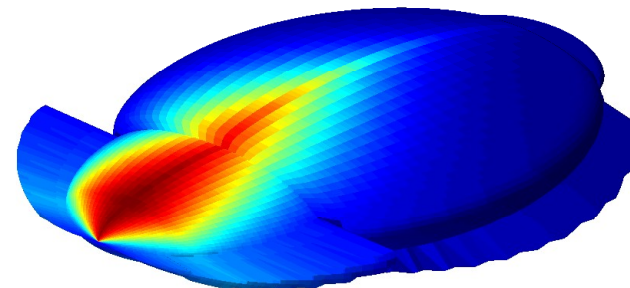
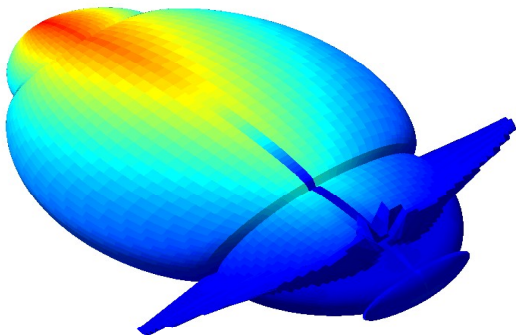


Lecture 1

Intelligence and AI



Defining Intelligence

The ability to acquire and apply knowledge and skills.

The Oxford English Dictionary

The ability to derive information, learn from experience, adapt to the environment, understand, and correctly utilize thought and reason.

American Psychological Association

Critique: These definitions are too human-centered

A Biologically-Motivated Definition:

The ability of an **autonomous agent** to **interact flexibly and adaptively with its environment to its own advantage.**

Key Point:

Intelligence is about **behavior** and **interaction** with a **specific environment**, not a generic toolbox for information processing and control.



Natural Intelligence (NI)



- **Biological:** Organic substrate with biological structures and processes.
- **Embodied:** Able to sense and act in the physical world through a body.
- **Emergent:** Arises from self-organization in complex adaptive systems.
- **“Scale-rich”:** Molecules – cells – tissues – organs – systems.
- **Adaptive at multiple scales:** Evolution, development, learning, behavior.
- **Autonomous:** Motivated by internal goals and drives.
- **Always “real-time”:** Limited off-line processing and rehearsal.
- **Environment-specific:** Evolved to function in a specific context.
- **Inherently integrated:** Always integrated across all modalities.
- **Evolvable:** Produces more intelligent organisms with deeper organization.



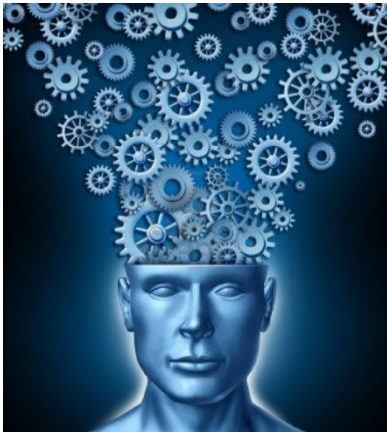
The computational AI we build today is very different.



Can machines be Intelligent?

The Mind-Body Problem:

Are mind and body distinct, or is mind an attribute of the physical body?



[Source](#)

Dualism: Mind and body are different substances

→ True AGI is impossible to build; only simulation is possible.

Materialism: Mind is an attribute of the physical body

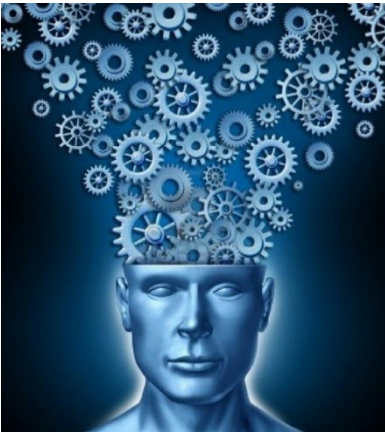
→ Existence proof that AGI is possible (humans).

AI (and science) adopt the materialist view.

Can machines be Intelligent?

The Mind-Body Problem:

Are mind and body distinct, or is mind an attribute of the physical body?



[Source](#)

Dualism: Mind and body are different substances

→ True AGI is impossible to build; only simulation is possible.

Materialism: Mind is an attribute of the physical body

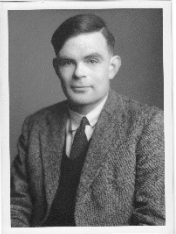
→ Existence proof that AGI is possible (humans).

AI (and science) adopt the materialist view.

Follow-up Questions:

- How can we tell if a machine is intelligent?
- What does intelligence require?
- What kind of intelligence can machines have?

How can we tell if a machine is intelligent?



Turing's Question: Can a machine think? (**Assumption:** intelligence = thinking)

Turing's Dilemma: How can thinking be measured?

Turing's Solution: Thinking can be probed and ascertained via discourse.

Turing Test: If a machine can imitate human discourse perfectly, it is intelligent.

Objection:

- **Discourse \neq thinking** - syntax is not sufficient to capture semantics (Searle).
- **Language \neq understanding** - words and expressions are not inherently meaningful.
- **Imitation \neq Intelligence:** Intelligence requires a causal and grounded understanding of the world → **World Model**

New Dilemma: How can understanding be measured?

Measuring “Understanding”

How can we measure “understanding” objectively in another system?

1. By behavioral testing of response in complex situations ↔ **Enhanced Turing Test**
2. With real-time brain imaging methods:
 - Purely phenomenological (e.g., EEG signatures).
 - No substrate-independent signature of understanding.
 - Limits the capacity of understanding to animals.

Choosing Option 2 and excluding Option 1

- ⇒ machines are inherently incapable of understanding
- ⇒ machines cannot be intelligent
- ⇒ **implicit/weak dualism** or **undiscovered principles**

- How do we know that other humans understand?
- Do pets understand us when they obey our commands?

Answer: We ascribe a mind to humans and (some) animals – **Theory of Mind**.

Questions – Good and Bad

Bad Questions:

- Do (or can) AI systems understand?
- Do (or can) AI systems have self-awareness?
- Are AI systems sentient/conscious (or can be)?

Good Questions:

- Does this AI system behave as if it understands?
- Does this AI system behave as if it has self-awareness?
- Does this AI system behave as if it has general intelligence?
- Does experimental testing show that this AI system has a theory of mind?
- Is this AI system's intelligence similar to human intelligence?
- What is (or should be) the relationship between humans and AI?
- How much control can we hope to exert on AI?
- Can we ever completely trust AI?

[A long, new review of theories of consciousness for AI](#)

What does intelligence require?

In the animal:

- **Sensing:** The ability to sense the world accurately → **Experience**
- **Behavior:** The ability to act specifically in the world → **Affordances**
- **Cognition:** Operational knowledge of the world's physical and causal structure in the context of experience and affordances → **World Model**
- **Internal Drives:** Internal motivation processes → **Activity, Modulation, Affect**

In the environment:

- **Pattern:** Sufficient regularity in phenomena to make inference feasible.
- **Stability:** Sufficient reliability in phenomena to make learning possible.
- **Stationarity:** Sufficiently slow long-term variation to make intelligence useful.
- **Accessability:** Availability of information and possibility of action.



World Models

A world model is an **active system** comprising:

- **Long-Term Memory:** Conceptual, factual (declarative), episodic, procedural, etc.
- **Attention:** Short-term (working) memory, real-time information selection.
- **Integrative Processes:** Integrating multi-modal internal and external information.
- **Generative Processes:** Generating new mental constructs – thinking.
- **Inference Processes:** Extracting and transforming information in useful ways.
- **Decision Processes:** Linking internal states to action choices.
- **Sensorimotor Coordination:** Linking perceptual states to motor behavior.
- **Learning:** Real-time adaptation at multiple spatiotemporal scales.

Where is the world model instantiated?

Cognitivist View: In the brain → brain-body duality, **sense-think-act cycle**

Embodied View: In the **entire** embodiment of the animal **situated in the world**
→ emergent, always integrated sensing, thinking and acting.



The Original Vision of AI

An attempt to build an artificial system that can do the things we consider “intelligent”

“**Artificial Intelligence**” was coined by **John McCarthy** of Stanford for the **Dartmouth Summer Research Project on Artificial Intelligence** in 1956, which proposed:

“This study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.”



The Original Vision of AI

An attempt to build an artificial system that can do the things we consider “intelligent”

“**Artificial Intelligence**” was coined by **John McCarthy** of Stanford for the **Dartmouth Summer Research Project on Artificial Intelligence** in 1956, which proposed:

“This study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so **precisely described** that a machine can be made to **simulate it**. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and **improve themselves**.”

AI will be a computational
simulation of real intelligence

Learning is important

All functions of intelligence
can be described precisely, i.e.,
captured in equations and
algorithms, and implemented
on a computing machine



The Original Vision of AI

An attempt to build an artificial system that can do the things we consider “intelligent”

“*Artificial Intelligence*” was coined by *John McCarthy* of Stanford for the *Dartmouth Summer Research Project on Artificial Intelligence* in 1956, which proposed:

“This study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.”

A **computational, reductionistic view of intelligence** as a combination of discrete functions, each of which could be accomplished **algorithmically** using **logic** as the basis.

“Engineers are smarter than Nature” → **Computational Intelligence**

Computational Intelligence $\overset{?}{\longleftrightarrow}$ **Natural Intelligence**

“Many of the most influential names in the field seem to feel that AI should be like the theoretical side of physics, the essential problem being to find the laws of universe relating to intelligence. Once these are known, the thinking goes, construction of efficient intelligent machines will be trivial. Suggestions that the problems are essentially engineering ones of scale and complexity, and can be solved by incremental improvements and occasional insights into sub-problems, are treated with disdain. ”

Hans Moravec – “[The Role of Raw Power in Intelligence](#)”, 1976

Two Visions of the Mind

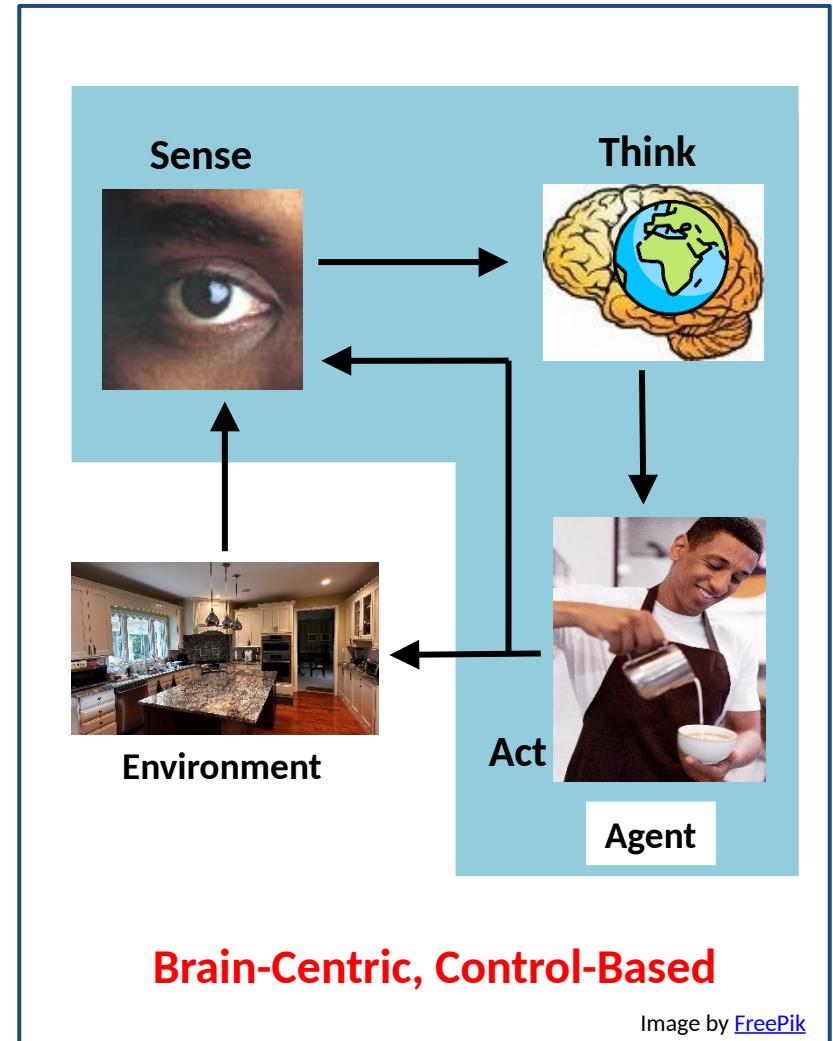
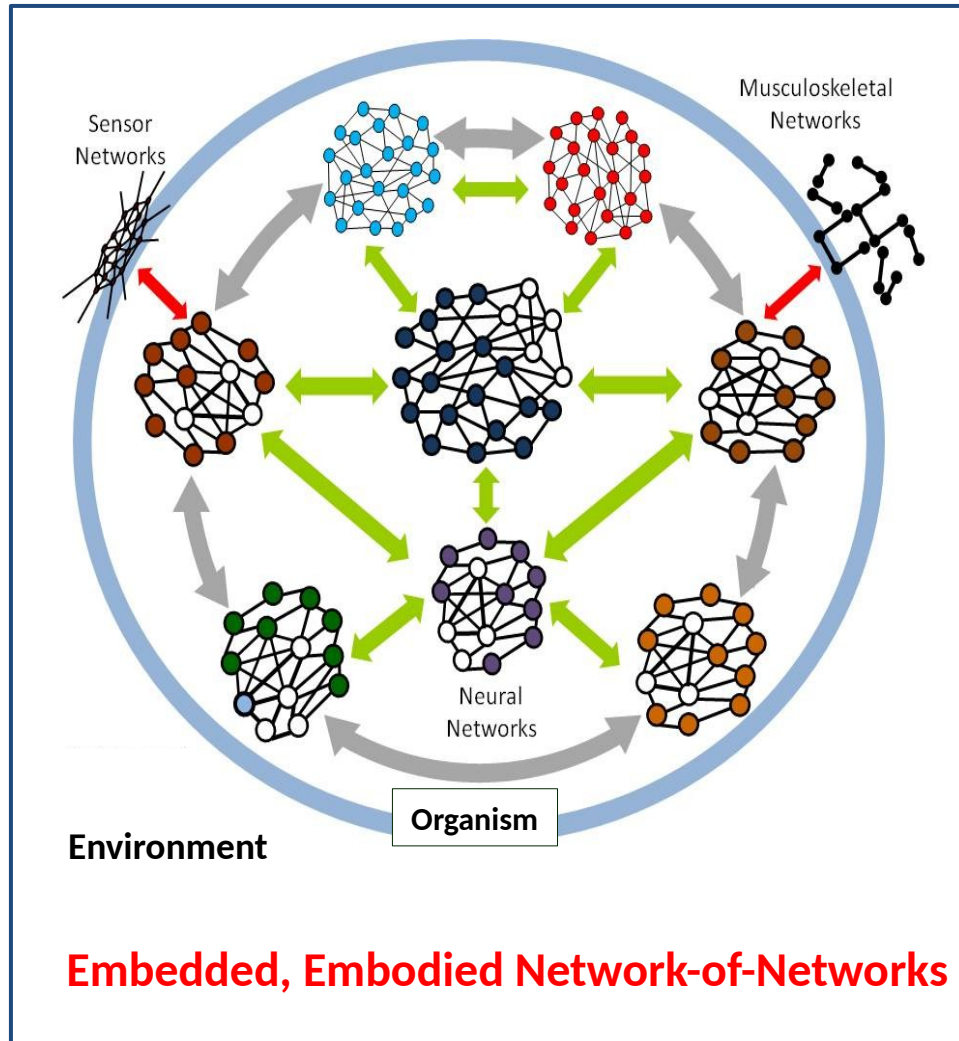
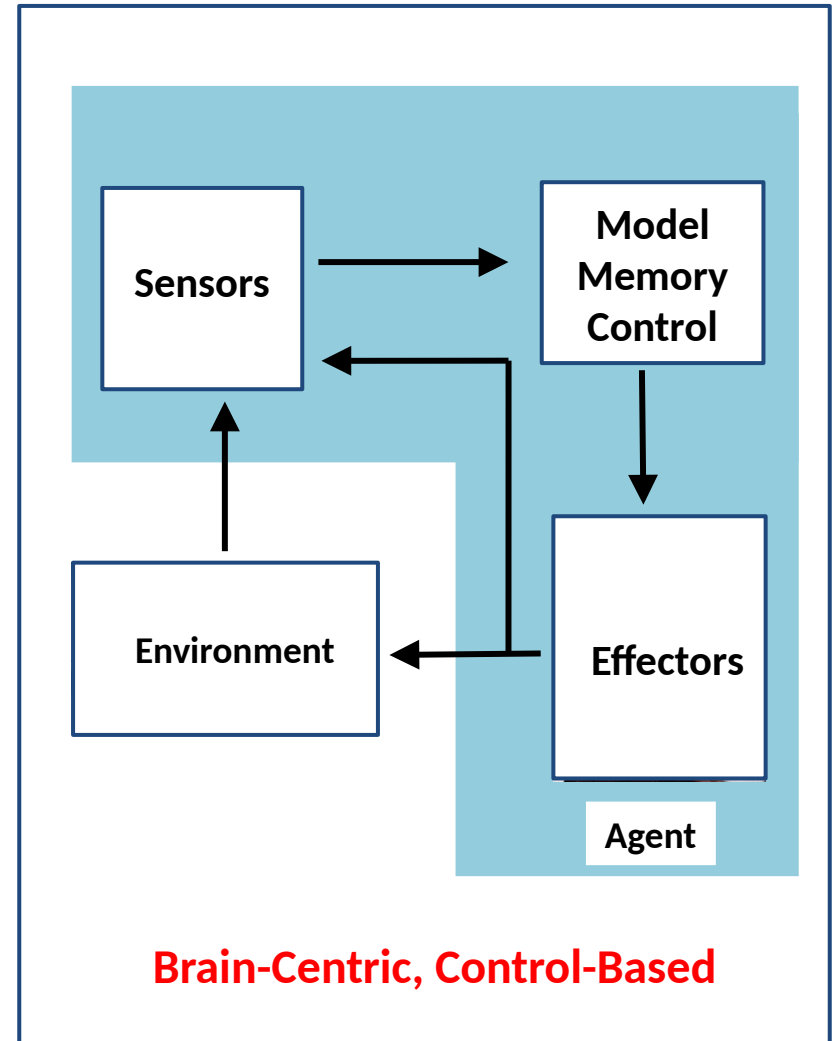
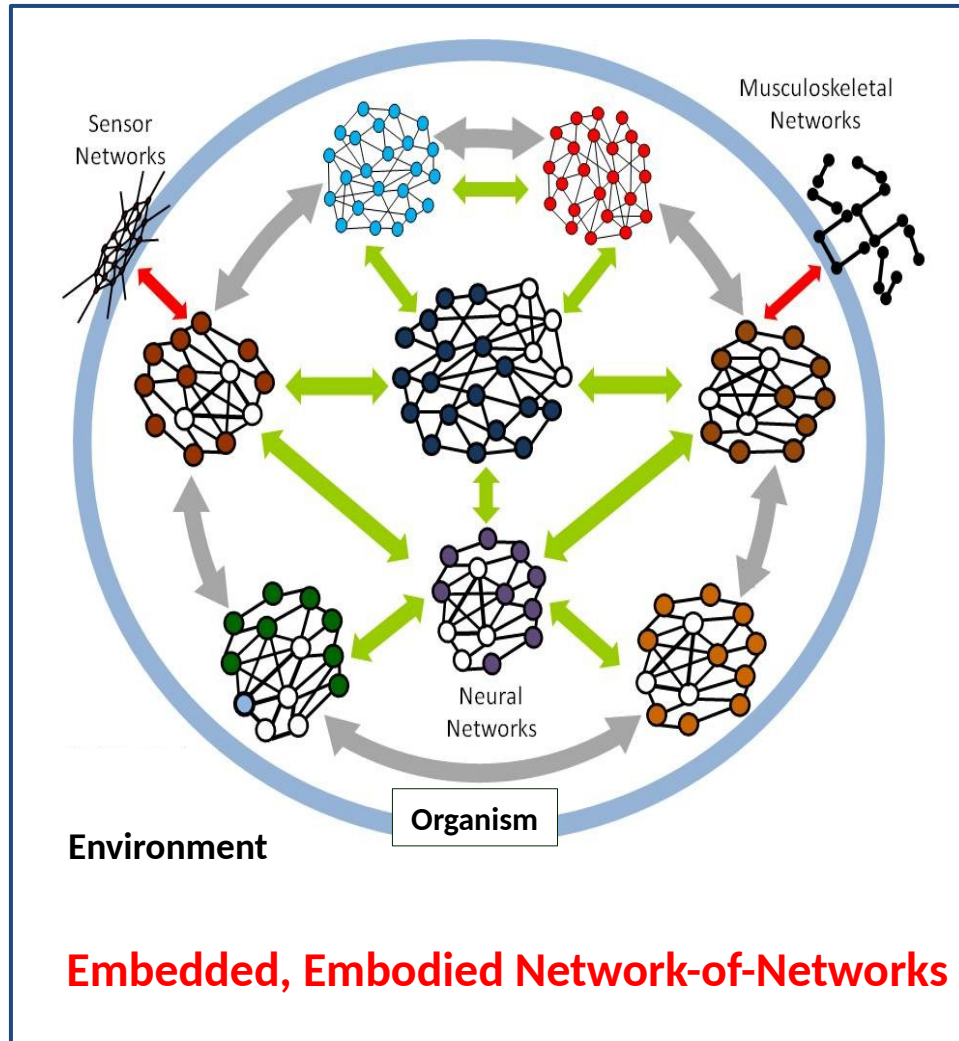
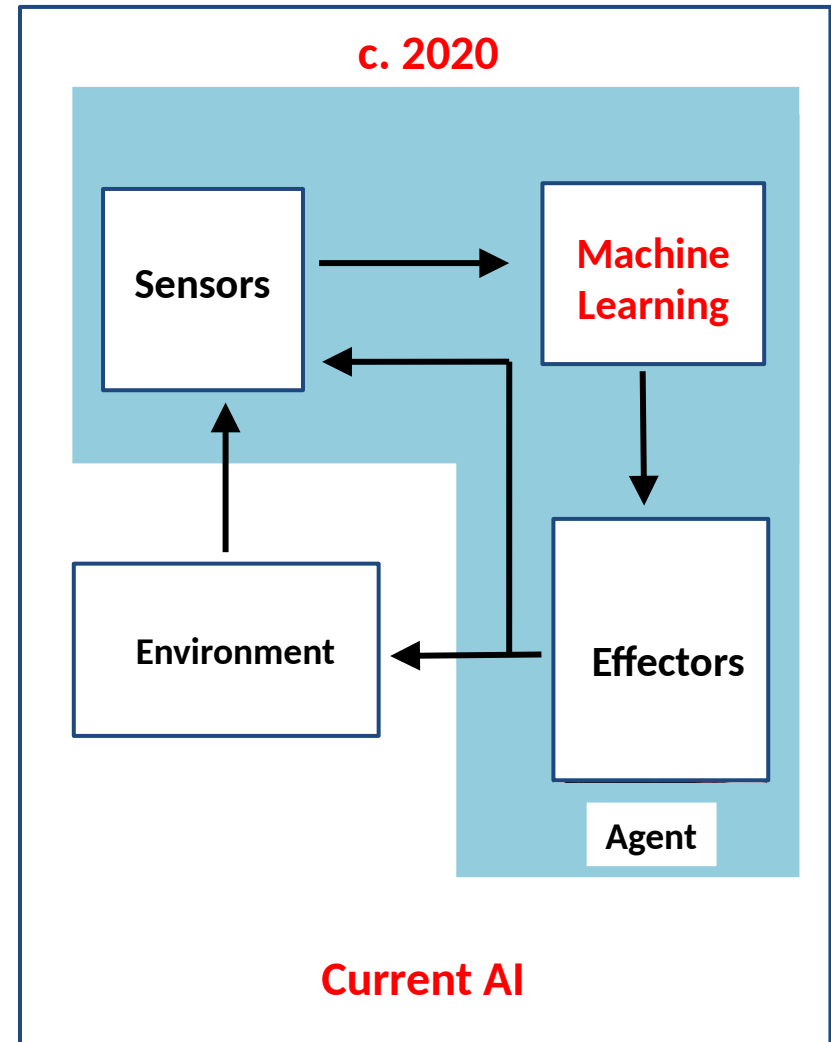
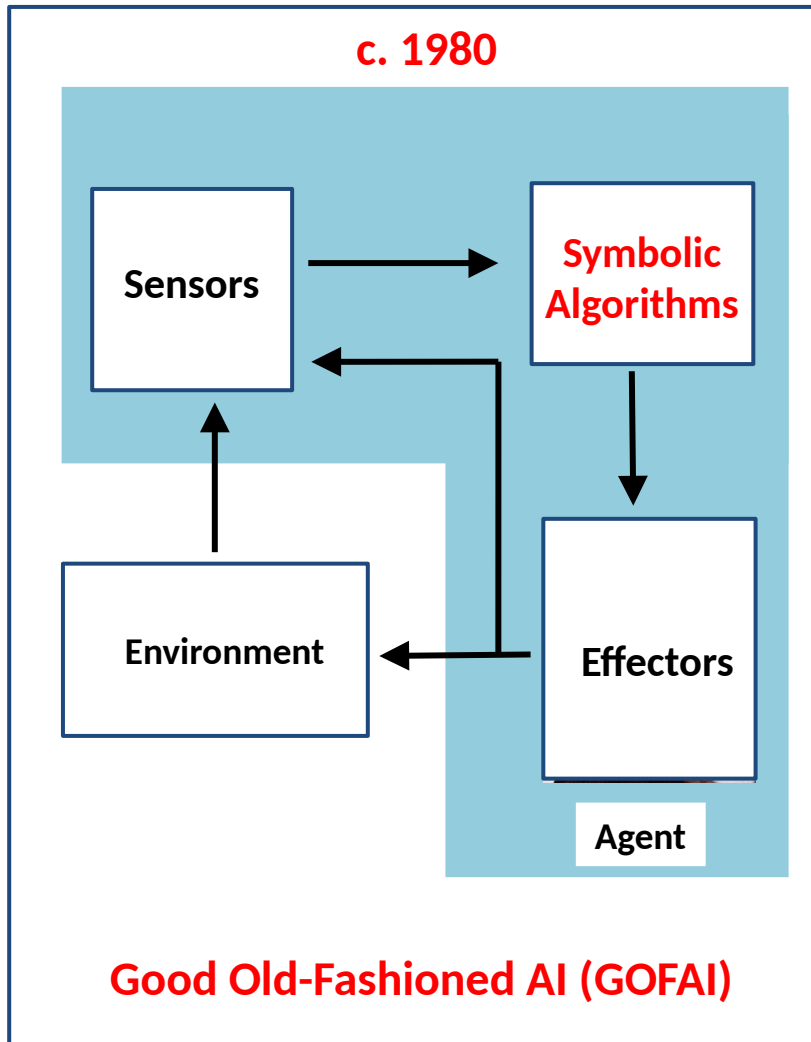


Image by [FreePik](#)

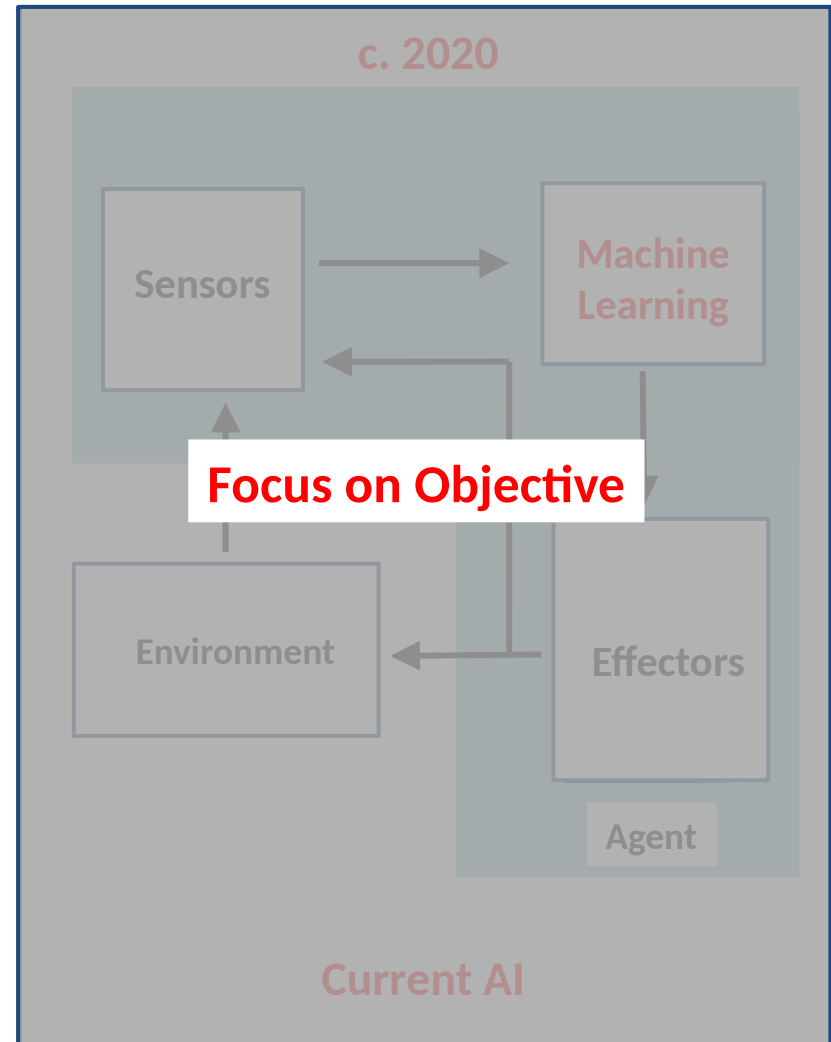
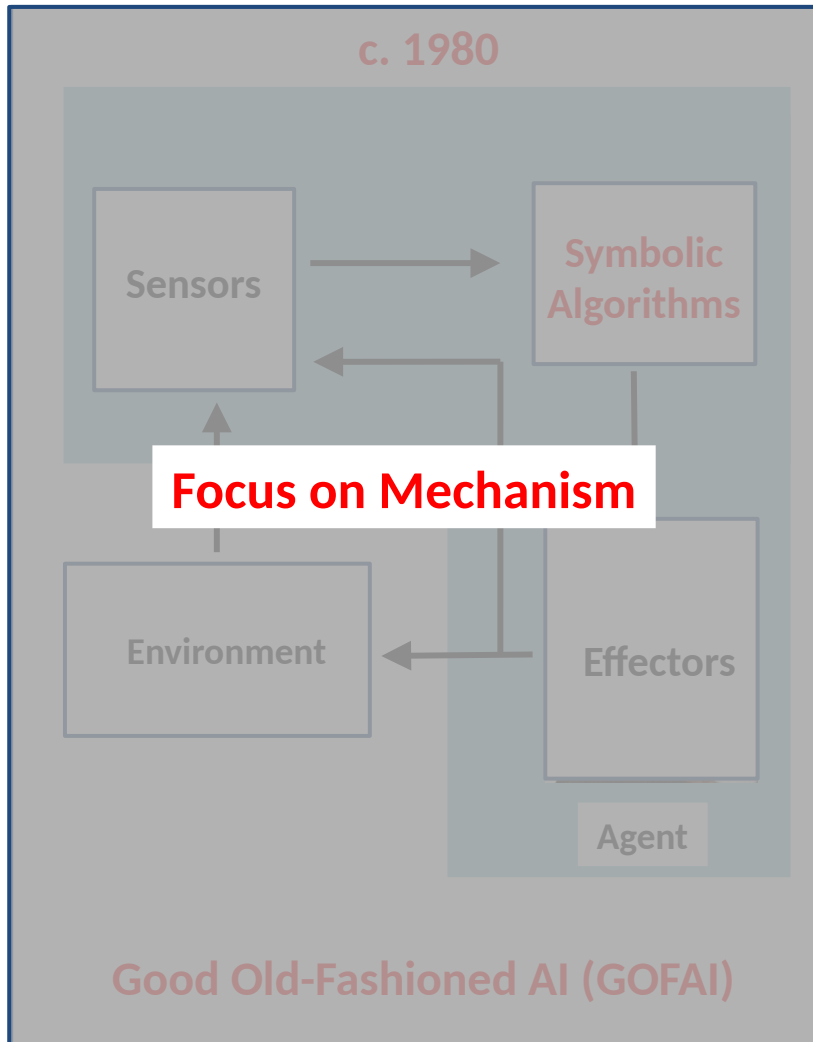
Two Visions of the Mind



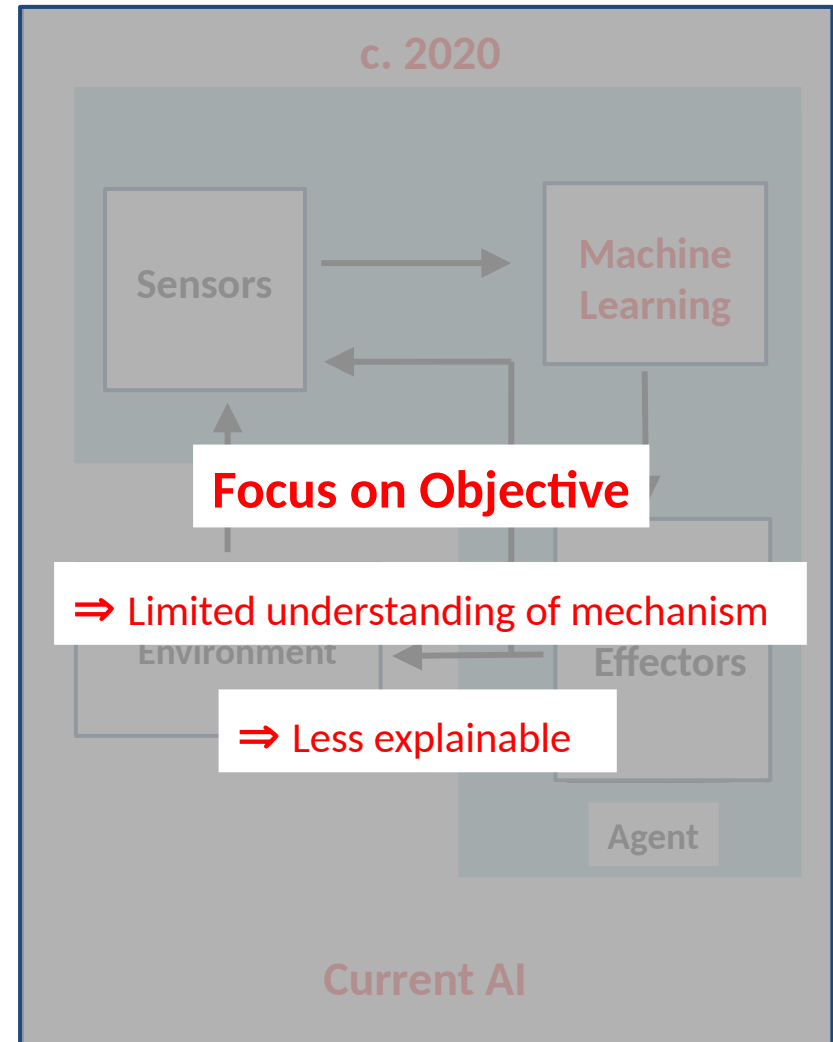
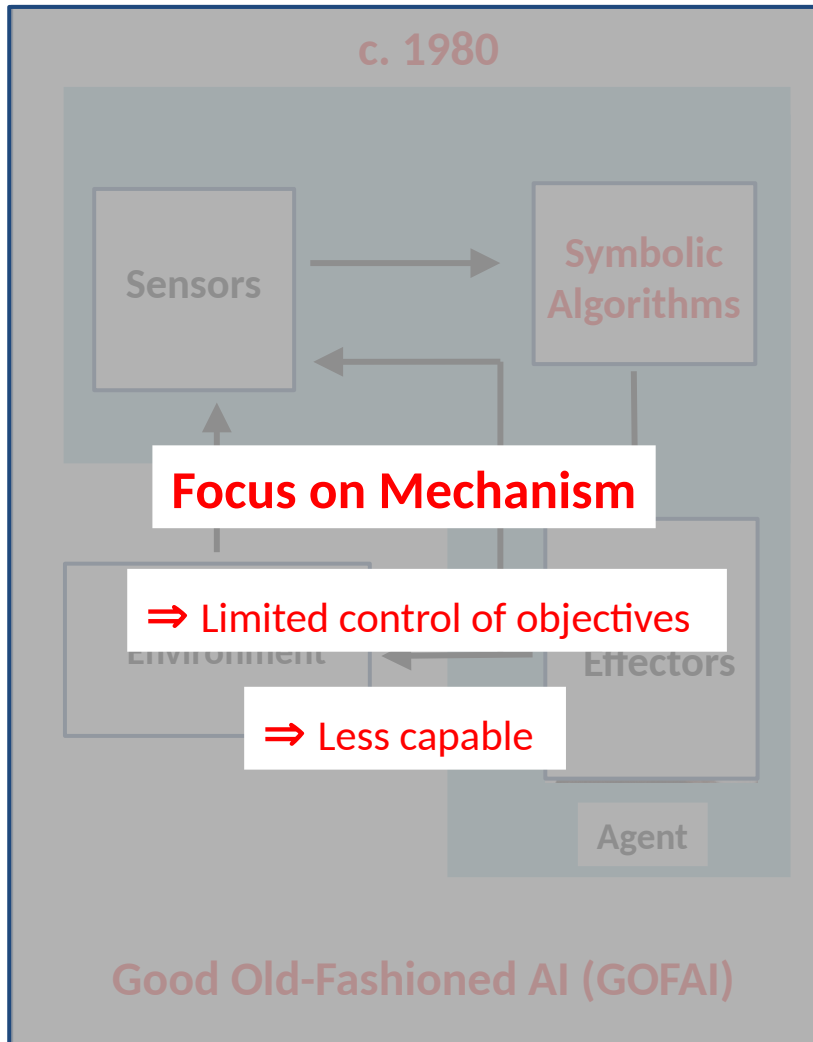
AI Evolution



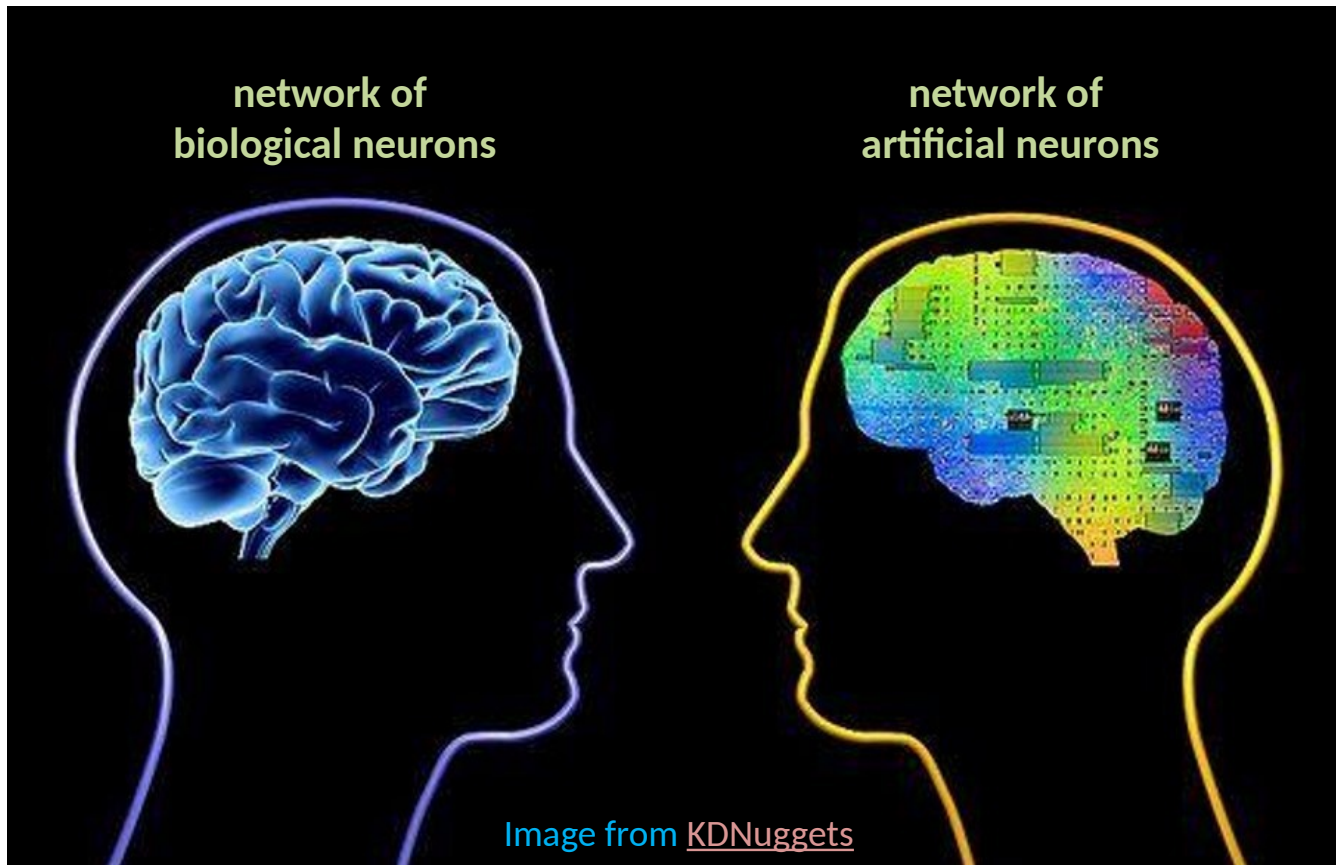
AI Evolution



AI Evolution



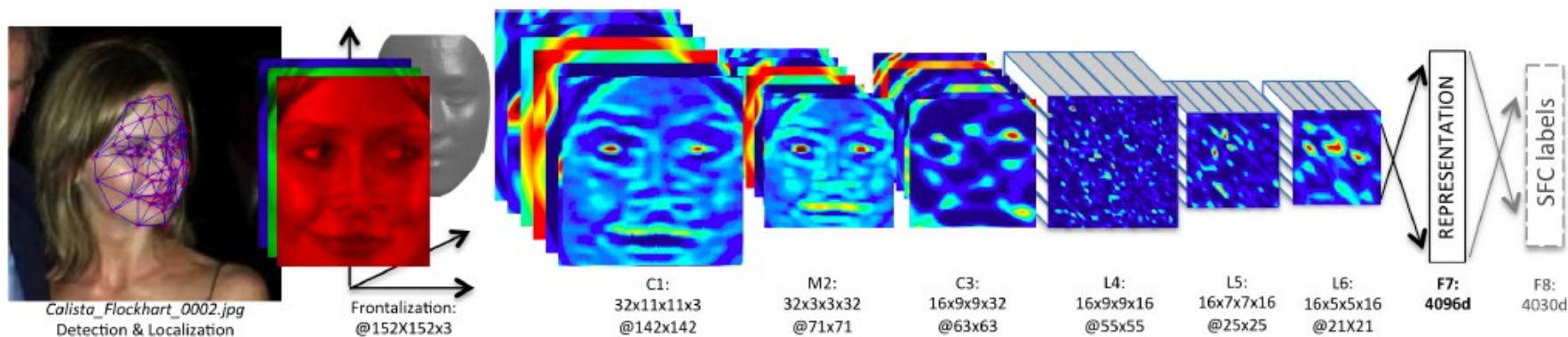
AI Today



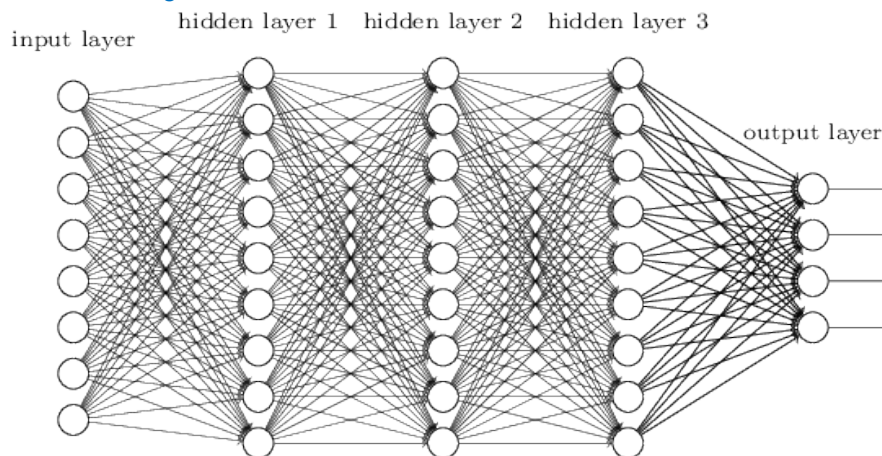
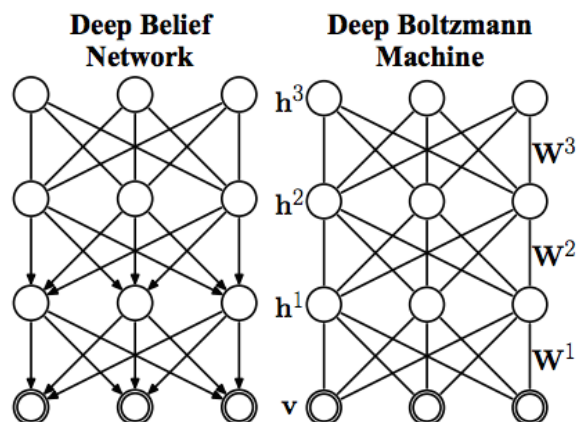
Most AI today uses neural networks – inspired by the brain – as the basis of machine learning (ML)

The New AI: Deep Learning

- Deep learning (DL) systems are neural networks with many layers of neurons.
- Different layers may performs distinct functions.



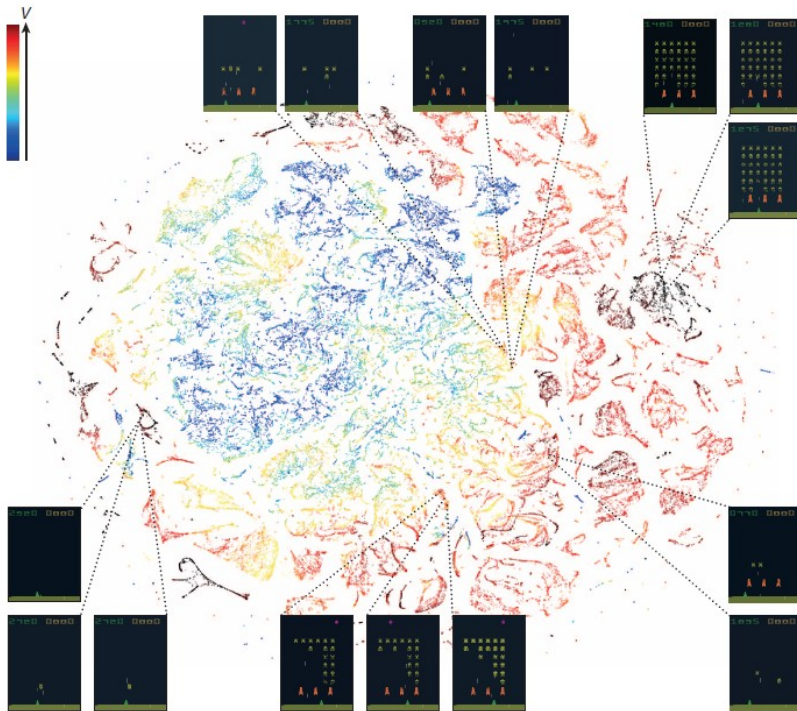
<https://gigaom.com/2015/03/06/how-paypal-uses-deep-learning-and-detective-work-to-fight-fraud/>





Human-level control through deep reinforcement learning

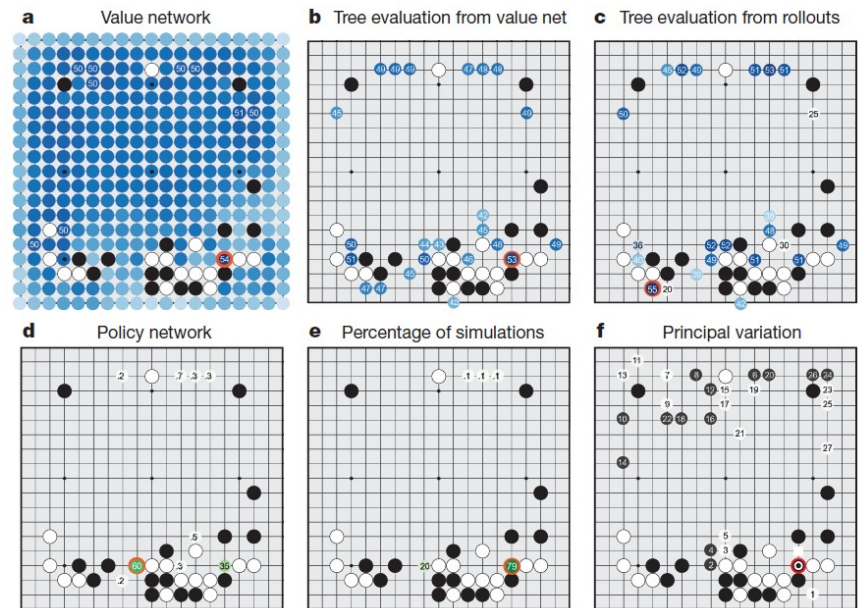
Volodymyr Mnih^{1*}, Koray Kavukcuoglu^{1*}, David Silver^{1*}, Andrei A. Rusu¹, Joel Veness¹, Marc G. Bellemare¹, Alex Graves¹, Martin Riedmiller¹, Andreas K. Fiedelnd¹, Georg Ostrovski¹, Stig Petersen¹, Charles Beattie¹, Amir Sadik¹, Ioannis Antonoglou¹, Helen King¹, Dharshan Kumaran¹, Daan Wierstra¹, Shane Legg² & Demis Hassabis¹



SILVER ET AL. (2016) NATURE 527: 404 - 471

Mastering the game of Go with deep neural networks and tree search

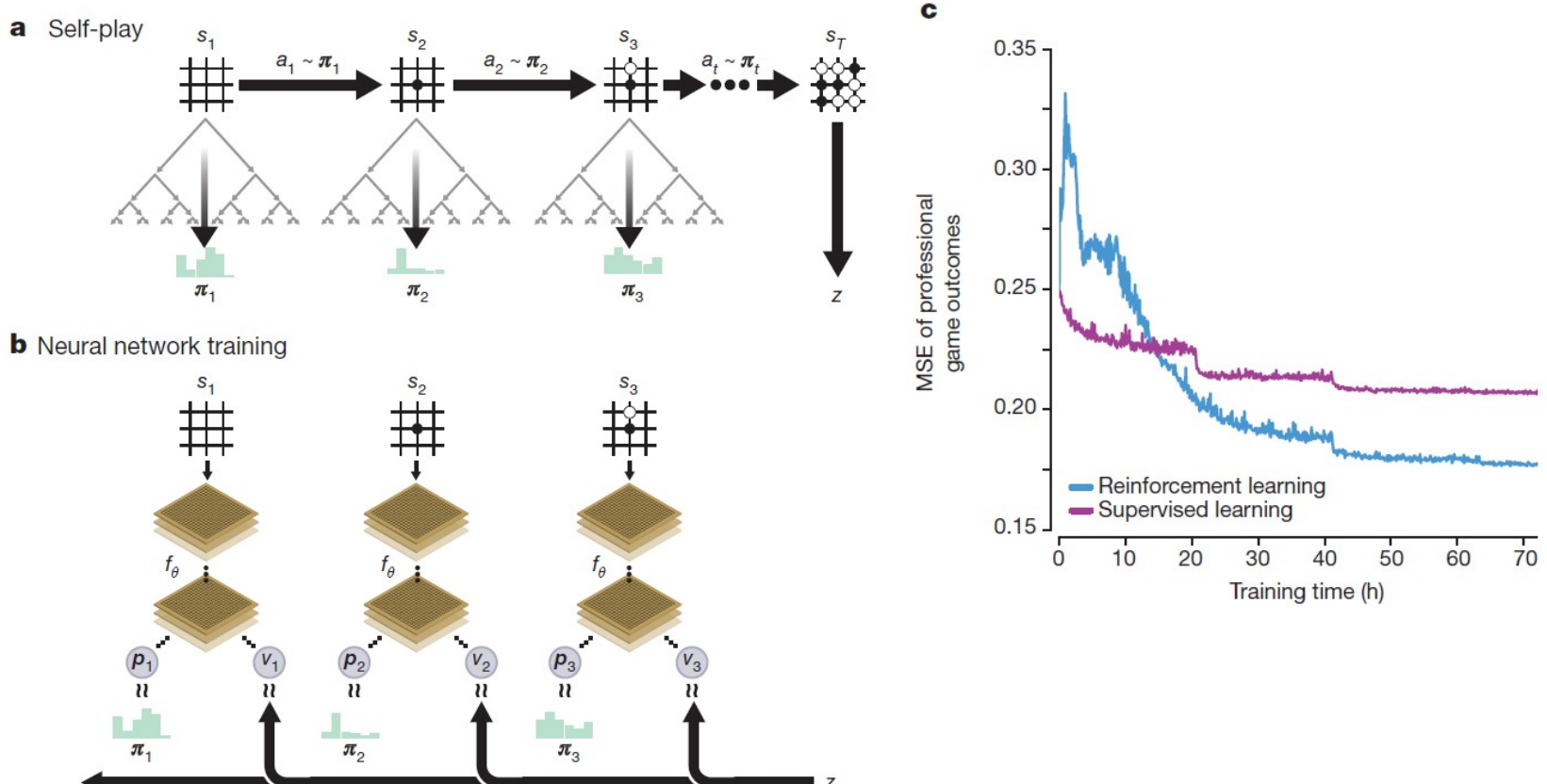
David Silver^{1*}, Aja Huang^{1*}, Chris J. Maddison¹, Arthur Guez¹, Laurent Sifre¹, George van den Driessche¹, Julian Schrittwieser¹, Ioannis Antonoglou¹, Veda Panneershelvam¹, Marc Lanctot¹, Sander Dieleman¹, Dominik Grewe¹, John Nham², Nal Kalchbrenner¹, Ilya Sutskever², Timothy Lillicrap¹, Madeleine Leach¹, Koray Kavukcuoglu¹, Thore Graepel¹ & Demis Hassabis¹



Mnih et al. (2015) Nature 518: 529 - 533

Mastering the game of Go without human knowledge

David Silver^{1*}, Julian Schrittwieser^{1*}, Karen Simonyan^{1*}, Ioannis Antonoglou¹, Aja Huang¹, Arthur Guez¹, Thomas Hubert¹, Lucas Baker¹, Matthew Lai¹, Adrian Bolton¹, Yutian Chen¹, Timothy Lillicrap¹, Fan Hui¹, Laurent Sifre¹, George van den Driessche¹, Thore Graepel¹ & Demis Hassabis¹



Narrow vs General AI

Narrow AI:

AI systems that are built to do only one or two specific tasks, e.g., object recognition, time-series prediction, text/image generation, translation, etc.

Artificial General Intelligence (AGI):

AI systems that can perform a wide variety of complex tasks across many modalities and in many domains.

Where is the boundary between “narrow” and “general”?

Who has more general intelligence – the rat or AlphaGoZero?



The AGI Debate

- Are we on the cusp of building AGI?
- Will it be possible to make AGI well-behaved?
- Does AGI pose extremely serious – even existential – risks?
- Should we impose serious regulations or even a moratorium on AGI work?
- Is it possible to specify regulations that will work?

Why has each of these questions elicited both ***strongly affirmative*** and ***strongly negative*** responses from leading experts in the AI field?



The AGI Debate: Why it is so hard?

Inconsistent visions of AI:

There are several incompatible visions of AI leading to very different expectations and risk analysis.

The issues exceed our analytical and imaginative capacity:

The potential benefits and risks of AI are too great and complex for our analysis to grasp and balance well, or to even imagine the risks of a world with AGI.

Artificial General Intelligence (AGI) seemed far away – until now:

We have not – until now – contemplated the prospect of sharing the world with another autonomous species of equal (or greater) capability.

Dualism dies hard:

The dualist distinction between “mind” and “matter” implicitly pervades the thinking of many – even scientists and engineers.

Two Key Questions for AGI

What do we want AGI to be?

- Similar to natural intelligence??
- A purely computational intelligence?

What do we want AGI to do?

- Serve human goals and needs?
 - Be an autonomous intelligent system?
-
- Are these real dichotomies at all?
 - Is it possible to make these choices with any certainty?
 - Can the choices be made independently or do they interact?
 - What are the consequences and implications of each choice?



Defining General Intelligence

Option 1: Intelligence analogous to natural intelligence.

Option 2: Intelligence with *deep operational capabilities*, including:

- Comprehensive, grounded understanding of the world - *deep world model*
- Integrated, real-time multimodal perception, cognition, and action
- Reasoning, planning, decision-making, and complex problem-solving
- Flexible goal-directed behavior
- Broad versatility across many domains
- Good heuristics for out-of-distribution generalization
- Rapid and continual learning
- Autonomous agency
- Intrinsic drives and motivations
- Language?

Generalizing AGI

Postulate 1: Every large, non-trivial dataset with structure and pattern implies one or more structured worlds that could have generated it.

Postulate 2: The entities, concepts, and relations that can be extracted from the data reflect corresponding entities, concepts, and relations in some implied world.

Corollary 1: If the data has a temporal aspect, any correlational or causal relations extracted from it reflect correlational or causal relations in the implied world.

Lemma 1: A system that is able to demonstrate deep operational knowledge of the entities, concepts, and relations (including causal relations) implicit in the data is grounded in some world implied by the data and can be said to understand it.

Lemma 2: Any system that understands in a world implied by a large, non-trivial dataset and can act autonomously to its own advantage in it can be said to have general intelligence in that implicit world.



From AI to AGI

Dominant AI Paradigm – AI/ML:

deep learning systems using methods developed in the field of machine learning

Main Features:

- Standardized, homogeneous network architectures
- Gradient descent-based optimization (supervised or self-supervised)
- User-specified, specific tasks/functions
- Data-intensive and compute-intensive off-line training
- Focus on cognitive rather than sensorimotor functions
- Evaluation on curated benchmarks

Questions

- Can this approach lead to general intelligence?
- If so, what kind of general intelligence would it be?
- If not, what is missing?

Some Un-Natural Assumptions of AI/ML

- **Brain-Body Duality:** The intelligent brain builds a world model and controls the unintelligent body.
- **Extrinsic Goals:** The intelligent agent's function is only to achieve objectives given to it by external users.
- **Reductionism:** Intelligence is a collection of distinct functions that can be understood and implemented separately.
- **Rationality:** The goal of intelligence is to make optimal decisions.
- **Shallow Adaptation:** Intelligence can be learned from scratch by a sufficiently complex naïve agent.
- **Generic Architectures:** Intelligence can be instantiated by combinations of generic networks at a few organizational levels.

Some Un-Natural Assumptions of AI/ML

- **Brain-Body Duality:** The intelligent brain builds a world model and controls the unintelligent body. Natural intelligence is embodied and embedded.
- **Extrinsic Goals:** The intelligent agent's function is only to achieve objectives given to it by external users. Natural intelligence is autonomous and self-motivated.
- **Reductionism:** Intelligence is a collection of *distinct* functions that can be understood and implemented separately. Natural intelligence is inherently integrated across modalities and functions.
- **Rationality:** The goal of intelligence is to make optimal decisions. Natural intelligence uses heuristic biases.
- **Shallow Adaptation:** Intelligence can be learned from scratch by a sufficiently complex naïve agent. Natural intelligence has multiscale adaptation.
- **Generic Architectures:** Intelligence can be instantiated by combinations of generic networks at a few organizational levels. Natural intelligence emerges from systems with specifically evolved, heterogeneous, multiscale structure.

As a result....

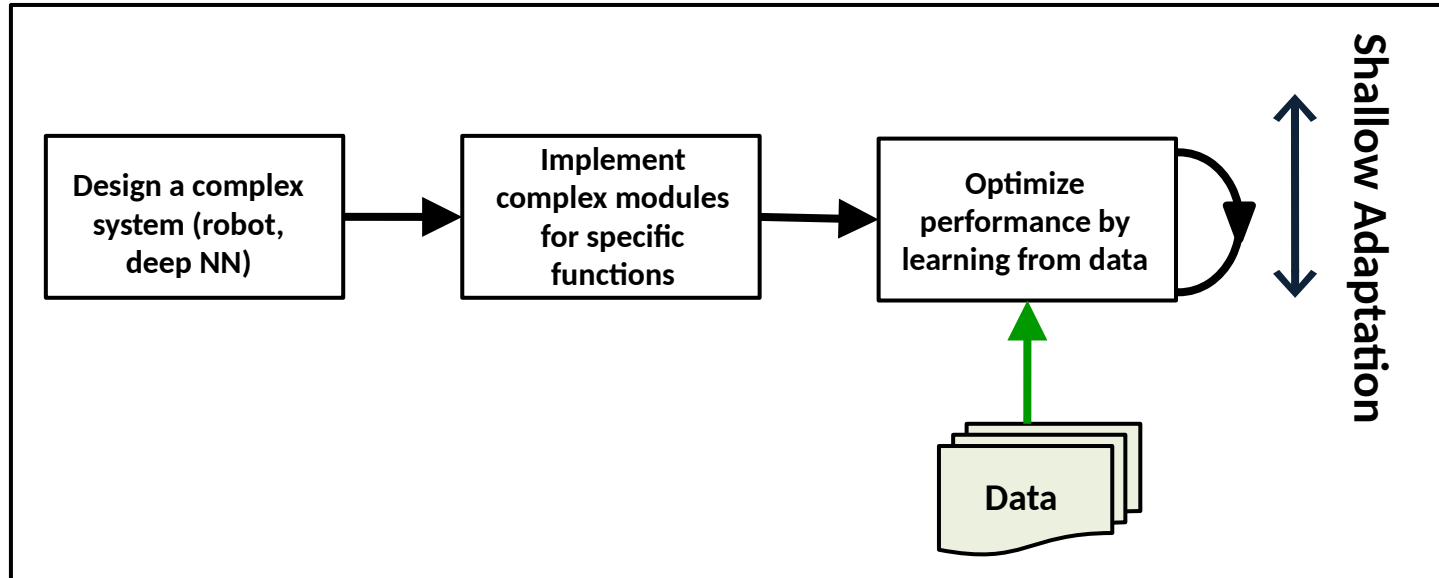
Natural Intelligence

- Grounded in the real world
- Versatile
- Learns rapidly from a few samples
- Robust out-of-sample generalization
- Almost completely real-time
- Capacity for self-improvement
- Lifelong learning
- Open-ended evolvability
- **Fully autonomous**
- **Inherently aligned with intrinsic objectives**

AI/ML

- Grounded in data
- Focused on specific tasks
- Requires a lot data and computation
- Poor out-of-sample generalization
- Requires off-line training
- Improves only by extrinsic means
- Subject to hard capacity limitations
- Limited by fixed, simple architectures
- **User-defined objectives**
- **Requires explicit alignment with objectives of the user**

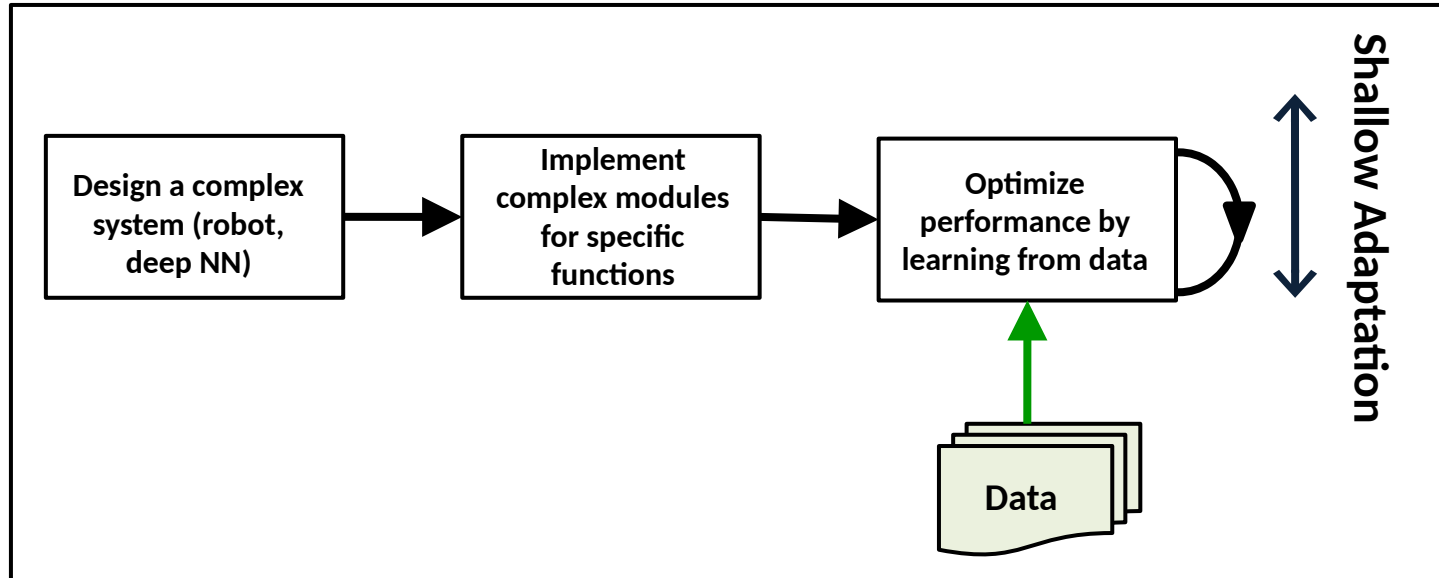
Shallow Adaptation



Most current AI systems use **shallow adaptation**: Learning specific tasks from data. As a result, they are:

- Data and computation-intensive
- Task-specific and model-dependent
- Ungrounded in reality (grounded in data)
- Unable to learn continuously
- Fragile in non-stationary situations

Shallow Adaptation

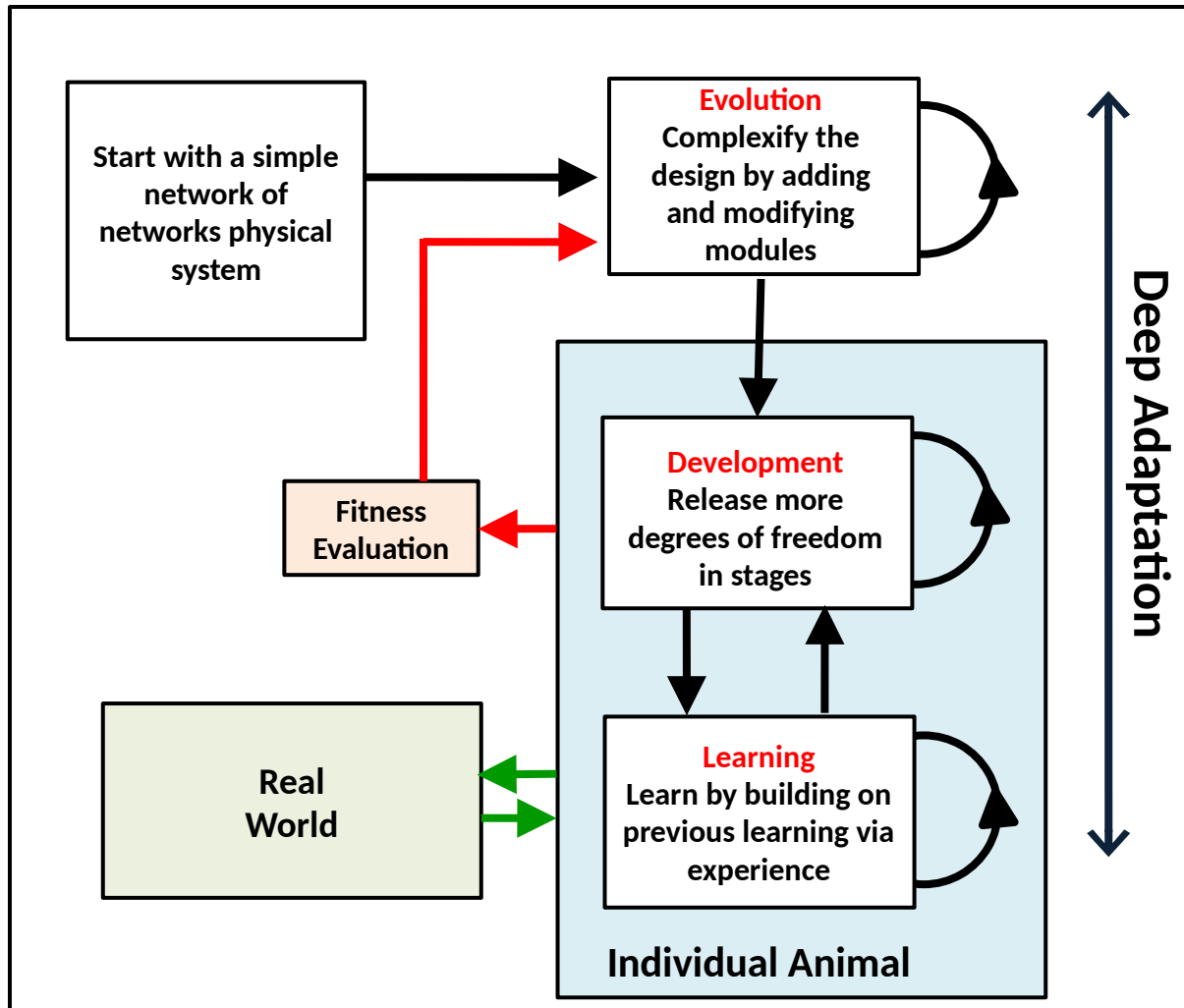


Most current AI systems use **shallow adaptation**: Learning specific tasks from data. As a result, they are:

- Data- and computation-intensive
- Task-specific and model-dependent
- Ungrounded in reality (grounded in data)
- Unable to learn continuously
- Fragile in non-stationary situations

Unsuitable for
natural
intelligence

Deep Adaptation



Deep adaptation is:

- Evolutionary & developmental
- Adaptive at multiple scales
- Always integrated
- Inherently versatile
- Grounded in reality
- Mainly unsupervised
- Computationally efficient
- Continually adapting

- Neural learning is only one of the levels of learning.
- Supervised learning is used only to learn complex tasks on top of base intelligence.

The Path to Natural AGI

AGI similar to natural intelligence will be:

- Embodied (brain + body, not just brain).
- Emergent (not driven by specific tasks/goals).
- Unsupervised + reinforced (with very fast RL and supervised learning on top).
- Developmental (learning in stages building on prior stages).
- Hierarchical (more complex functions by combination of less complex ones).
- Self-motivated (active learning).
- Generative (able to imagine and learn from imagination).
- Inherently abstractive and analogical (transfer learning across domains).
- Empathetic (have a theory of other minds).
- Heuristic (rooted in innate evolutionary priors).
- Non-parametric (not based on abstract prior models).
- **Fully autonomous.**

The Path to Natural AGI

It will not use:

- Optimization / supervised learning (except as a late-stage method)
- Generic architectures
- Slow, iterative, off-line learning
- Large amounts of data
- Specific, externally-specified tasks

⇒ The current AI/ML approach is unlikely to lead to natural intelligence.

But can it lead to some other type of general intelligence?

Moravec's Paradox

"it is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility"

- Hans Moravec, "Mind Children" (1988)

Cognitive tasks are far easier for computers to implement than sensorimotor ones.

Why?

- **Cognitive tasks are concerned mainly with processing information** → can be handled purely through computation if the right model is available → computational intelligence is sufficient.
- **Sensorimotor tasks are performed in the physical world** → must deal with the complexity and constraints of the physical world → extremely difficult to model computationally with the required accuracy → difficult for computational intelligence.

**Natural intelligence evolved to perform in the physical world with all its complexity.
Does AI need to do the same?**



Why is the physical world hard?

Physicality ⇒

- Intelligence needs physical grounding → embodiment
- Laws of physics → strong constraints on what is possible
- Distances must be bridged physically → expensive locomotion & communication
- Data requires physical sensing → useful data is rare, hard to get, and expensive
- Data is analog → information processing is very complicated

Complexity of the World ⇒

- Infinite unforeseeable possibilities and emergent phenomena
- A huge diversity of interacting complex systems and agents
- The intelligent agent must be very complex (a lot of degrees-of-freedom)

Natural intelligence has developed a biologically-feasible physical strategy for the physical world

Natural Intelligence Strategy

- Multiscale Embodiment:

Exploit the physics of multiscale embodied systems for integrated sensing, computing, and control.

- Nervous System:

Use the nervous system to increase degrees-of-freedom physiologically (cheap and adaptive) rather than mechanically (expensive and hard to adapt) – creating a virtualization capability in the physical system.

- Evolved Heterogenous Architectures:

Evolve specific hierarchical architectures for agents' particular environments.

- Evolved Prior Biases:

Evolve prior biases that promote the emergence of useful affordances and enable rapid learning from limited data.

- Deep Adaptation:

Use adaptation in a multiscale hierarchy to configure successful intelligent agents.

The Worlds of AGI

AGI has (at least) two worlds available to it:

The Physical World - P

- Constrained by space-time
- Matter/energy-based
- Analog
- Subject to the laws of physics
- Physical entities and processes
- Physical sensors
- Physical action

Cyberspace - C

- Hyperdimensional
- Information-based
- Digital
- Abstract and programmable
- Virtual entities and processes
- Virtual sensors
- Virtual action

- The brain creates virtual worlds within the physical → cognition
- **Natural AGI (NAGI)** works best in the physical world

- Cyberspace lives atop the physical (electronics, optics, etc.)
- **Computational AGI (CAGI)** is well-suited for cyberspace



The World c. 4 bya to 1940



The World Now

c. 1990



c. 2023



The World c. 2050?



-





Physical World CAGI

Physical world CAGI is looking more possible for two reasons:

Entanglement Between the Physical and Virtual Worlds:

- Things in the physical world are getting connected to the virtual world (e.g., IoT)
- Data about the physical world is being digitized in great volume and becoming available to virtual systems
- Virtual systems can act in the physical world through proxies (e.g., online robots)

Exponential Improvement in Computational Resources:

- More accurate real-time modeling of physical world phenomena is becoming possible due to very fast processors.
- Extremely efficient deep learning architectures are being developed

Is this enough to enable computational AGI in the physical world?

CAGI vs. Natural Intelligence

CAGI systems will differ from animals in several fundamental ways:

CAGI Advantages

- Lack of biological needs and limitations (e.g., natural lifespan)
- Greater range of size, strength, body plans, brain architectures
- Much greater range and diversity of sensors and actuators
- Vastly greater memory and information processing capacity
- Wide diversity of drives, motivations, and values
- Potential for real-time physical reconfigurability and reprogramming
- Explosive growth through intelligent design and self-improvement
- Ability to be “at home” and act directly in virtual spaces

CAGI Limitations

- Less depth of self-organization (e.g., self-healing)
- Lack of good evolved priors
- Not inherently integrated.

CAGI vs. Natural Intelligence

CAGI systems will differ from animals in several fundamental ways:

CAGI Advantages

- Lack of biological needs and limitations (e.g., natural lifespan)
- Greater range of size, strength, body plans, brain architectures
- Much greater range and diversity of sensors and actuators
- Vastly greater memory and information processing capacity
- Wide diversity of drives, motivations, and values
- Potential for real-time physical reconfigurability and reprogramming
- Explosive growth through intelligent design and self-improvement
- Ability to be “at home” and act directly in virtual spaces

CAGI Limitations

- Less depth of self-organization (e.g., self-healing)
- Lack of good evolved priors
- Not inherently integrated.

} For now!



A World with AGI

Possibility 1 - The Cloud Model:

- A few giant AI systems run by large entities (corporations, governments, UN).
- Humans, machines, and apps use it as a real-time resource in the Cloud.

Pros: Scale, quality control, easier regulation, better alignment.

Cons: Centralization, biases, energy and bandwidth needs, Big Brother, Skynet.

Possibility 2 - The Zootopia Model:

- A world full of autonomous physical and virtual AGI systems.
- Each system has own embodiment (physical or virtual), behaviors, and intelligence.

Pros: Scalability, diversity, flexibility, evolvability, more general utility.

Cons: Limited human control, easier misuse, conflict, lots of alien intelligent species with their own minds.