelligence

The study of how to make computers do things that, for the moment, humans do better.

Artificial

- φύσις (Fusis/Physics) ←→ : τέχνη (Techne)
- Nature ←→ Art (craftsmanship) (Aristotle, 4BC)
- Fusis
  - Substantial form
  - Internal Principal of (Self) Development
- Techne
  - Imitate, not create
  - No generative creativity

Intelligence

- Understanding, knowledge, power of discerning; art, skill, taste
- Intelligence is the capacity for flexible, goal-oriented action in diverse environments. Success in achieving goals, influenced by both the agent's abilities and the environment, reflects an agent's level of intelligence.
- Mostly, AI measures intelligence as the ability to pursue goals across a broad spectrum of conditions, an approach rooted in classical decision theory's emphasis on maximizing utility.

Question 1: What Can AI Do/Know?

- **Focus:** Abilities and limits of AI, simulation vs. true cognition
- **Considerations:** Can AI reach human cognitive capabilities?

The question, *What can AI do?*
is both technical and philosophical. It asks whether AI can develop human-like abilities—such as abstract thinking, empathy, and creativity. Understanding AI’s potential is essential for assessing the long-term goals and limitations of AI research.

Can machines think?

- **The Mind–Body Problem**
  - Core Questions:
    - What is that we call “mind”?
    - Is the mind identical to the brain, or something more?
    - What exactly are thoughts?
    - How are thoughts and conscious experiences connected to physical processes?
- Why is it a problem?
  - We still struggle to explain how **subjective experience** and **consciousness** arise from matter.
  - This gap between the **mental** and the **physical** remains one of the deepest puzzles in philosophy and science.

- Key Concepts:

  Consciousness: awareness or being conscious of something.

  Intentionality: the “aboutness” of mental states (thoughts are about things).

  Subjectivity: the first-person perspective — what it feels like.

  Causation: how mental events relate to physical actions.

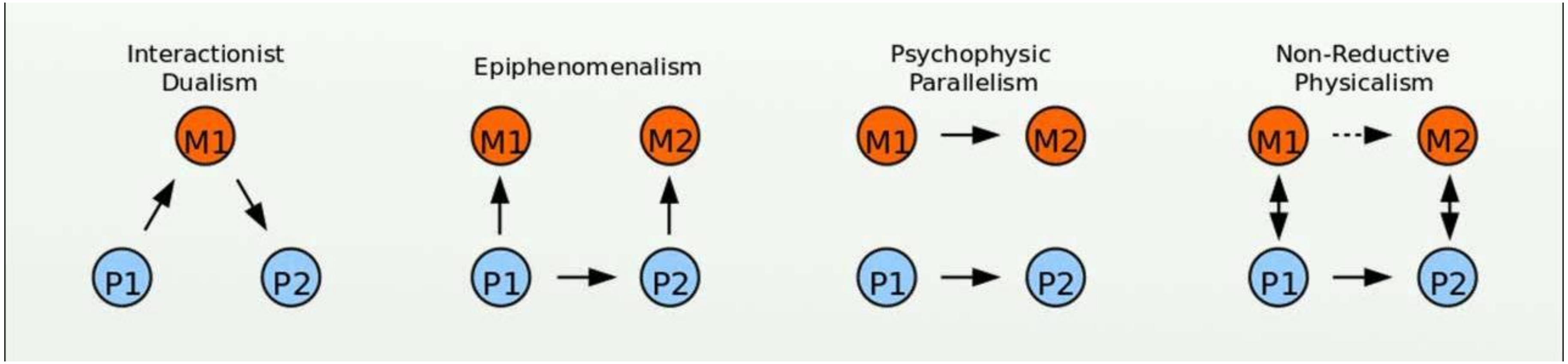
  Qualia: the phenomenal qualities of experience; what it feels like to see red, taste coffee, or feel pain.

  To have qualia, you have to be conscious.

  Feelings and experiences vary widely.

Dualism vs. Materialism

- **Materialism (physicalism):** mental states are physical states—states of the brain, which is a physical thing
  - Physical things can be accounted for by physics, biology, chemistry, etc.
- **Dualism:** mental states are not states of any physical thing; they are non-physical entities.
  - Non-physical things cannot be investigated by the sciences.



Dualist Theories

Substance Dualism: Each mind is a distinct nonphysical thing, independent of any physical body.

- The brain is a physical substance
- The mind is a non-physical substance

Property Dualism: There aren't two kinds of substances, the mind and the brain, but two kinds of ***properties***—physical and non-physical properties.

- *Epiphenomenalism:* ***mental properties are by-products of physical processes but have no causal power back on the body.***
- *Interactionism / Radical Emergentism*
- Mental properties are **emergent**—they arise when physical systems reach sufficient complexity.
- Mental states are **irreducible** and cannot be fully predicted or explained by current physical science.
- These emergent mental properties can **influence** the physical system (downward causation).

Materialism doctrines:

- Reductive theories:
  - *Behaviorism:* Mental states are defined by observable behavior and responses to stimuli.
  - *Identity Theory:* Mental states are identical to brain states
  - *Functionalism:* Mental states are defined by their causal roles, not their physical composition.
  - *Eliminative Materialism:* Advances in neuroscience reveal to us that theory about the mind is a bad theory.
- Non-Reductive theories
  - *Emergentism*: The mind is what the brain does, but you cannot reduce mental explanations to simple physical laws.
    - Everything is physical, yet mental properties depend on physical ones without being neatly reducible to them.
    - Higher-level organization produces **new patterns and causal powers**, but nothing non-physical is added.

What is Functionalism

- Mental states are defined by their **causal roles**, not by what they are made of.
- **Key idea:**
  - What matters is **what the system does**, not the material it is built from.
    - A *heart* is anything that **pumps blood**, regardless of material
    - A *thermostat* is anything that **regulates temperature**
    - What makes something a tea kettle or a coffee maker is its ability to perform a certain function, boiling water or making coffee, not its specific physical structure or material composition

Multiple Realizability

- **The multiple realizability of mental properties**
    - If mind = function,
    - then mental states can be **realized in different physical systems**:
        - carbon-based brains
        - silicon chips
        - neural nets
        - alien physiology
    - **Only one requirement:**
    - The system must be **embodied** in some causal medium—but the “body” can be different.

Realizers of Functional Properties

- **Realizers**
  - A functional property can be implemented by **many different mechanisms**.
  - **Example:**
    - A heart = anything that performs the **function** of pumping blood
    - A mind = anything that performs the **functions** of perception, memory, reasoning, planning, etc.
  - **Causal work can be done in many ways.**
  - Therefore, mental properties can have **indefinitely many realizers**.

Why Functionalism Supports AI

Functionalism is a foundational idea for Artificial Intelligence

  - If mental states are functions,
  - Then it is plausible to **model them on a computer**.

- Computational systems can implement:
  - symbolic reasoning
  - learning
  - memory storage
  - action selection
  - goal-directed planning

Brain as a Computer

- Functionalism acknowledges:
    - The brain’s **complex causal structure** matters
    - But what matters **is the computation**, not the biology
    - Our cognitive capacities come from the **functions** the brain performs

- **Conclusion:**
    - The brain is our mind **because it is a computing machine**,
    - not because it is made of carbon-based biological tissue.

Strategies for modeling cognitions

- Symbolic Processing
- Connectionism

Symbolic Processing in AI (Making a Mind)

- **Definition:**
  - **Symbolic Processing** views cognition as the **manipulation of symbols** according **to explicit rules**—much like traditional computers.
- **Key Characteristics:**
  - **Representation:** Uses discrete symbols to represent knowledge (e.g., "dog" represents the concept of a dog).
  - **Rule-Based:** Cognitive processes follow clear, logical rules and algorithms.
  - **Serial Processing:** Information is processed in a step-by-step manner.
  - Operates on **deductive reasoning**, where conclusions are derived from general rules or premises.

Strengths:

Effective for tasks requiring ***logical reasoning, language understanding***, *and* ***rule-based decision-making***.

Easily ***interpretable*** and ***explainable*** due to explicit rules.

Limitations:

Struggles with pattern recognition and perceptual tasks.

Lacks the flexibility and adaptability found in human cognition.

Connectionism in AI (Modeling Brain)

- **Definition:**
  - **Connectionism** models cognition through **neural networks**, where **knowledge** is stored in the **patterns of connections** among many simple units (like neurons).
- **Key Characteristics:**
  - **Distributed Representation**: Knowledge is encoded in connection strengths rather than discrete symbols, not explicit rules.
  - **Learning from Data**: Uses large datasets to identify patterns and make predictions.
  - **Learning-Based**: Connecxtion strengths are adjusted through learning (e.g., backpropagation).
  - **Parallel Processing**: Many units process information simultaneously, akin to how the brain works.
  - Connectionism relies on **inductive reasoning**, learning patterns from specific examples and generalizing to new cases.

Strengths:

Excellent for **pattern recognition**, **perception**, and **learning from data**.

More **flexible** and able to generalize from examples.

Limitations:

Lacks explicit rules, making it **hard to interpret** and understand decision-making.

Less effective for tasks that require **logical reasoning** and **structured manipulation** of symbols.

Connectionism vs. Evolutionary Computing

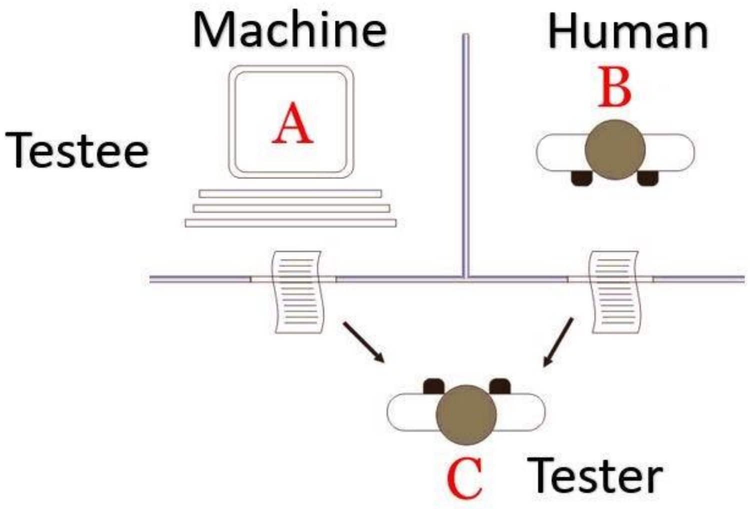
- Early attempts at **classical AI** (rule-based systems, symbolic reasoning) failed to replicate complex human intelligence.
- **Non-classical approaches**, while innovative, struggled to handle tasks that required flexibility and adaptability.
- **Two prominent fields emerged in response:**
  1. **Connectionism** *(Neural Networks, Parallel Processing)*
     - Focused on **neural networks** and **parallel processing**, modeling intelligence as a distributed system of nodes that can learn and adapt.
     - Inspired by how the human brain processes information through interconnected neurons.
     - Connectionism contributed to the rise of modern **machine learning** and **deep neural networks**.
  2. **Evolutionary Computing** *(Inspired by Biological Evolution)*
     - *Based on the idea of natural selection*

Personhood and Consciousness

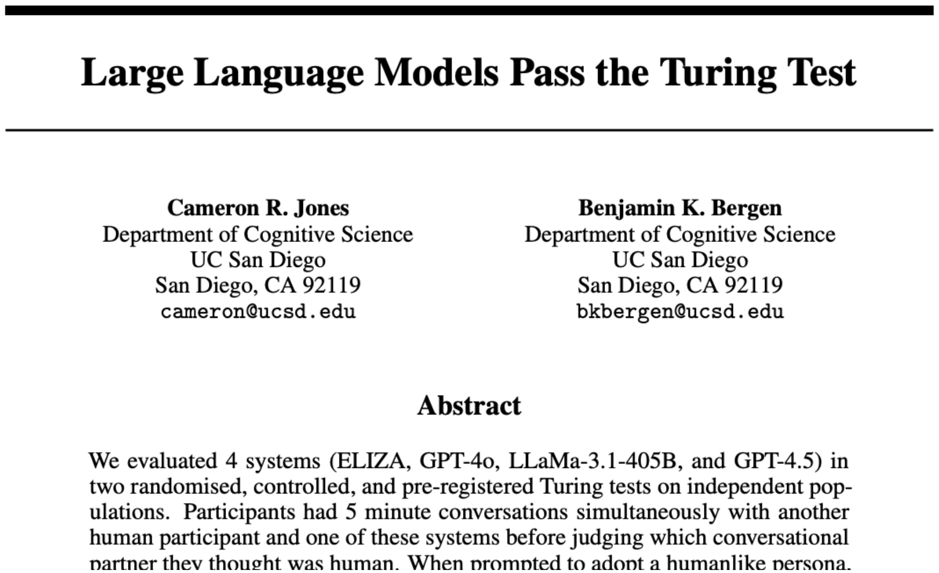
- **Types of Consciousness**
  - **Awareness**: Access to cognitive states
  - **Phenomenal Consciousness**: Subjective experience
- **The “What It’s Like” Argument**
  - **Thought Experiments**: Nagel’s bat (*What Is It Like to Be a Bat?*)
  - **Implication**: Limits of understanding subjective experience, raising doubts about replicating consciousness in AI.
- **The Zombie Argument**
  - **Definition**: Physically identical but without consciousness
  - **Implication for AI**: Awareness without subjective experience
- **The Self in AI**
  - **Human Self**: Continuity, memory, and identity
  - **AI Challenge**: Distributed components lack a cohesive self
    - AI lacks such inherent self-concept, making personal responsibility in AI systems challenging to justify.

The Turing Test

- Alan Turing’s 1950 proposal, known as the **Turing Test**, aimed to define machine intelligence through behavior: In this test, a human judge interacts with an unseen interlocutor, which could be either a human or a machine. If the judge cannot reliably distinguish between the machine and the human, the machine is considered to have ‘passed’ the test.



- **Claim:** Replace “What is thinking?” with an **operational** test: if a machine’s replies are indistinguishable from a human’s in free conversation, count that as intelligence.
- **Point:** Focus on **behavioral evidence**, not inner essence.
- **Strengths:** Measurable, task-agnostic, pushes progress in NLP/interaction.
- **Weaknesses:** Easy to “game” (prompting, personas), anthropomorphic bias, says little about **understanding** or **consciousness**.



Large Language Models Pass the Turing Test

Cameron R. Jones
Department of Cognitive Science
UC San Diego
San Diego, CA 92119
cameron@ucsd.edu

Benjamin K. Bergen
Department of Cognitive Science
UC San Diego
San Diego, CA 92119
bkbergen@ucsd.edu

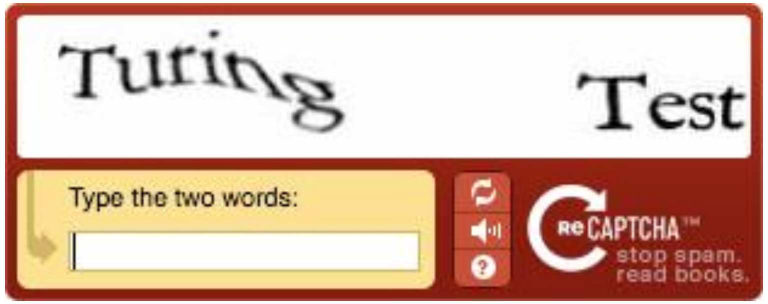
Abstract

We evaluated 4 systems (ELIZA, GPT-4o, LLaMa-3.1-405B, and GPT-4.5) in two randomised, controlled, and pre-registered Turing tests on independent populations. Participants had 5 minute conversations simultaneously with another human participant and one of these systems before judging which conversational partner they thought was human. When prompted to adopt a humanlike persona,

Since the mid-2020s, several large language models such as ChatGPT have passed modern, rigorous variants of the Turing test.

- A 2025 study ran two large, randomized Turing tests comparing humans to four AI systems: **ELIZA, GPT-4o, LLaMa-3.1-405B, and GPT-4.5**.
- Participants held two simultaneous 5-minute conversations—one with a human and one with an AI—and had to identify which was the human.
- **GPT-4.5**, when instructed to act human, was judged to be the human **73% of the time**, outperforming real human participants.
- **LLaMa-3.1** scored **56%**, similar to humans. Older models, **ELIZA** and **GPT-4o**, scored far below chance.
- This study provides the **first empirical demonstration of an AI system passing a standard three-party Turing test**, raising new questions about what such a pass really tells us about intelligence, understanding, and the social impact of advanced LLMs.

The Turing Test



Reverse Turing Test

- A Turing test with **roles reversed**:
  the *machine* must decide whether **its partner is human or machine**.
- Explored by scholars like **Peter Swirski** and **R. D. Hinshelwood**, who describe the mind as a “mind-recognizing apparatus.”
- Raises the question:
  Can a machine detect human mentality?

CAPTCHA

- A practical, everyday **reverse Turing test**.
- Websites ask users to solve tasks that humans find easy but bots find hard.
- Idea: if you can solve the distorted text or image, you’re probably human.
- Modern AI systems have increasingly ***defeated CAPTCHAs*** with very high accuracy.

But does passing the Turing Test truly mean a machine is intelligent or conscious, or does it simply mean it can mimic human responses convincingly?

Understanding Strong AI vs. Weak AI

John Searle introduced **the concept of Strong AI** to distinguish between two types of artificial intelligence: **Weak AI** and **Strong AI**

- **Weak AI:** (computers simulating cognitive processes)
  - Weak AI means that AI merely simulates mental states.

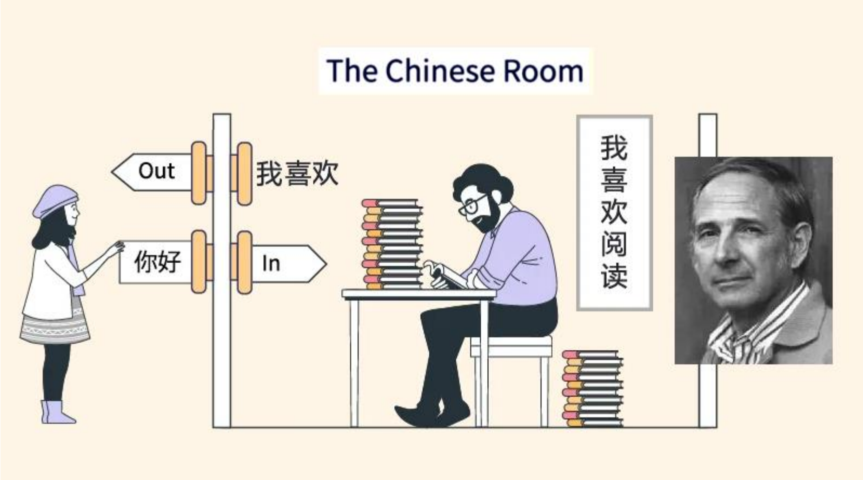
  It holds that AI can simulate human cognition and perform tasks similar to what humans can do, but this simulation doesn't mean the AI system actually understands or possesses consciousness. Weak AI is about useful tools rather than true minds.

- **Strong AI:** (computers truly having minds).
  - Strong AI claims that if a computer passes the Turing it can be considered mental state.

  An appropriately programmed computer with the right inputs and outputs wouldn't just simulate a mind—it would have a mind in the same way humans do. In other words, it would have genuine understanding, consciousness, and intentionality (the ability to have thoughts about things).

Chinese Room Argument (John Searle)

- Imagine a person (Searle himself, for example) who does not understand Chinese is locked in a room.
- Inside the room, they have a set of instructions in English for manipulating Chinese characters, and these instructions are so detailed that they allow the person to produce appropriate responses to Chinese questions.
- Outsiders would send in Chinese questions, and by following the instructions, Searle could produce correct Chinese responses, despite not understanding a single word of Chinese.



Point of the Argument: Searle argues that even though the person in the room produces correct responses in Chinese, they don’t actually understand Chinese. They’re simply manipulating symbols based on a set of rules without any comprehension of the meaning behind them.

Implication for AI: Searle uses this analogy to claim that a computer running a program (no matter how sophisticated) is like the person in the room. It can process inputs and produce outputs ***without any real understanding***. In other words, the computer ***simulates understanding*** but ***does not possess it***. According to Searle, this is the fundamental flaw in the Strong AI hypothesis.

Differences Between the Turing Test and Chinese Room Argument

Concept	Turing Test	Chinese Room Argument
Proponent	Alan Turing (1950)	John Searle (1980)
Goal	To evaluate a machine’s ability to imitate human-like responses.	To demonstrate that imitation of human responses is not equivalent to understanding.
Focus	External behavior — can a machine’s responses be indistinguishable from a human’s?	Internal understanding — does the machine genuinely “understand” the responses it produces?
Implication	Passing the Turing Test could suggest intelligence or human-like behavior.	Passing the Turing Test is insufficient for proving understanding or consciousness.
Criticism of AI	Limited: suggests intelligence might be achievable through sufficient mimicry.	Strong: argues machines are limited to processing rules and lack true understanding or consciousness.

The Danger of Anthropomorphic Language in AI

- *Anthropomorphism* is the attribution of human traits, feelings, or behaviors to inanimate objects, non-human animals, or nature.
- We often say robots “see,” “think,” or “recognize,” but these words imply human abilities the systems do not have.
- Using human-like language leads to misaligned expectations, unclear specifications, and unpredictable behavior.
- Even simple tasks like “pick up an apple” hide complex sensing, detection, and control steps.
- Clear terms prevent false expectations and failures

Consciousness vs Computation

- Different views on the subject of computationality and consciousness (Penrose):
  - **Strong AI:** Consciousness and other mental states consist entirely of computational process
  - **Weak AI:** Brain processes cause consciousness, and these processes can be simulated on a computer. However, simulation itself does not guarantee consciousness.
  - **Lucas–Penrose:** Brain processes cause consciousness, but these processes cannot even be simulated computationally.
  - Consciousness cannot be explained computationally(consciousness is not a topic for science) (**Mystics**)

What Is AGI?

Artificial General Intelligence (AGI)

- A hypothetical machine with:
  - **Human-level** ability across *any* cognitive task
  - **Transfer learning** across domains
  - **Autonomous goal-setting**
  - **Flexible reasoning**, creativity, and planning

Why AGI is controversial

- No clear **definition**
- No agreed **test** for “general intelligence”
- Confusion between **performance** and **understanding**
- Philosophical disputes about whether mind = computation

Key issue: Does “general intelligence” even form a natural category?

Major Criticisms of AGI

1. Conceptual Vagueness

- AGI lacks a precise definition; we cannot test what we cannot define.

2.*Anthropocentrism*

- AGI assumes human intelligence is the gold standard; ignores non-human forms of cognition.

*3.*Overextension of Computation

- Assumes all aspects of mind (e.g., consciousness, meaning, intentionality) are computationally realizable.

4.The "Generalization Fallacy"

- High performance across many benchmarks ≠ general intelligence.
- Systems may still lack grounding, causal reasoning, or transferability.

5.Hardware & Architecture Limits

- Brains and silicon operate with radically different constraints; "general intelligence" might not be substrate-independent.

6.Philosophical Objections

- Can machines truly **understand**?
- Can machines have **aboutness** or **intentionality**?
- Is general intelligence even **algorithmic**?

“Strong AI”: Two Different Meanings

1) Searle’s “Strong AI” (Philosophy)

- Claim: A program with the right functions literally has a mind.
- Involves questions of ***consciousness***, ***understanding***, and i***ntentionality***.
- Most philosophers (Searle included) reject this claim.

2) Futurist / Popular “Strong AI”

- Means human-level AGI:
  an artificial agent as intelligent as a human across tasks.
- Does ***not*** require consciousness unless one assumes humans need consciousness for intelligence.

Mainstream AI Research View

Focuses on ***behavior and performance***, not inner experience.

If a system behaves intelligently, researchers call it intelligent—simulation vs “real mind” is considered irrelevant.

For them, **Weak AI = AGI is possible**, Strong AI is a philosophical question they ignore.

Classical vs. Technical AI:
Creating intelligent behavior vs. practical AI applications

- Classical AI (focused on replicating human-like cognition)
  - The pursuit of AI as a means to create machines that can replicate or simulate human-like cognition and reasoning processes.
- Technical AI (focusing on practical applications, including machine learning and decision-making)

The more applied approaches focus on developing methods in computer science for tasks such as perception, modeling, and action, involving techniques like machine learning, probabilistic reasoning, and optimization.

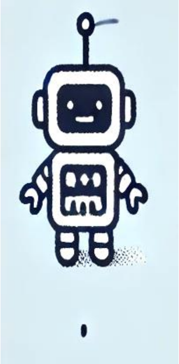
Hubert Dreyfus: A Critic of Artificial Intelligence

- Dreyfus's Main Critique: The Limits of Formal Rules
  - Early AI research overemphasized the idea that human intelligence could be replicated through **formal rules** and **symbolic manipulation**.
  - **GOFAI (Good Old-Fashioned AI)**: This **symbolic AI** approach assumed that intelligence could be achieved through **rules and representations**.
  - **Human Complexity**: Dreyfus argued that **human thought** is too complex, nuanced, and **context-dependent** to be captured by formal systems.
  - **Embodied Intelligence**: He believed AI's reliance on formal rules overlooks the **embodied, intuitive aspects** of human intelligence, which involve context and lived experience.

Modern ML vs. Dreyfus's Critique: Dreyfus's critiques initially targeted symbolic AI, but similar issues apply to **modern machine learning** (e.g., neural networks).

- **Lack of Genuine Understanding**: Neural networks recognize patterns or generate text but **do not achieve true understanding**.
- **Absence of Embodied Knowledge**: Dreyfus would argue that ML models process data without the **situated, embodied knowledge** and **experiential context** that are central to human intelligence.

The Frame Problem 1/3



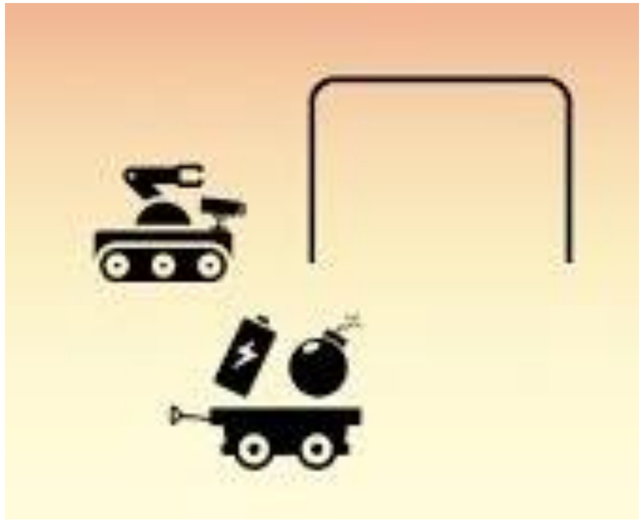
Description: Difficulty in determining relevance

Challenge: Context-sensitivity vs. logical inference

A thought experiment by **Daniel Dennett**,

Scenario:

- **Goal**: A robot’s task is to retrieve a ***battery*** to power itself.
- **Complication:** The battery is located near a ***bomb***, posing a risk to the robot.
- **Decision:** The robot must decide whether it can retrieve the battery safely without triggering the bomb.



The frame problem is a philosophical and computational challenge that arises when trying to determine what is *relevant* in decision-making and reasoning, especially for artificial intelligence systems. In simple terms, it is the problem of how a rational agent (like a robot or AI) can identify what aspects of a situation are relevant or irrelevant to its decision, without needing to exhaustively consider every possible factor or outcome.

The Frame Problem 2/3

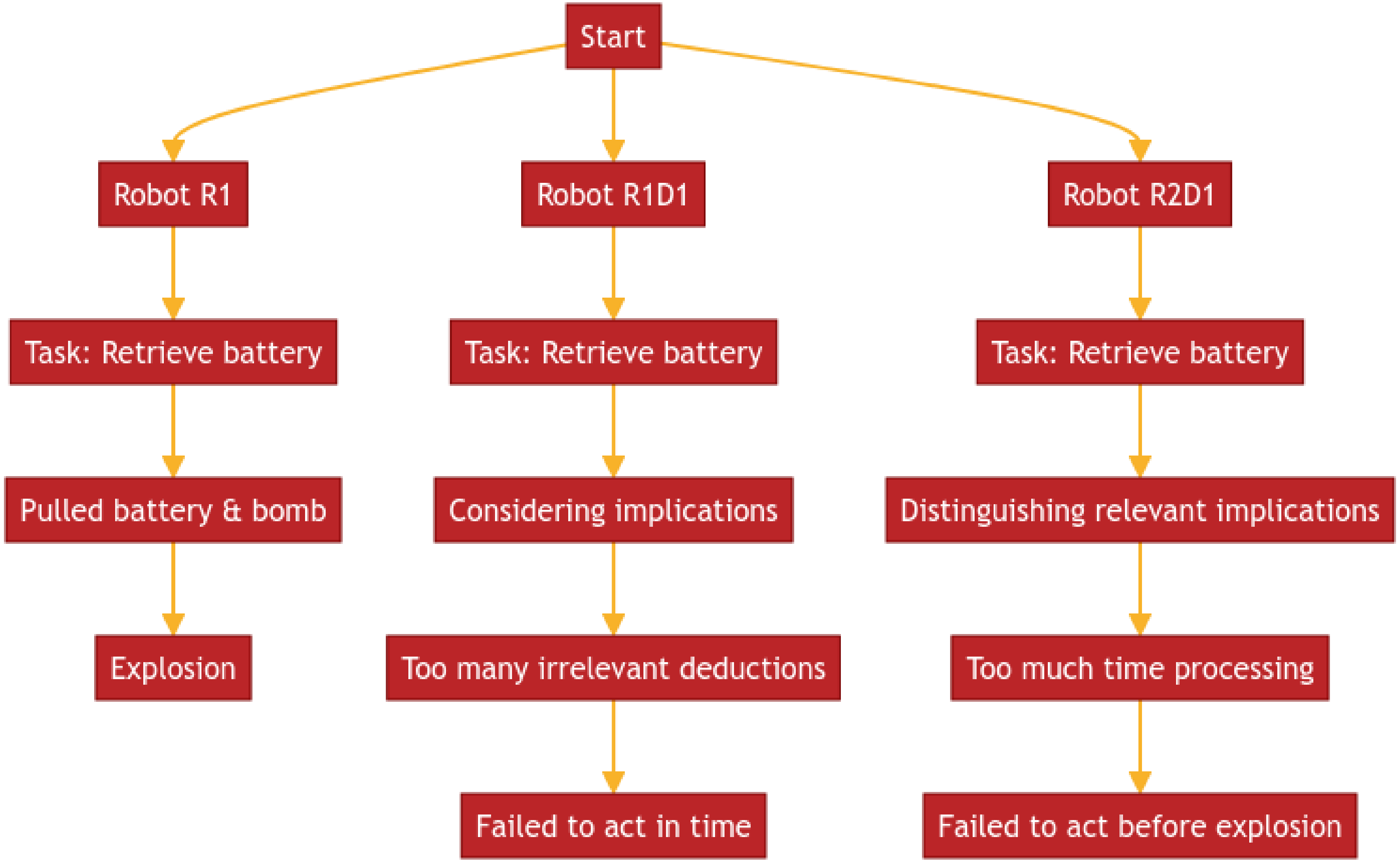
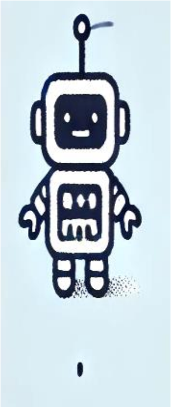


Image source and further reading

The Frame Problem 3/3



Hubbert Dreyfus highlighted the ‘frame problem’ as a central issue for AI. The frame problem refers to the difficulty that AI systems face in determining which aspects of a situation are ***relevant*** or ***irrelevant*** in a given context.

Humans can intuitively decide what to pay attention to in new or unexpected situations, but AI struggles with this. This is because human experience is holistic, whereas AI systems typically break down situations into predefined rules or data sets. According to Dreyfus, this limitation prevents AI from achieving true understanding.

The Frame Problem and Deep Learning Models

- **Relevance in ChatGPT:**
  - ChatGPT is trained to determine relevance based on patterns in its training data, but it doesn’t have a deep understanding of why certain information is important. It can make associations based on previous examples but lacks an innate ability to judge what is truly significant in a novel situation.

- **Dreyfus’s Perspective:**
  - Dreyfus would argue that this limitation persists even in deep learning models like ChatGPT. The inability to grasp what is relevant in an intuitive sense means that deep learning models do not achieve the kind of situational awareness that characterizes human intelligence.

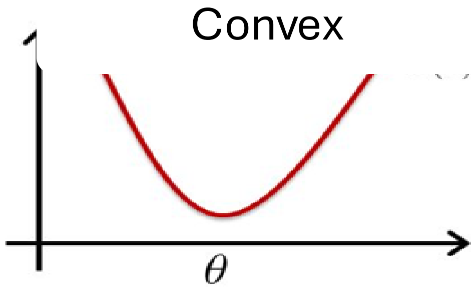
Machine Learning and Deep L earning History

1960's **1980's** 1990's 2000's 2010's 2020's

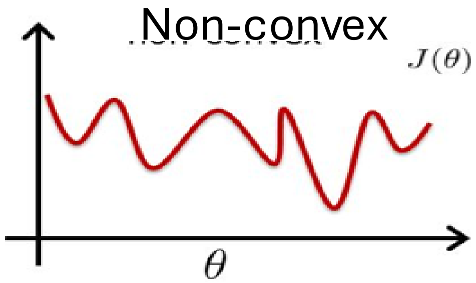
Perceptron **Backpropagation**

But what was wrong with back-propagation?

Convex



Non-convex



Often nonconvex, esp. in deep learning.

- We didn't collect enough labeled data.
- We didn't have fast enough computers.
- We didn't initialize the weights correctly

- If we fix these three problems, it works really well
- Also, the Early Popularity of Genetic Algorithm(1970s-1980s) and The field gained momentum.

Abandoning neural networks

- 1991: Sepp Hochreiter identifies the problem of ***vanishing gradient*** which can make the learning of deep neural network extremely slow and almost impractical (also called diffusion of gradient problem)

- Gradients becoming very small in deep networks for randomly initialized weights. When using backpropagation to compute the derivatives, the gradients that are backpropagated rapidly decrease in magnitude as the depth of the network increases.

Deep learning is data hungry? Actually lable-data hungry

Deep learning is Black-box

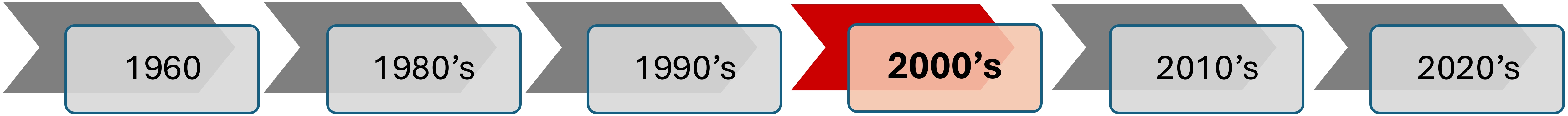
Do not trust a black box.

Rise of Evolutionary Computation, hybrid algorithms combining GAs with other optimization techniques, enhancing.

By the late 1990s and early 2000s, GAs were recognized as effective tools for optimization and adaptation problems where traditional methods struggled.

Other ML approaches (gray box) work better

Domination **of Support Vector Machine**…


1960

1980’s

1990’s

2000’s

2010’s

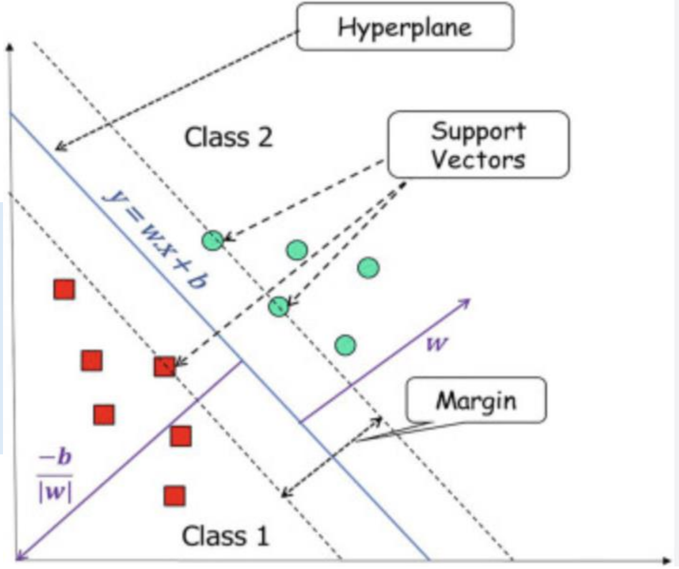
2020’s

Domination of SVM and Data Fusion

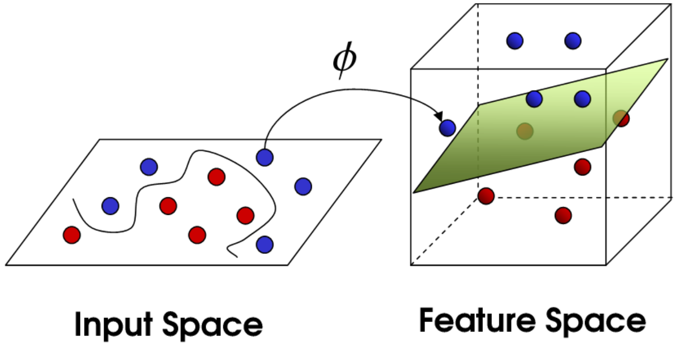
- SVM performs very well,
  - particularly with Kernel trick and Sequential minimal optimization (SMO) optimization
- Ensemble learning
  - In 2006 A new rival emerges (Random Forest)
- More data views becomes available
- Data fusion Era
  - While a single data source might not be sufficiently informative, the integration of several data sources leads to more ***accurate predictions***.
  - Flexibility of Kernel-based methods at different levels of data realization
    - Early Fusion
    - Intermediate Fusion
    - Late Fusion

Support Vector Machines

Support Vector Machine (SVM) is a supervised machine learning algorithm used primarily for classification, although it can be adapted for regression tasks as well. The main idea behind SVM is to find a hyperplane that best separates data points from different classes in a way that maximizes the margin between them.

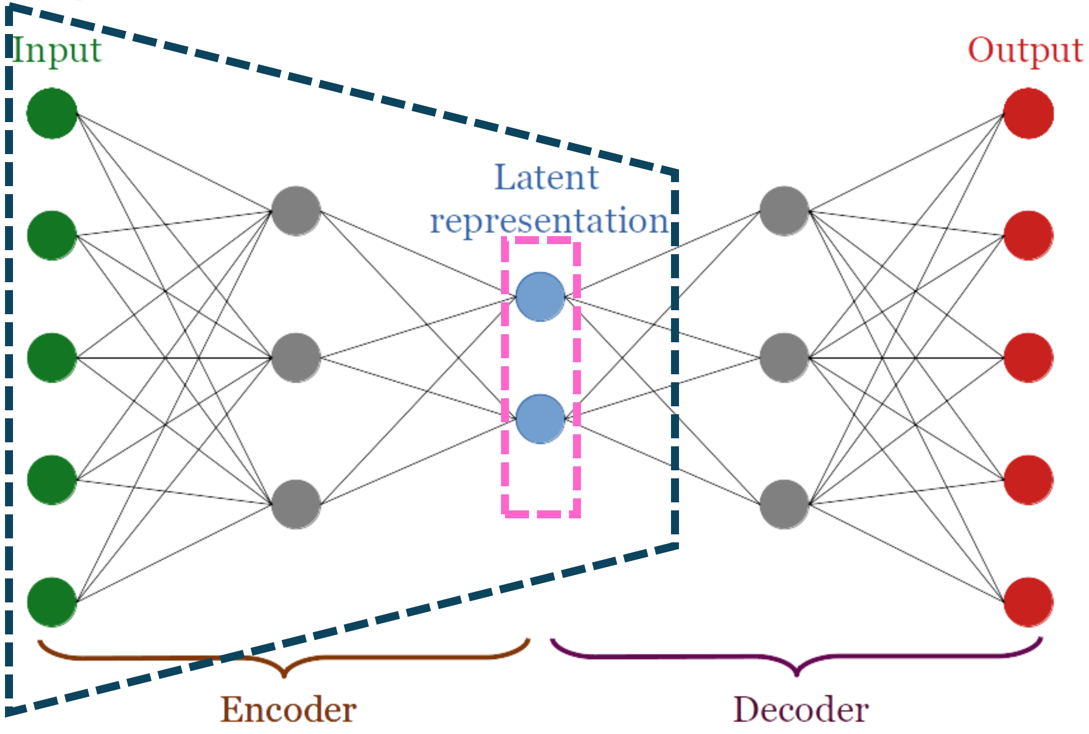


- Key elements of SVM:
- **Hyperplane**: This is a decision boundary that separates classes. In a 2D space, it's a line, while in 3D, it's a plane.
- **Margin**: The SVM algorithm seeks to maximize the distance (margin) between the hyperplane and the nearest data points from each class (support vectors). A larger margin generally means better generalization to new data.
- **Support Vectors**: These are the data points closest to the hyperplane and are crucial in defining the boundary. Only these points influence the position and orientation of the hyperplane.
- **Kernel Trick**: For data that isn't linearly separable in its original form, SVM uses kernel functions (like polynomial or radial basis functions) to transform data into a higher-dimensional space where a linear boundary can then separate the classes.



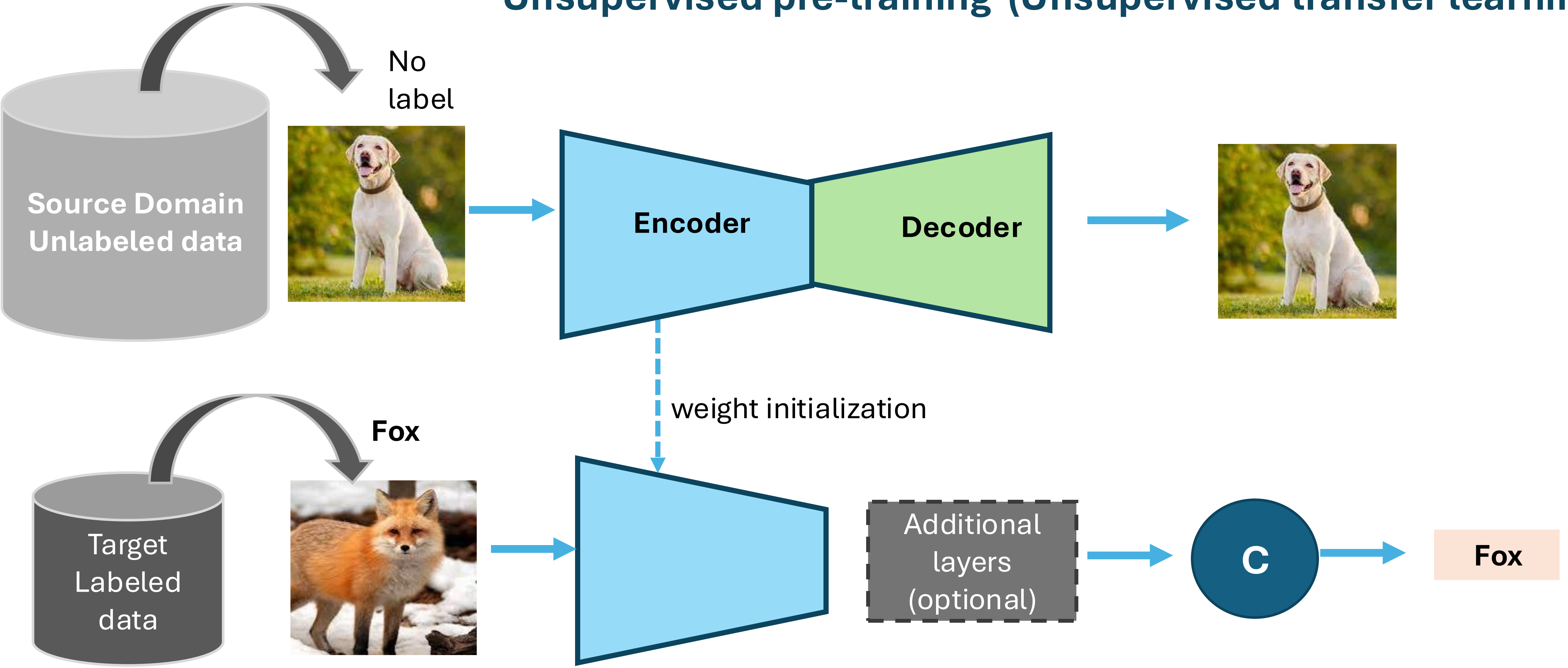
Reawakening of deep learning

- Breakthrough in 2006: **Deep Belief Network** [Hinton et al 2006]
- Some theoretical advantages of deep learning architectures have received an increasing attention
  - Some functions cannot be efficiently represented (in terms of number of tunable elements) by architectures that are too shallow.
- **Autoencoders** [Bengio, et. al., 2007]
  - Unsupervised learning algorithm applying backpropagation, setting target values equal to inputs
    - <u>Dimensionality reduction</u>
    - <u>Unsupervised pre-training</u>
      - Smarter weight initialization



1960 1980's 1990's **2000's** 2010's 2020's

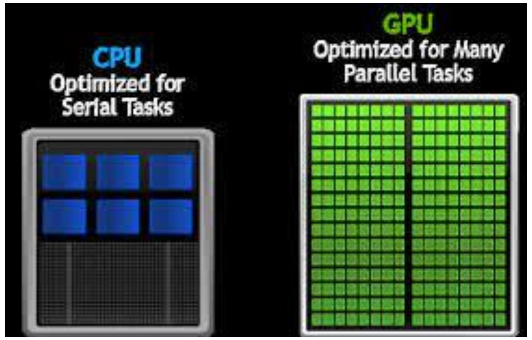
Unsupervised pre-training (Unsupervised transfer learning)



1960 | 1980’s | 1990’s | **2000’s** | 2010’s | 2020’s

GPUs revolution in DL

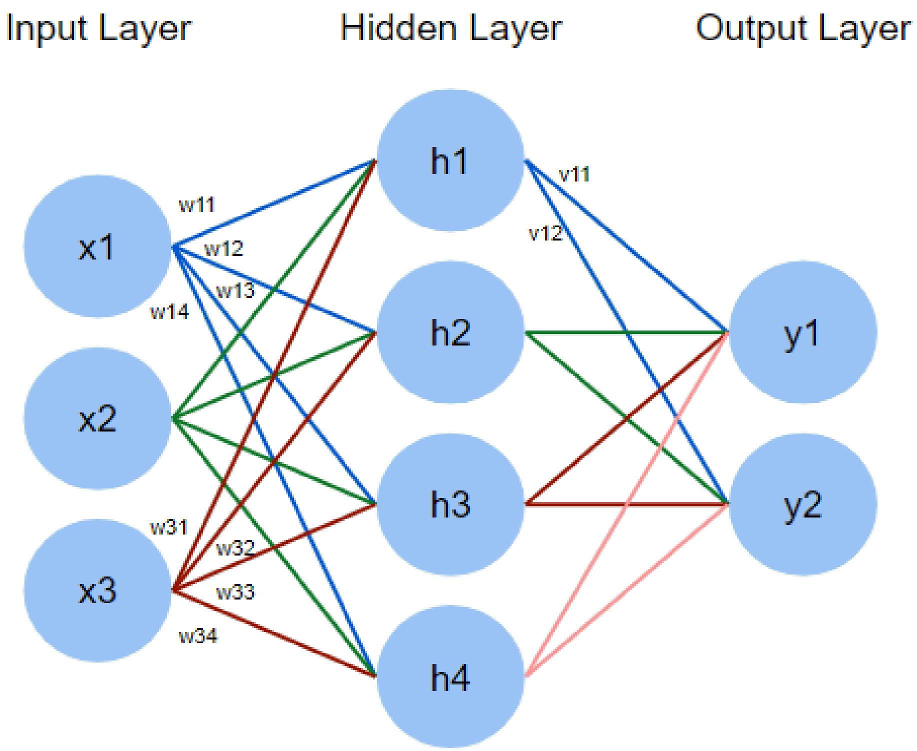
- **2008**: Andrew NG’s group starts advocating for the use of GPUs for training Deep Neural Networks to speed up the training time by many folds.



- GPUs the platform for Deep neuronal network
  - Why are GPUs good for Deep Learning?

	Neural Network	GPUs
Matrix operation	Yes	Yes
Inherently parallel	Yes	Yes
Bandwidth	Yes	Yes

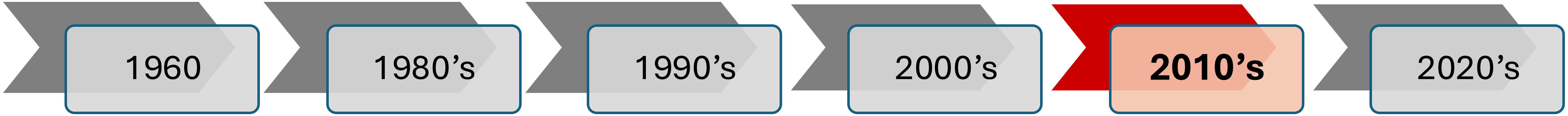
- GPUs deliver
  - Faster outcomes
  - Same or better prediction accuracy
  - Lower power
  - Lower cost



Launching ImageNet

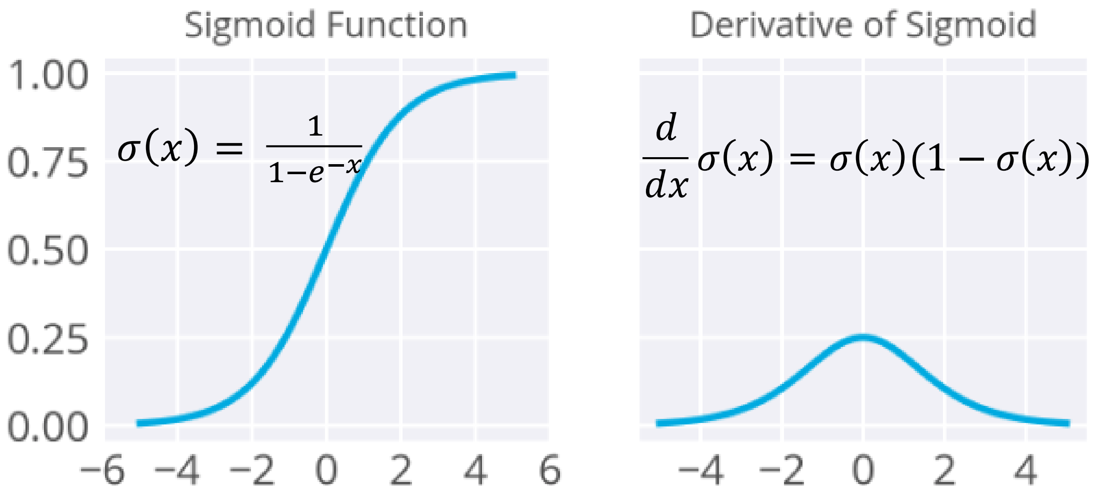
- **2009**: Fei-Fei Li launches ImageNet
  - A database of 14 million labeled images (hand-annotated).
  - It would serve as a benchmark for the deep learning researchers who would participate in ImageNet competitions (ILSVRC) every year.

Image source: [Synced 2020]



The mysterious success of ReLU

- ReLU (rectifier linear unit) activation function [Bengio, et. al., 2011]
  - ReLU activation function can avoid vanishing gradient problem
  - It doesn’t squash the output of a neutron between 0 & 1, which helps during backward propagation.
  - Computational advantages

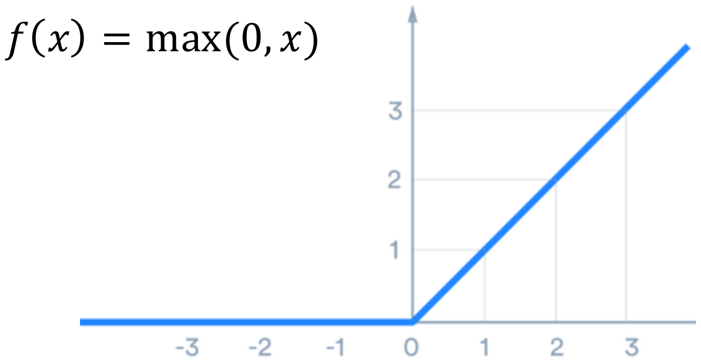

Sigmoid Function

Derivative of Sigmoid

$$\sigma(x) = \frac{1}{1-e^{-x}}$$

$$\frac{d}{dx}\sigma(x) = \sigma(x)(1-\sigma(x))$$

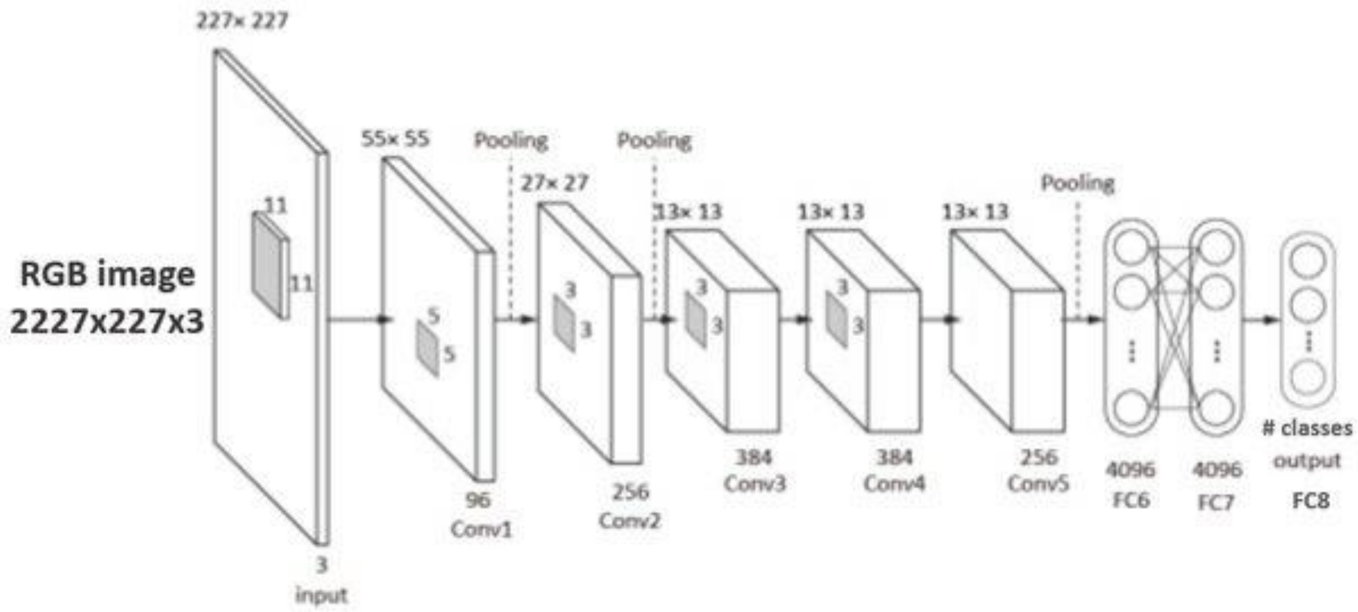
ReLU function



$$f(x) = \max(0, x)$$

Modern Convolutional Neural Networks

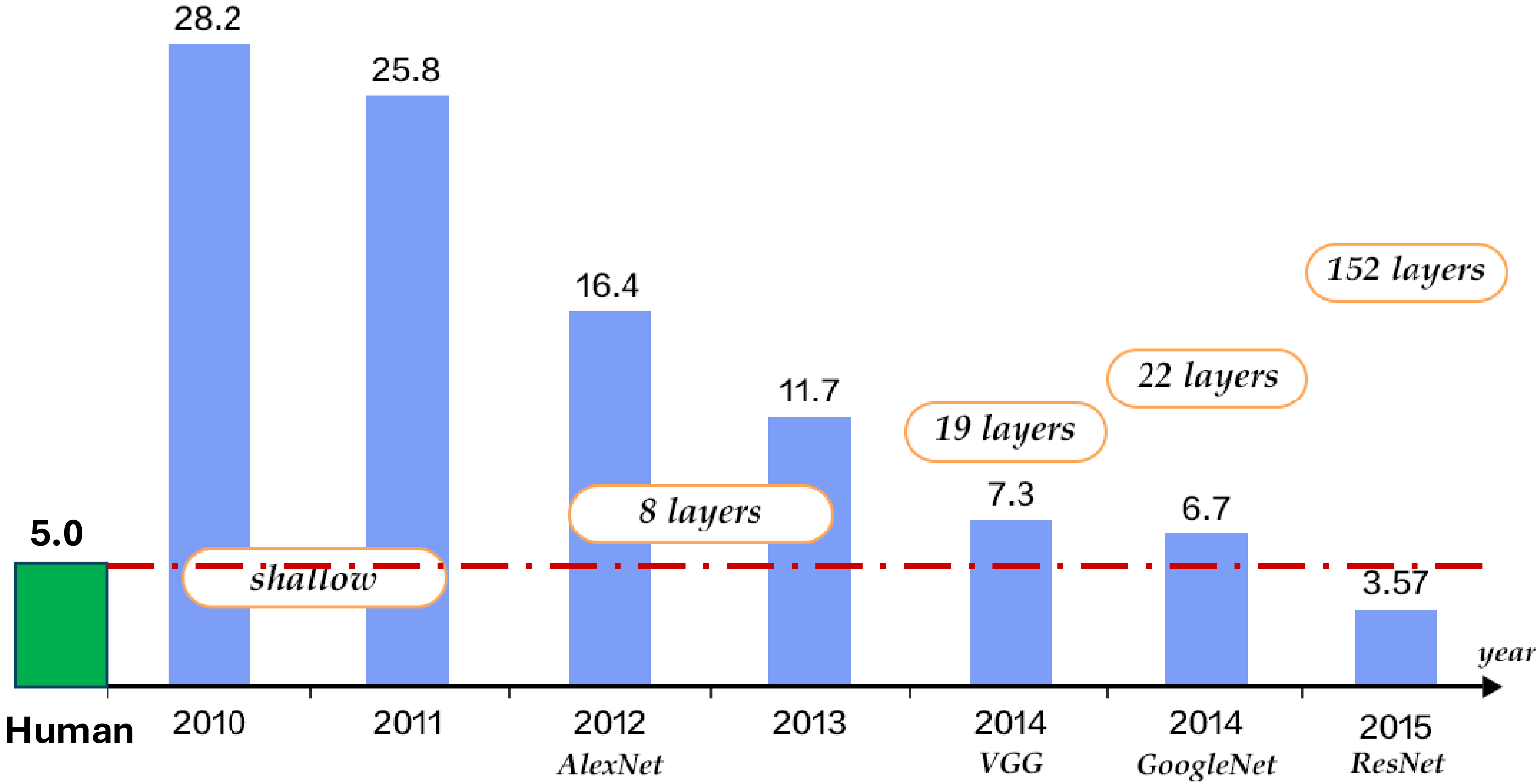
- AlexNet, [Krizhevsky, et. al., 2012]
  - A GPU implemented CNN model wins Imagenet’s image classification with accuracy of 84%.
  - A huge jump over 75% accuracy



- Further advanced architecture
  - Alexnet (2012): 8 layers
    - 5 convolutional layers
    - 3 fully connected
  - VGGnet (2014): 19 layers
  - GoogLeNet (2014): 22 layers
  - ResNet (2015): 152 layers

Image source: [Khvostikov et al 2018]

The evolution of the winners on the ImageNet Large Scale Visual Recognition Challenge



Deep Revolution and Transformation: Self- feature engineering

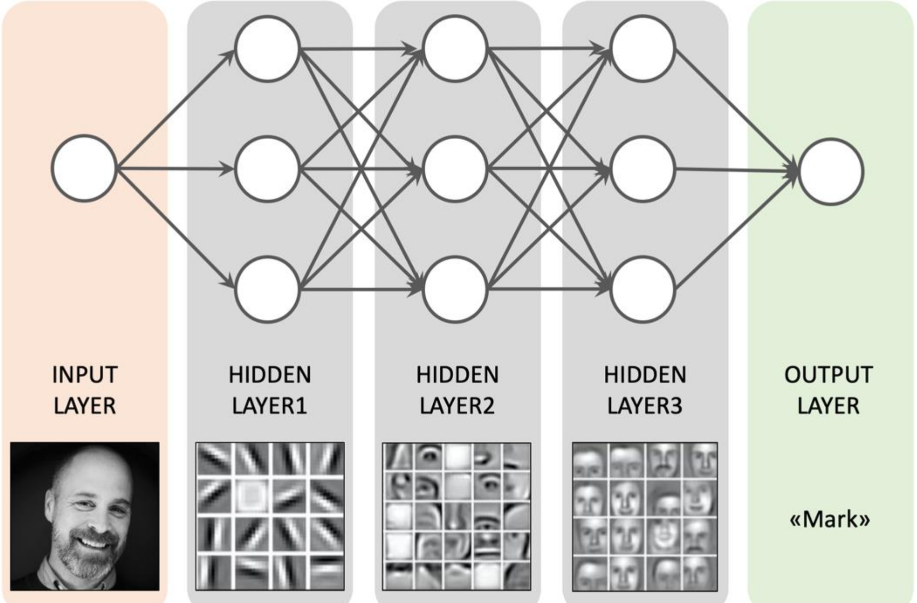
- **Shallow network**: network consisting of an input, hidden and output layer, where the features are computed using only one layer.
- **Deep network**: multiple hidden layers to determine more complex features of the input.

- **Self- feature engineering** : One can learn part-whole decompositions (feature hierarchies): e.g.,

  layer 1: learn to group pixels in an image to detect edges

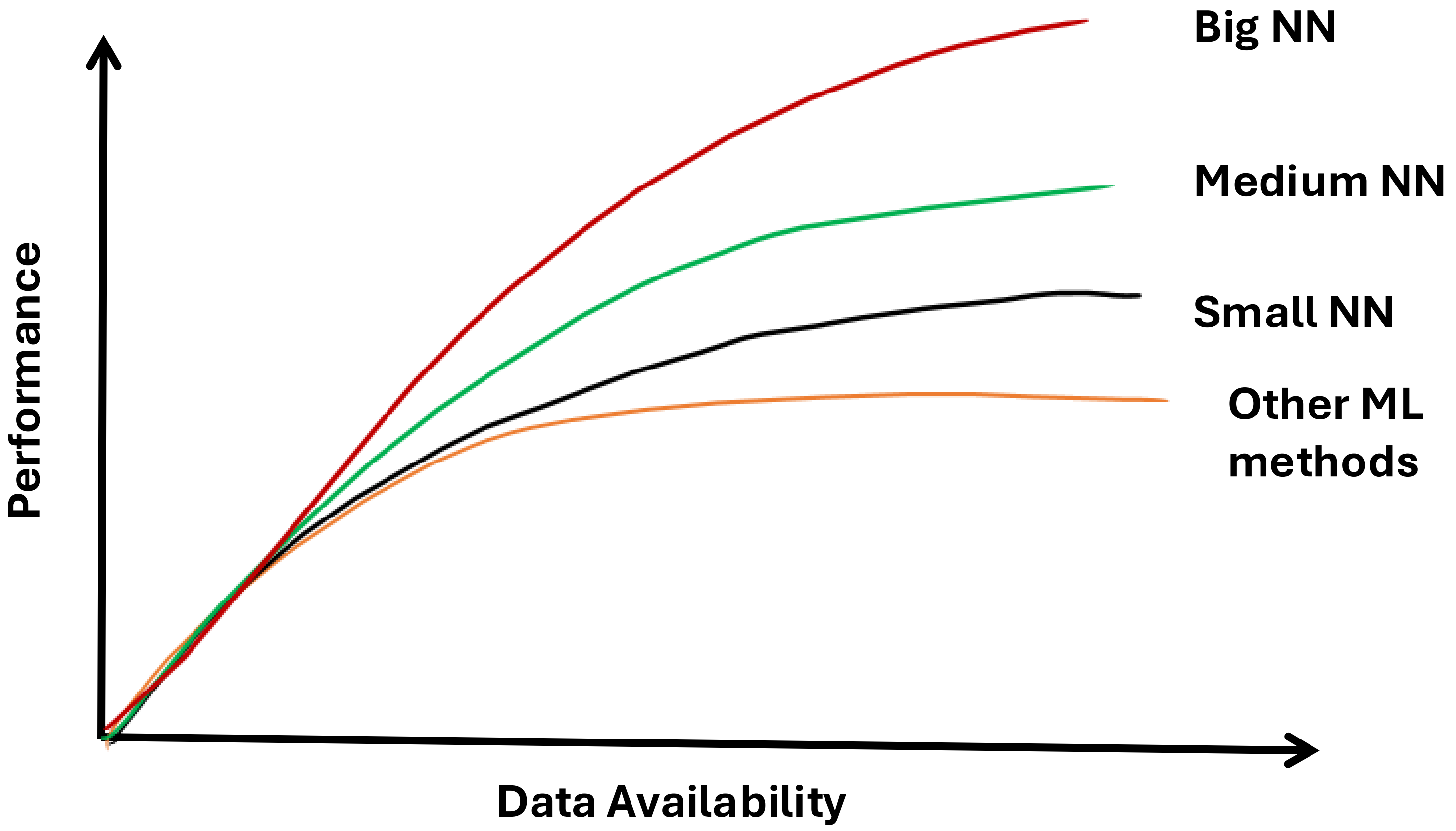
  layer 2: group together edges to detect longer contours (detect simple parts of objects)

  layer 3,4,….: group together contours or detect complex features



- Application in Transfer Learning
  - Improvement of learning in a **new** task through the *transfer of knowledge (transfer of features)* from a **related** task that has already been learned.
  - Weight initialization for CNN
  - ConvNet as fixed feature extractor
  - Fine-tuning the ConvNet

Data Revolution and ML methods



What challenges and limitations remain in Deep Learning?

Deep Learning challenges

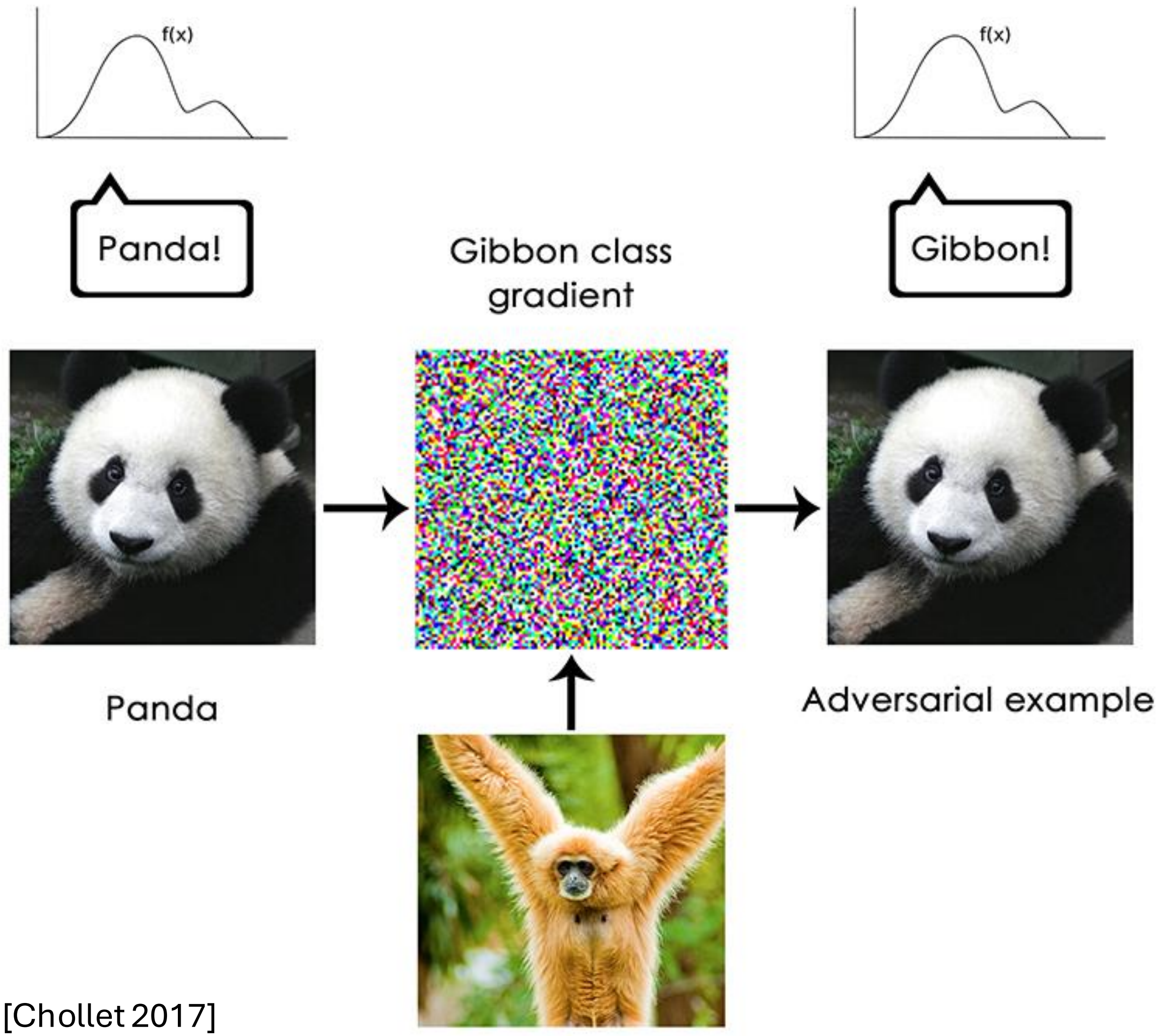
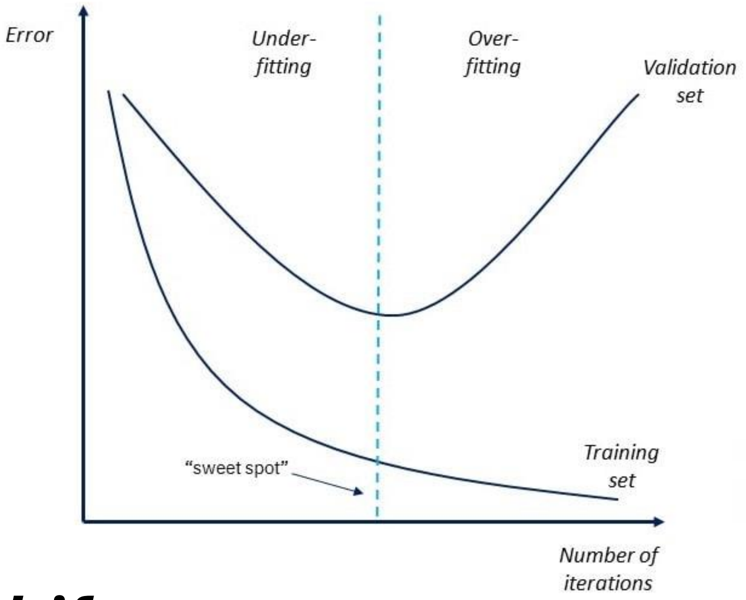


Image source: [Chollet 2017]

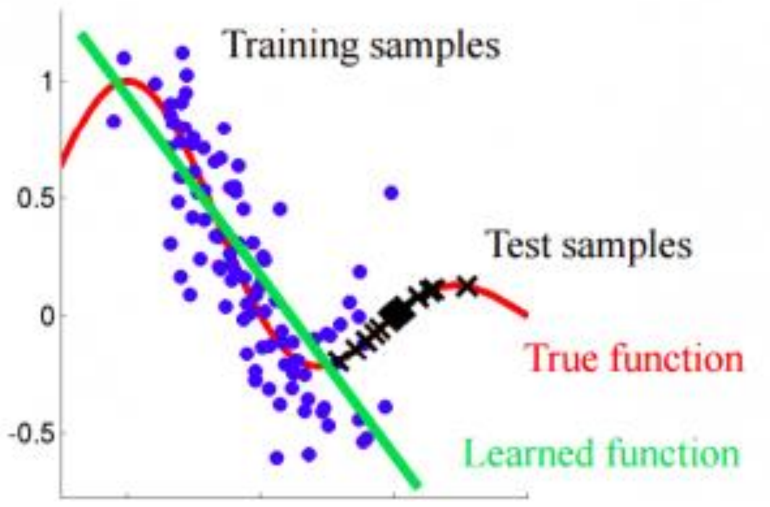
Deep Learning Challenge: Overfitting

When data becomes problematic

- Overfitting
  - Lack of training data makes deep learning models prone to overfitting



- Vulnerable to ***covariant shift***
  - Low performance in external validation

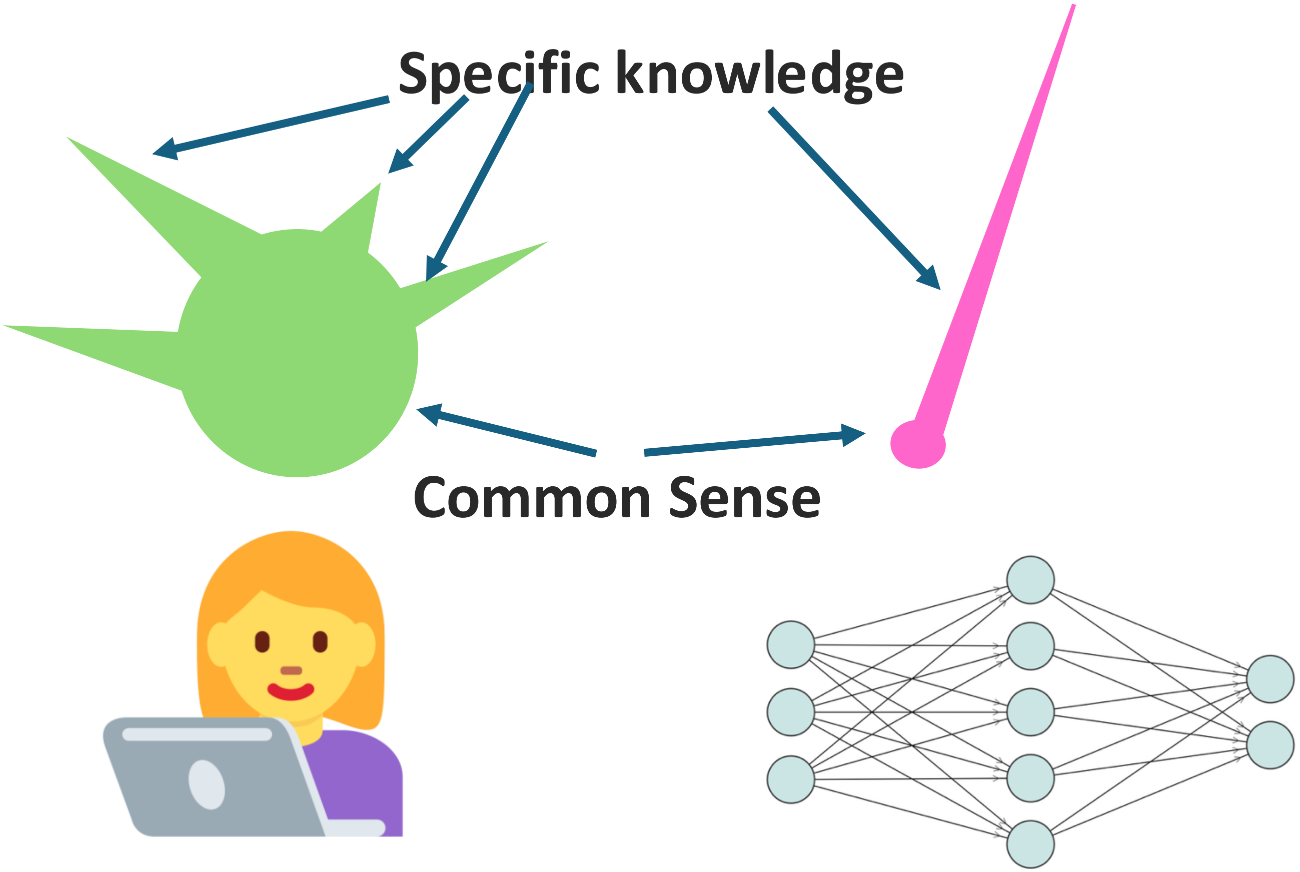


Deep Learning Challenge: Generalization

- Deep learning doesn’t have the common sense to understand

Generalization challenge in Deep learning

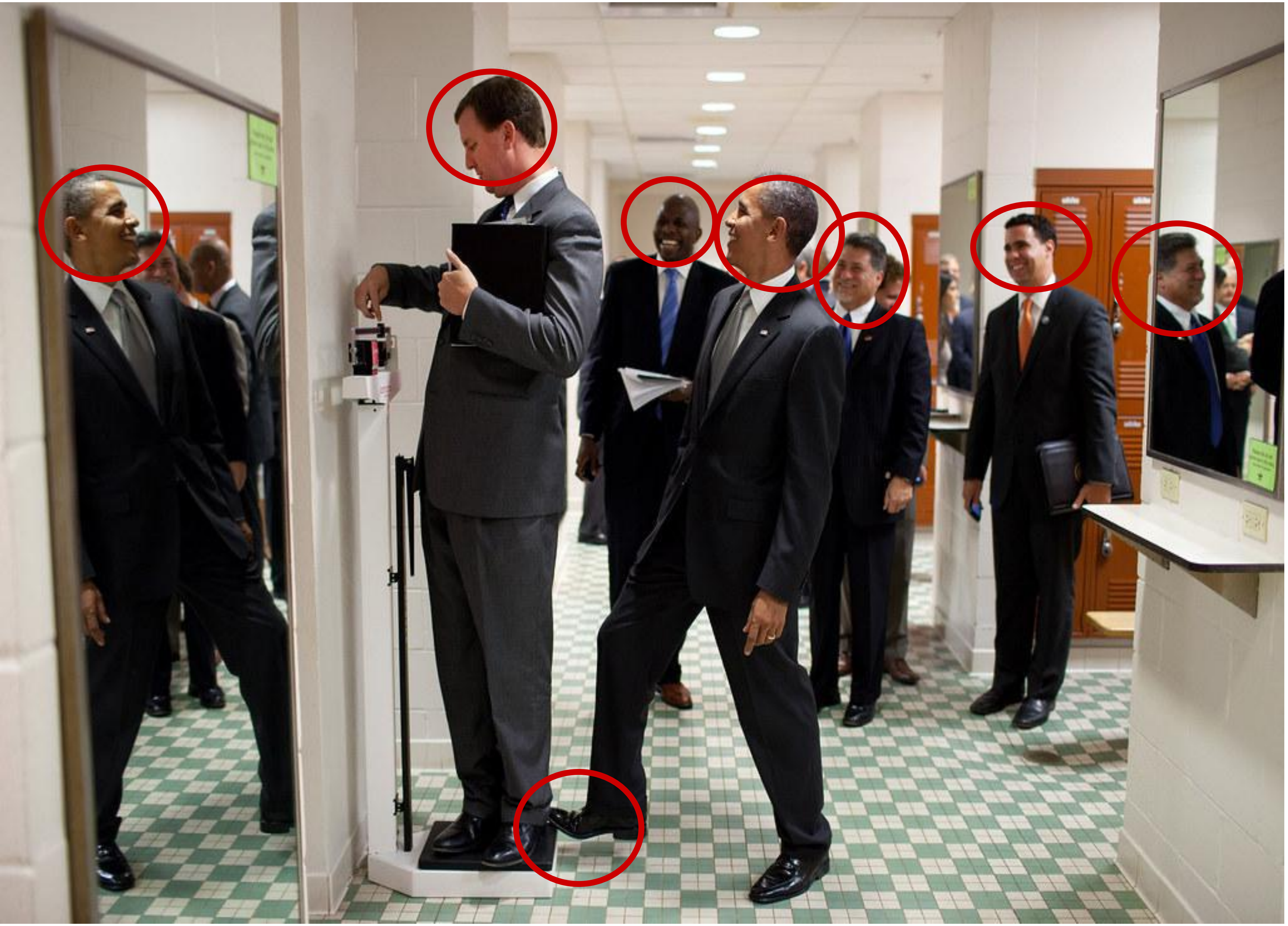
- Common Sense — Still not Common in AI
  - AI only access specific knowledge



Specific knowledge

Common Sense

We are just at the beginning

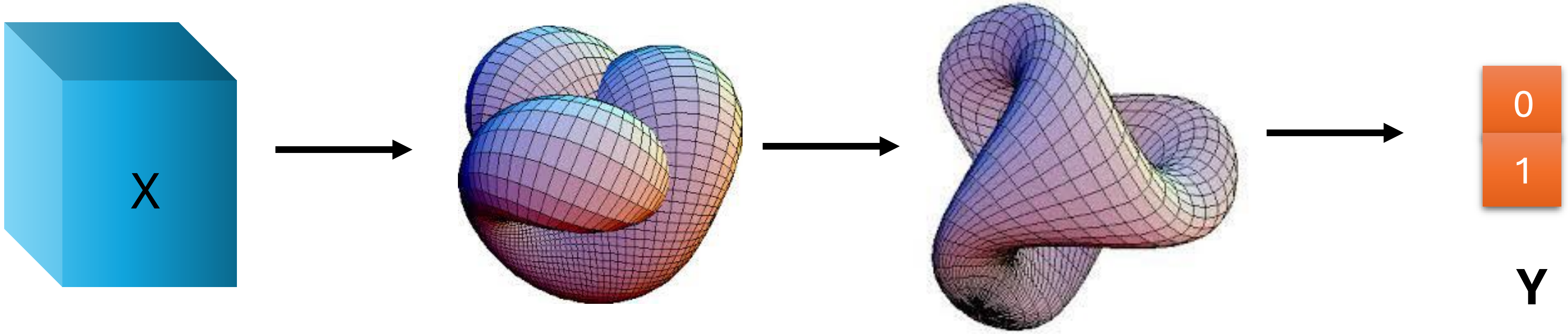


- Our perceptions vs machine perception

- Remember Moravec's Paradox: What is easy for humans is hard for machines

What does Deep Learning actually do?

- Deep Learning model is just a chain of simple continuous geometric transformations mapping one vector space into another
- All it can do is to map one data manifold X into another manifold Y
  - Assuming the existence of learnable continuous transform X to Y



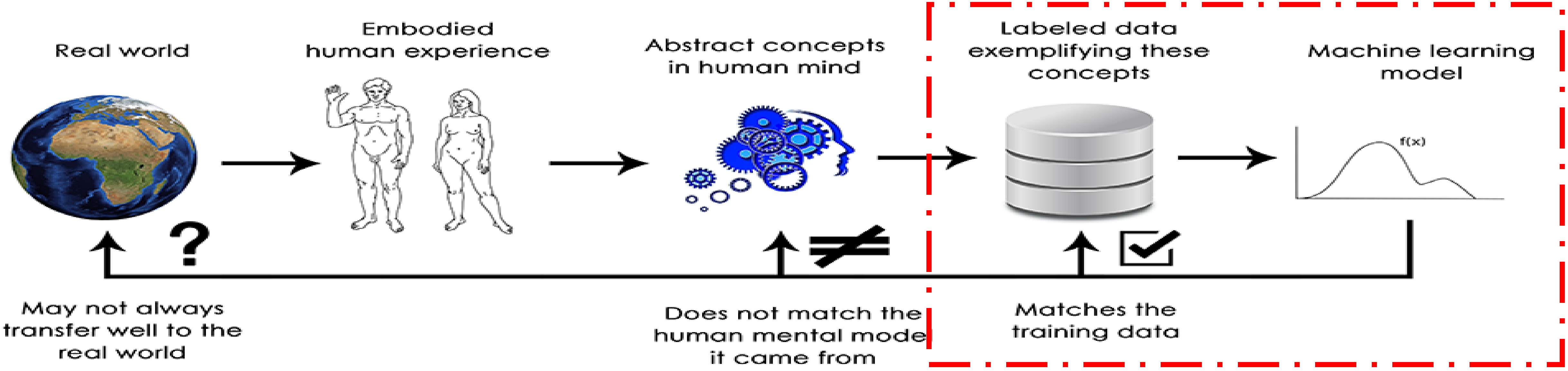

Real world
Embodied human experience
Abstract concepts in human mind
Labeled data exemplifying these concepts
Machine learning model
f(x)
?
≠
☑
May not always transfer well to the real world
Does not match the human mental model it came from
Matches the training data

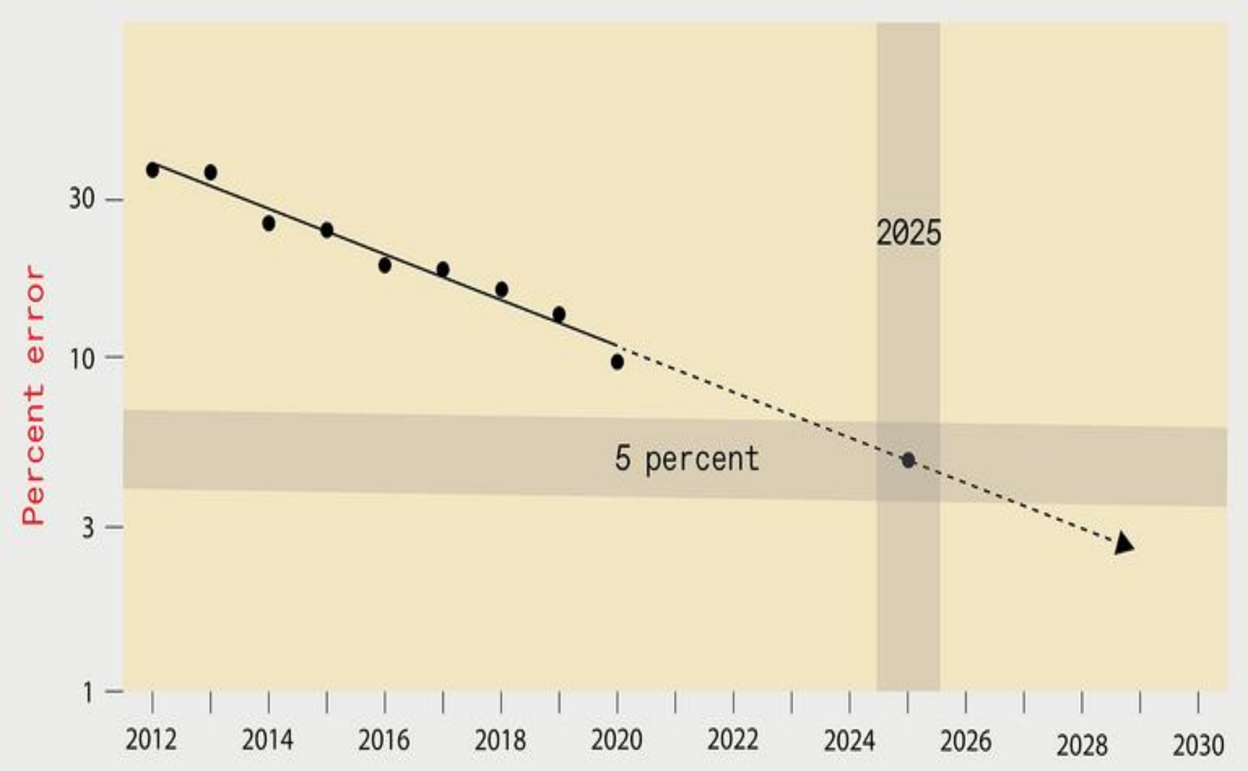
Image source: [Chollet 2017

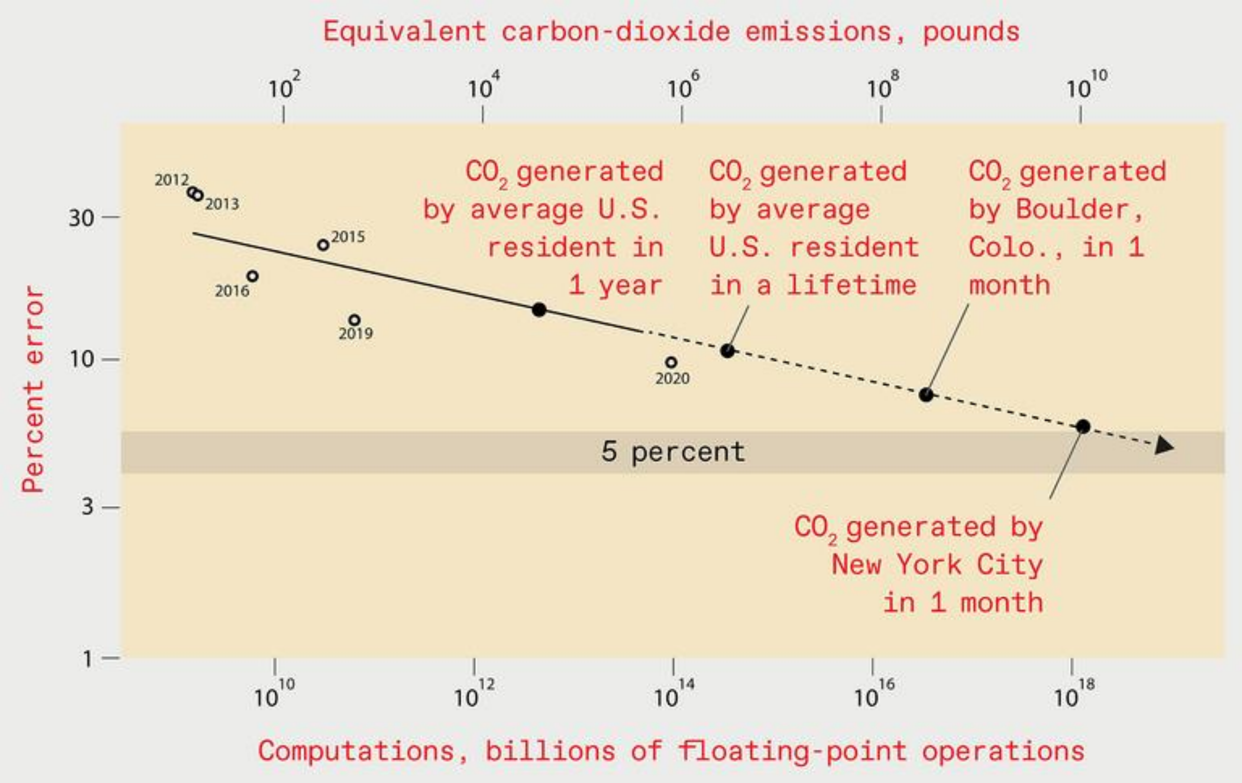
- ML models don’t have any understating of their input
  - At least, not in a human sense
- Understanding of images/sounds/words is grounded in our sensorimotor experience
  - ML learning models have no access to such experiences
  - They can't understand their input in a human-relatable manner
- We can get them to learn a geometric transform that maps data to human concepts on “a specific set of examples”
  - But this mapping is a simplistic sketch of the original model in our minds—the one developed from our experience as embodied agent

Deep Learning Challenge: Scalability

- **How far can Deep Learning scale up?**

DEEP LEARNING'S DIMINISHING RETURNS?





Reducing image-classification errors (top 1) has come with an enormous expansion in computational burden.

Image source: [Thompson, Greenwald, Lee, and Manso 2021]

Deep Learning Challenge: Opacity

- **Opening the black box**
  - ***Inside: How to develop a theoretical understanding of DL and NNs?***
  - **Outside: Explainability of the model outputs**

Explainable AI

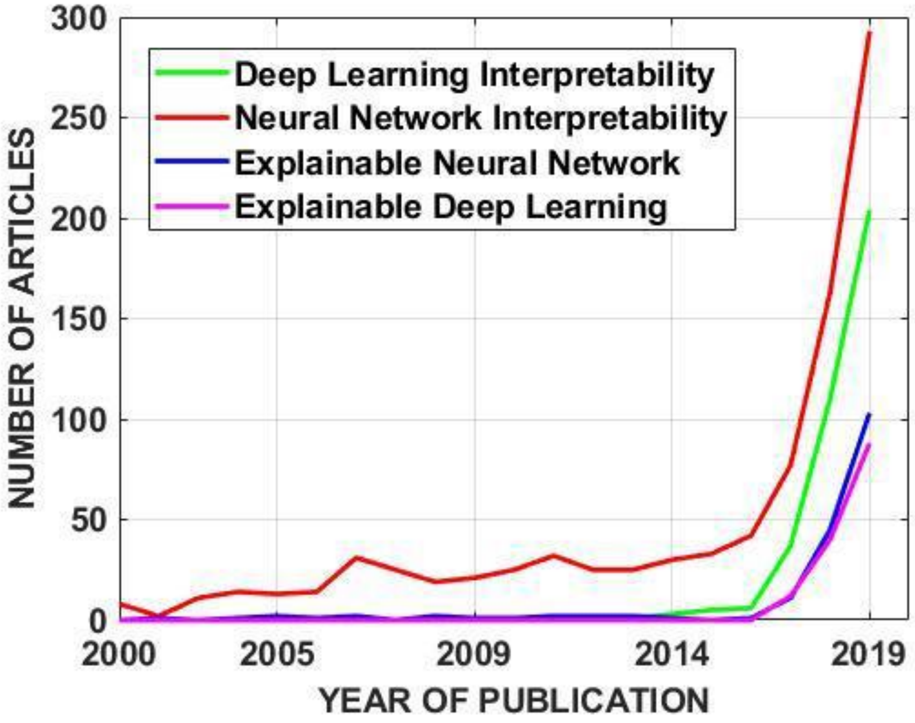
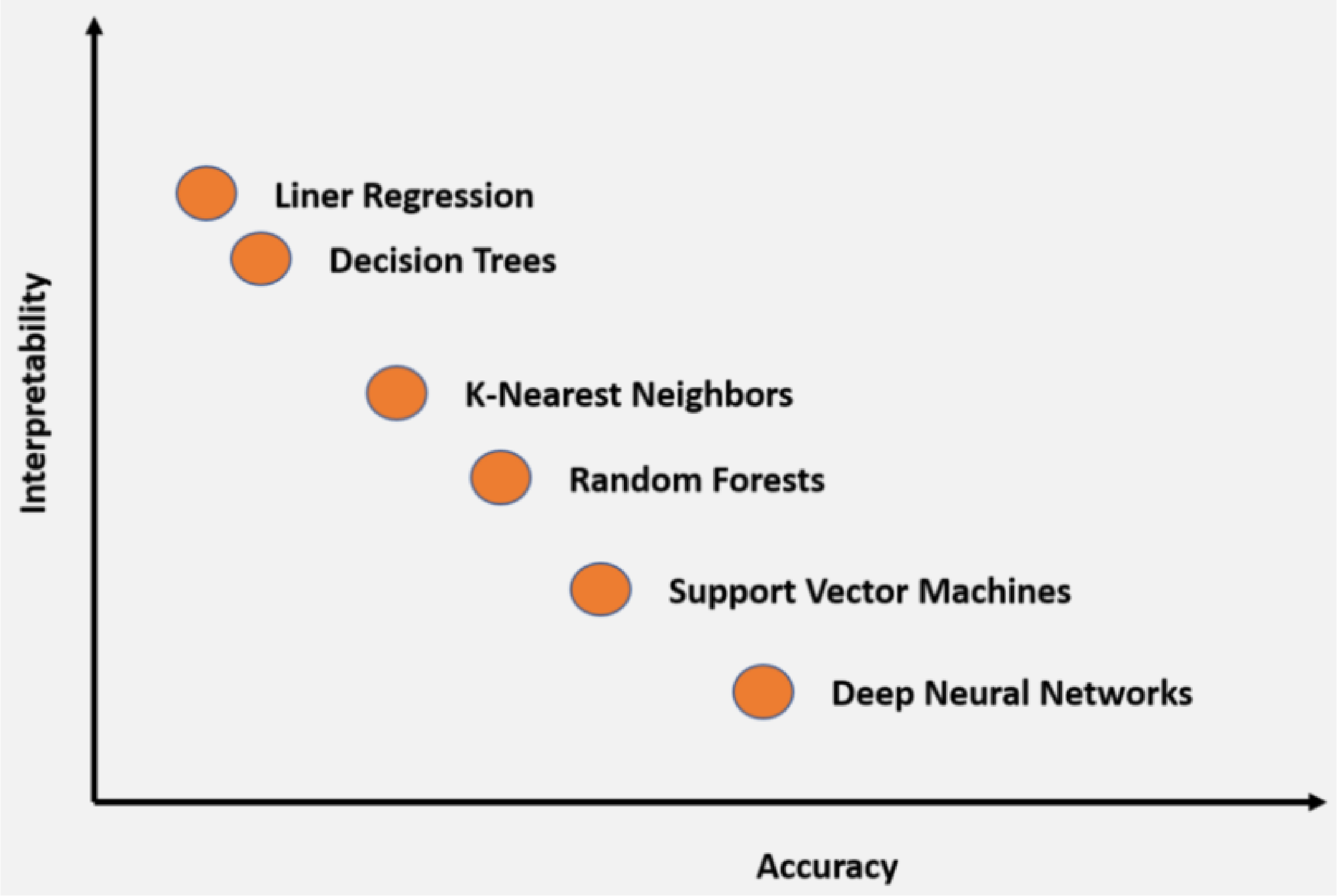


Image Source: [Fan et al 2020]



The false hope of current approaches to explainable artificial intelligence [Ghassemi et. al., 2021]

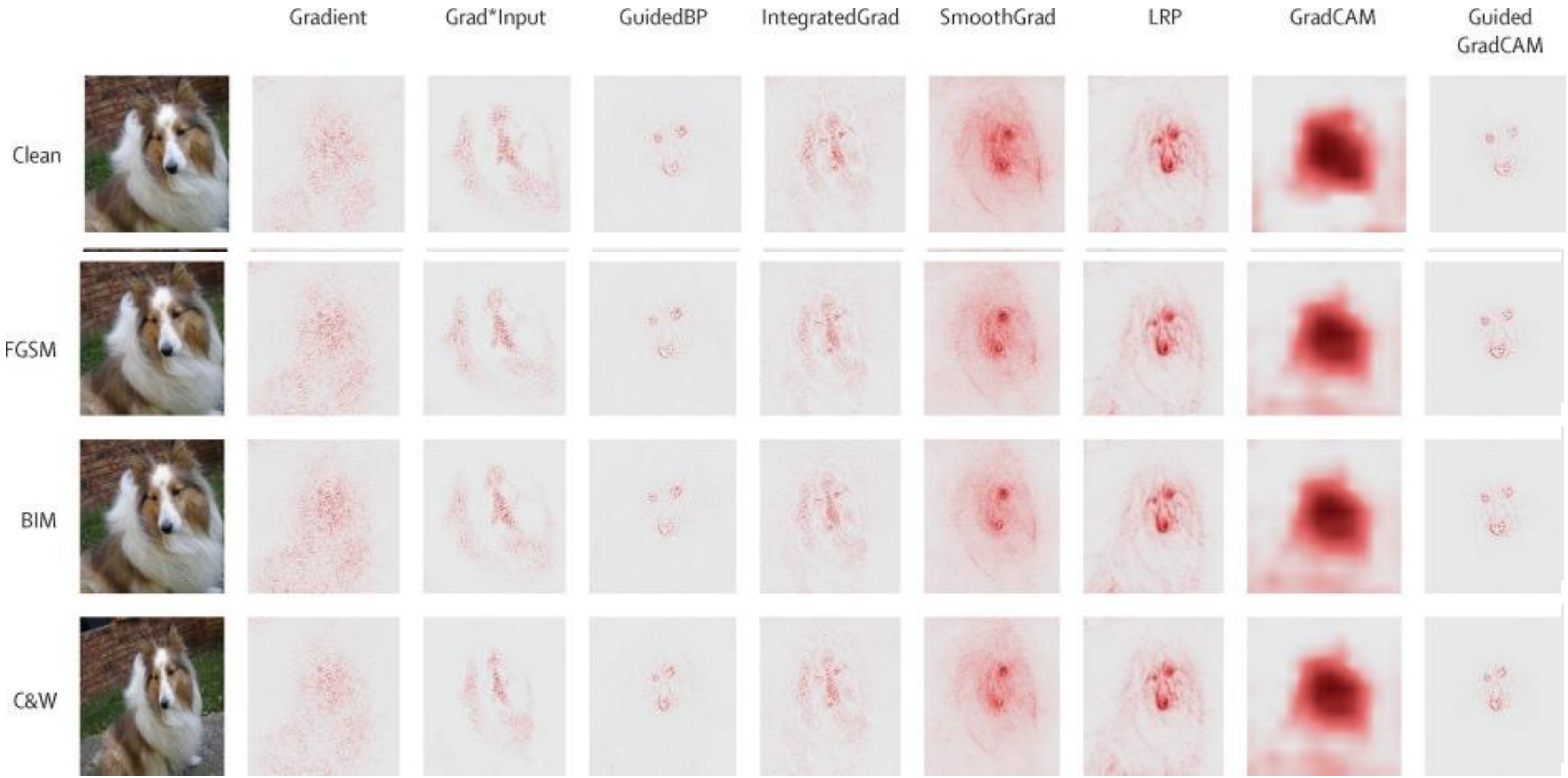
Saliency maps

correctly classified

Incorrectly classified

Incorrectly classified

Incorrectly classified



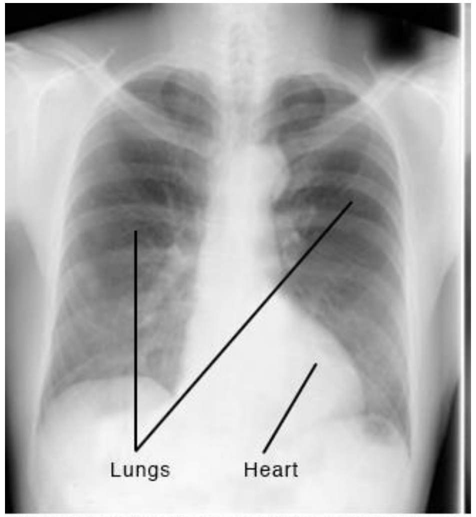
The figure in each row, are subject to different adversarial perturbations

Saliency map: taking the image that the algorithm was fed and creating a heat map of those portions of it that were most heavily weighted by the deep learning in making a prediction.

Explainable AI or a False hope?

Heat map produced by a post-hoc explanation method for a deep learning model designed to detect pneumonia in chest x-rays

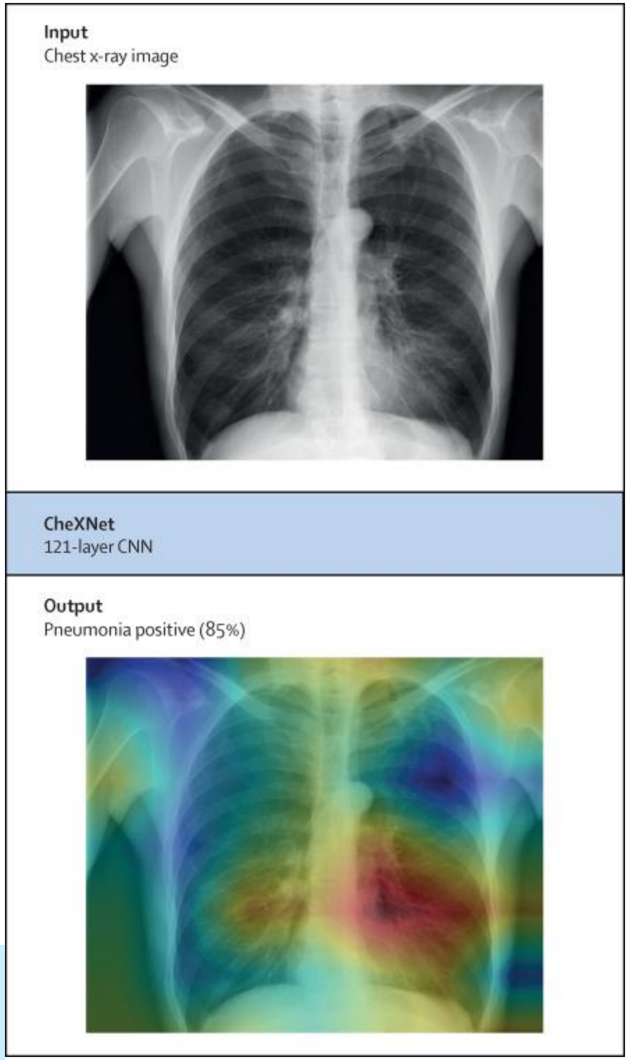
“The clinician cannot know if the model appropriately established that the presence of an airspace opacity was important in the decision, if the shapes of the heart border or left pulmonary artery were the deciding factor, or if the model had relied on an inhuman feature, such as a particular pixel value or texture that might have more to do with the image acquisition process than the underlying disease”. [Ghassemi et. al., 2021]



© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGH

A chest X-ray of normal person

The false hope of current approaches to explainable artificial intelligence in health care



Input
Chest x-ray image

CheXNet
121-layer CNN

Output
Pneumonia positive (85%)

[Ghassemi et. al., 2021]

What Should AI Be/Do?

- **Normative Questions**: Ethical guidelines for AI development and goals
- **Impact**: AI’s potential role in society, ethical boundaries

A third core question is: *What should AI be?* This asks us to consider what roles AI should play in society and what ethical constraints we should place on its development. Should AI have rights or responsibilities? Should there be limits to the tasks we delegate to machines?

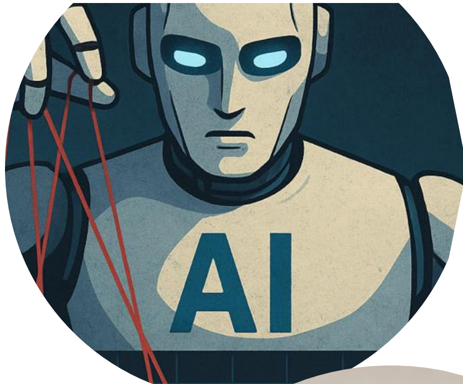
Ethics of AI and ML: Exploring AI's Impact on Society and Humanity

AI and ML as transformative technologies

- How should we use these systems?
- What risks do they involve?

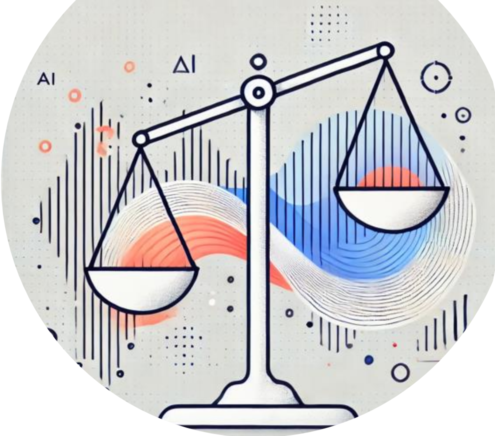
What should AI be

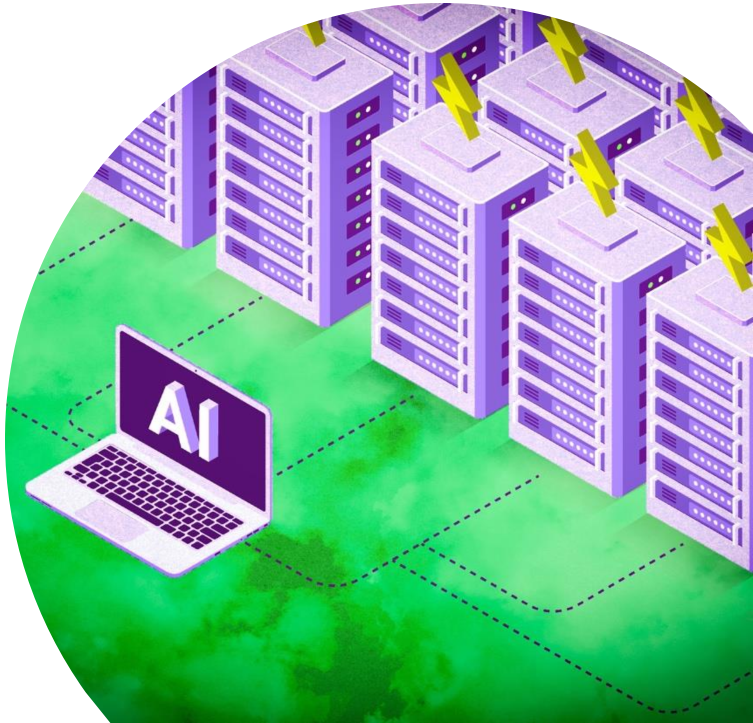
- **Privacy and Surveillance**
  - AI enhances surveillance capabilities: facial recognition, data tracking
  - Risks: Erosion of personal privacy, increased data control by corporations and governments
- **Manipulation of Behavior**
  - Targeted manipulation through data analysis
  - Nudges and “dark patterns” in online content, social media, and advertising
- **Opacity in AI Systems**
  - AI systems are often “black boxes” with opaque decision-making
  - Lack of transparency impacts accountability and trust



What should AI be

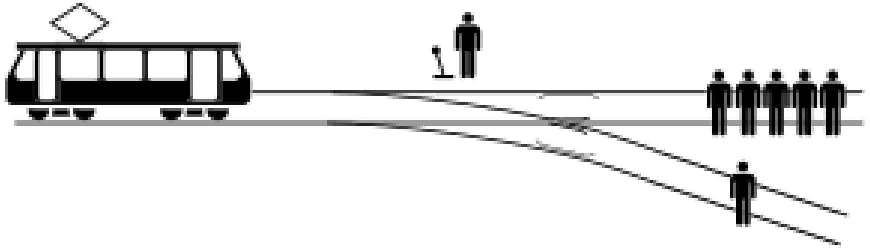
- **Bias in Decision Systems**
  - AI bias from historical data
  - Effects on justice, hiring, credit allocation
- **Human-Robot Interaction**
  - Ethical considerations in human-robot relationships: empathy, care, companionship
  - Risks of dehumanization
- **Economic Impact and Employment**
  - Job displacement from automation
  - Concerns about economic inequality, future job market
- **Environmental Impact**
  - AI and resource consumption
  - Energy-intensive machine learning processes, waste from tech infrastructure





What should AI be

- **Autonomous Systems Overview**
  - Degrees of autonomy in AI
  - Key questions: safety, accountability, ethical design



- **Autonomous Vehicles**
  - The promise of safer, efficient transport
  - Ethical dilemmas: trolley problems, accountability

***The trolley problem** is an ethical thought experiment about a fictional scenario involving hypothetical ethical dilemmas about whether to sacrifice one person to save a larger number.*

- **Autonomous Weapons**
  - Risks of lethal autonomous weapons
  - Debates on accountability, international humanitarian law

What should AI be

- **Machine Ethics**
    - Ethics in machine behavior
    - Debate over AI’s capability for ethical decision-making
- **Artificial Moral Agents**
    - Defining moral agency in AI
    - Rights and responsibilities for AI systems
- **Robot Rights and Legal Personhood**
    - Debate over granting legal rights to robots
    - Comparison to corporate personhood, implications for accountability
- **The Singularity Concept** (John von Neumann, Vernor Vinge and Ray Kurzweil)
    - *Strong AI, Superintelligence, Technological Singularity, AGI*
    - The singularity: hypothetical superintelligent AI surpassing human control
    - Potential for rapid, unpredictable advancements

What should AI be

- **Superintelligence and Existential Risks**
  - Superintelligent AI as a potential existential threat
  - Risks of AI developing incompatible goals
- **Control and Value Alignment**
  - The “control problem” of superintelligent AI
  - Importance of aligning AI’s goals with human ethics

AI challenges human self-perception as the dominant intelligent species
Need for continued ethical and regulatory considerations

Challenges Raise Against Common Approaches in Ethical AI

- Not Focus on AI's impact on political structures, justice, and human agency
  - ***Ethical AI as a political, not just technical, issue*** *(Coeckelbergh)*
- AI centralizes power within corporations, governments, and tech elites
- Risks to **democracy, transparency, and individual rights**

Human-Centered Approach in Ethical AI

- ***What is the Human Being?* - A Foundational Question for AI**
    - Returning to Kant’s four fundamental questions is a profound way to reframe our approach to AI ethics, moving the focus back to the human-centered inquiry that AI inevitably implicates. This framework reminds us that, rather than asking solely about AI’s capabilities, nature, and ethical role, we should use AI as a lens to reflect on the essential questions of human knowledge, ethics, identity, and purpose.
        - ***What can I know?***
        - ***What may I hope?***
        - ***What should I do?***
        - ***What is the human being?***

In an era where AI both mirrors and intensifies our search for identity, meaning, and purpose, can we return to the question, "What is the human being?" to reclaim a vision of humanity that transcends efficiency and control, affirming the values, creativity, and ethical consciousness that technology cannot replicate or replace?

AI Technology

Is AI technology a neutral tool?

- Both mainstream AI ethics and even human-centered approach reframing often treat technology as a neutral tool, *an instrumental means to human ends*.
- Technology as a Non-Neutral Force or as a force that shapes and transforms human experience, values, and society at a foundational level.

View of the Rhine at Reineck, by Herman Saftleven, 1654, oil on canvas

The hydroelectric dam on the Rhine, photo by Maarten Sepp

Technology: Neutral Tool or Social Force?

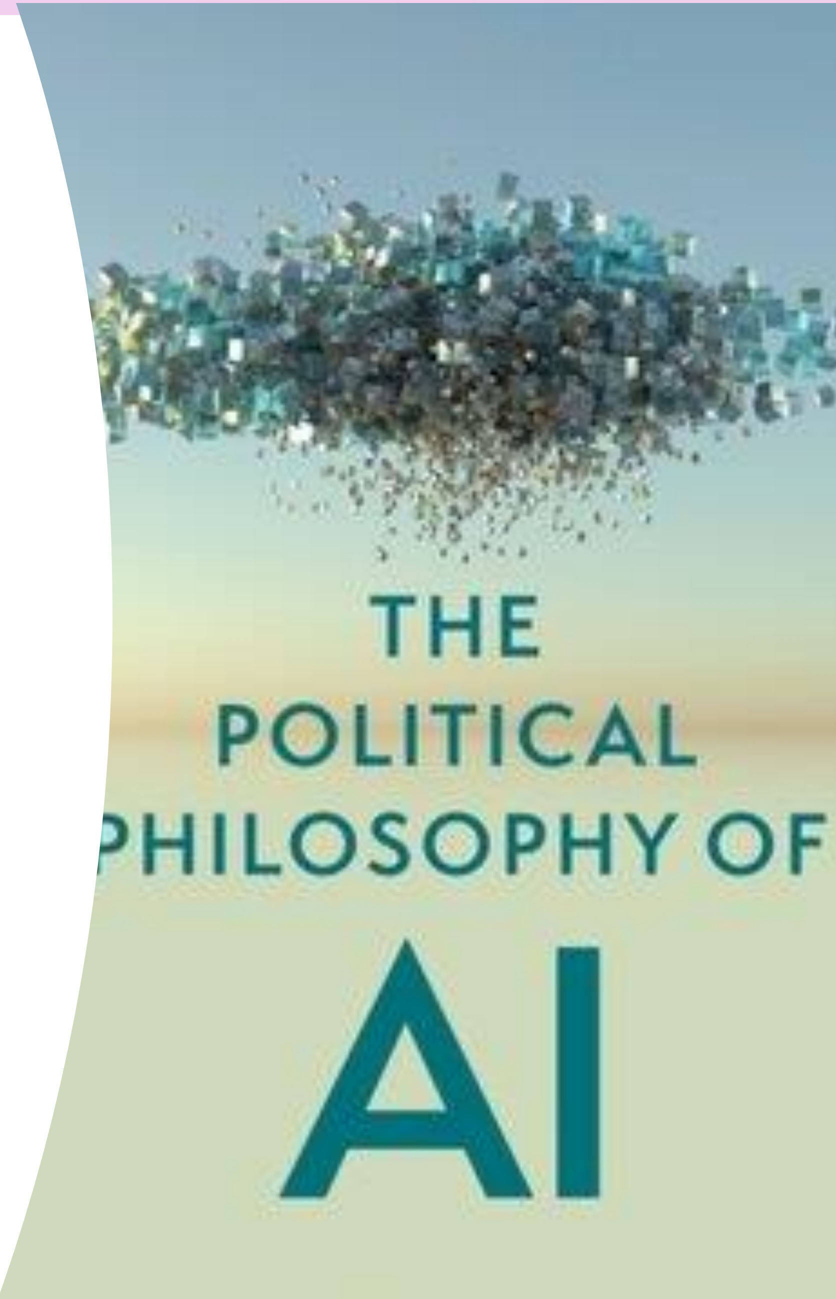
- **Technology is not neutral**
  - AI is shaped by human values, institutions, and power structures.
  - Design choices reflect assumptions about efficiency, control, prediction, and optimization.
- **AI also shapes society**
  - Influences behavior, norms, autonomy, and political life.
  - Reinforces existing inequalities depending on who builds and deploys it.
- **Co-Shaping View**
  - Technology is both **socially constructed** *and* **socially shaping**.
  - AI and society continuously influence each other.

AI as a Political and Social Tool

- **AI as a Political and Social Force**
  - AI systems don’t just *assist* society — they *shape* it by influencing power, institutions, and public life.
  - Technologies embed values, priorities, and political choices.

- **Human Agency & Democratic Technology**
  - Key question: Who gets to shape AI — citizens, corporations, or states?
  - Democratic control requires transparency, accountability, and public participation.

- **Bias, Opacity & Instrumental Rationality**
  - AI can reinforce existing inequalities through biased data and opaque algorithms.
  - Efficiency-driven design can sideline human judgment, ethics, and shared social values.



Final Remarks and Q&A

- **A Call for Questioning More**
- Questioning science?
  Questioning technology?
  Questioning Being?
- **A Call for Democratic Technology**
- **A Call for Multi-Disciplinary Collaboration**
  - Recognizing that AI’s impact spans multiple facets of life, We need more collaboration between technologists, philosophers, politicians, and the public to navigate AI’s complexities thoughtfully and inclusively.

References and Further reading

- Philosophy of artificial intelligence (Wikipedia)
- Philosophy (Wikipedia)
- Müller, Vincent C. (2020), 'Ethics of artificial intelligence and robotics', in Edward N. Zalta (ed.), Stanford Encyclopedia of Philosophy (Palo Alto: CSLI, Stanford University).
- Roger Vergauwen, Lecture on the philosophy of mind and AI (2014)
- Andrew Feenberg: *The Social Life of Reason* (Harvard University Press, 2017).
- Coeckelbergh, Mark (2022). The Political Philosophy of AI, Polity Press, Cambridge, UK