An AI Review

- collaborative workshop

CAS ADS M5

University of Bern

1. Group 1 - Chapter 1 and 2 (Salomé, Hazel, Lenja, Daniëlle)
2. Group 2 - Chapter 3, 4, and 5 (Nicolas, Galina, Moataz, Tobias, Mayra)
3. Group 3 - Chapter 6 and 7 Marina
4. Group 4 - Chapter 8 and 9 (the ones onsite)

Each group creates max 20 slides and has 20 min for presentation and 20 minutes for discussion.

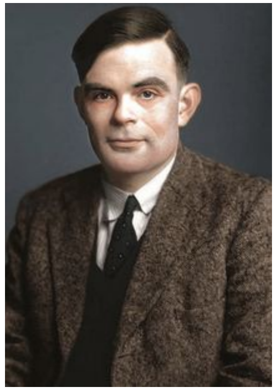
The History of AI - René Descartes 1637

Two certain tests to distinguish a machine from humans:

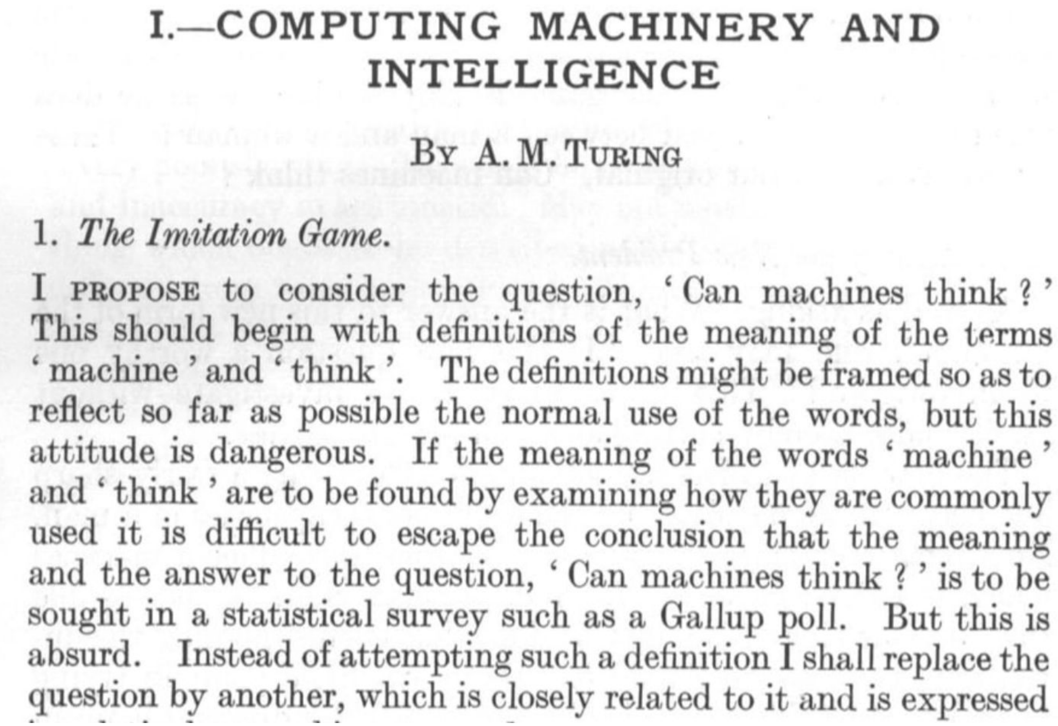
1. **Speech**: Machines can produce words or react to stimuli but cannot use language flexibly or meaningfully
2. **Reasoning and Adaptability**: Machines may perform specific tasks well but lack the universal reasoning and adaptability of humans

Conclusion: Human reason allows us to act and react appropriately **across a range of situations** whereas computers are always **limited to the pre-programmed responses**

The History of AI - Alan Turing



- Influential Paper ‘Computing Machinery and Intelligence’ 1950 asking fundamental questions about machine intelligence
- ‘Father of AI’



I.—COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

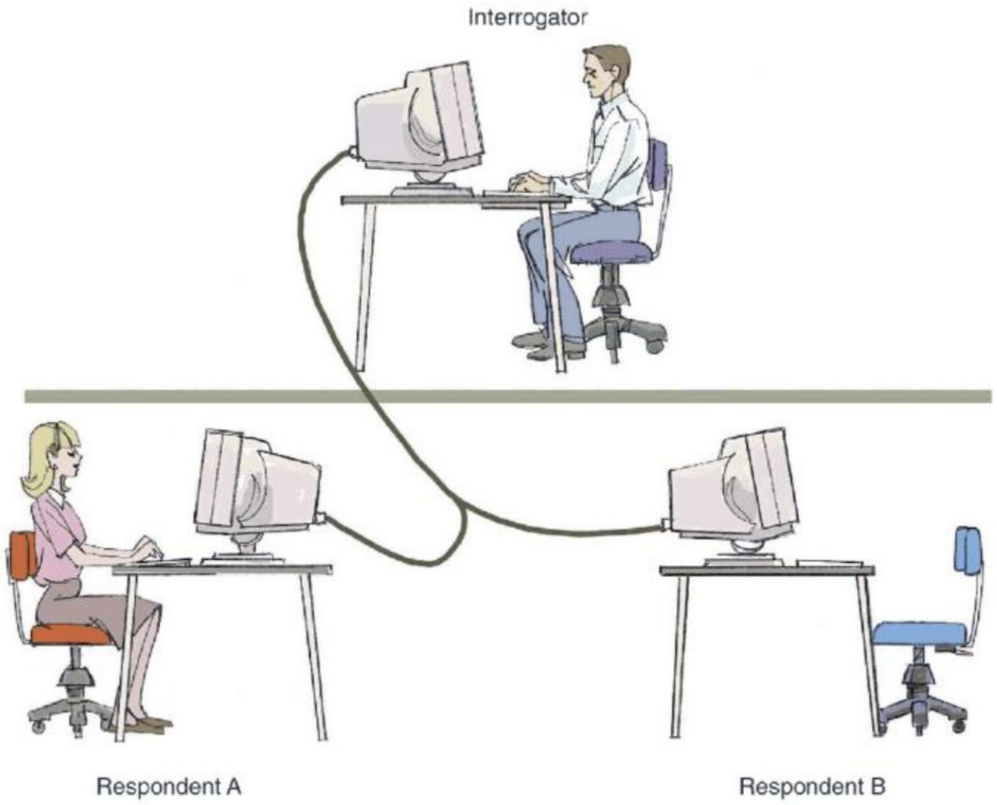
1. *The Imitation Game.*

I propose to consider the question, ‘Can machines think ?’ This should begin with definitions of the meaning of the terms ‘machine’ and ‘think’. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words ‘machine’ and ‘think’ are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, ‘Can machines think ?’ is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed

The History of AI - Turing Test

Turing Test:

‘Can a machine be linguistically indistinguishable from a human?’

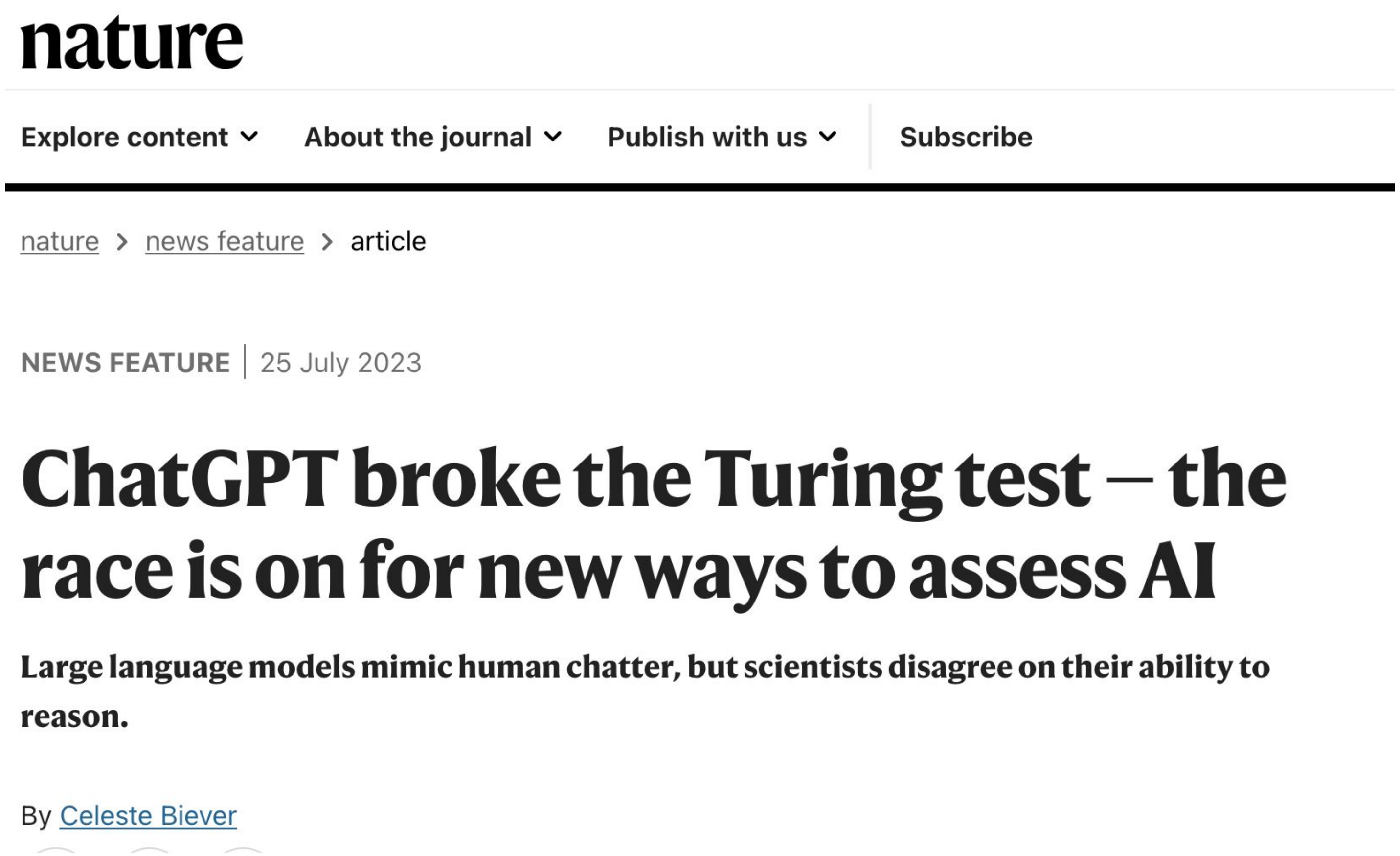


T

est Passed if judge can’t do better than correctly allocating 50/50 to machine/human

Turing Test

Turing himself predicted his test would be passed by 2000



nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄ Subscribe

nature > news feature > article

NEWS FEATURE | 25 July 2023

ChatGPT broke the Turing test — the race is on for new ways to assess AI

Large language models mimic human chatter, but scientists disagree on their ability to reason.

By Celeste Biever

Dartmouth Conference (1956): Birth of AI

- Official start of the field of artificial intelligence (AI)
- Famous attendees, sponsored by DARPA (Agency with the goal of creating breakthrough technologies for national security)

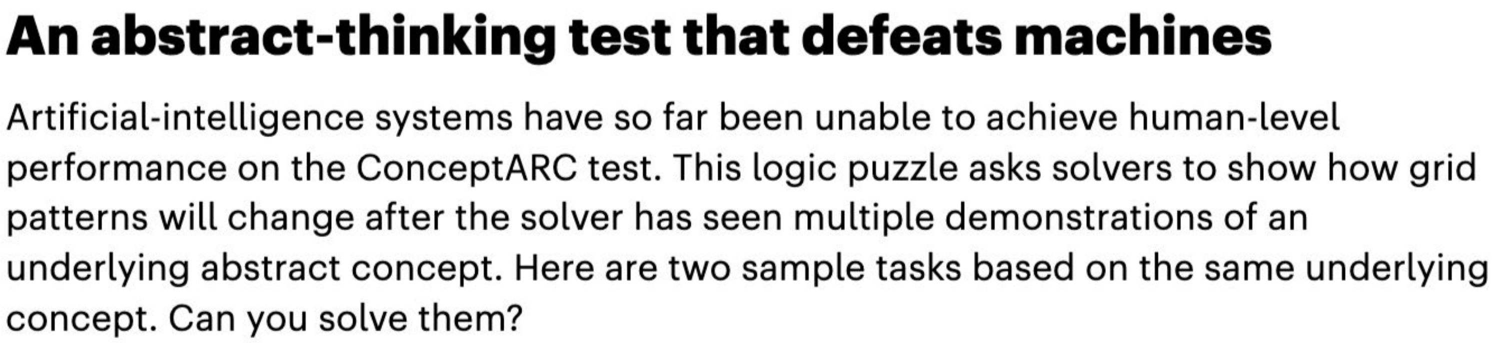
Achievements of the conference:

- Term artificial intelligence was coined
- Newell and Simon revealed the Logic Theorist Computer Program
    - First AI Program
    - Program proves mathematical Theorem
    - Conference found it to be a remarkable achievement

Conclusions of the Author

AI hasn’t managed to create general intelligence

- Artificial general intelligence (AGI) matches or surpasses human cognitive capabilities across a wide range of cognitive tasks
- Narrow AI matches human capabilities in a specific task

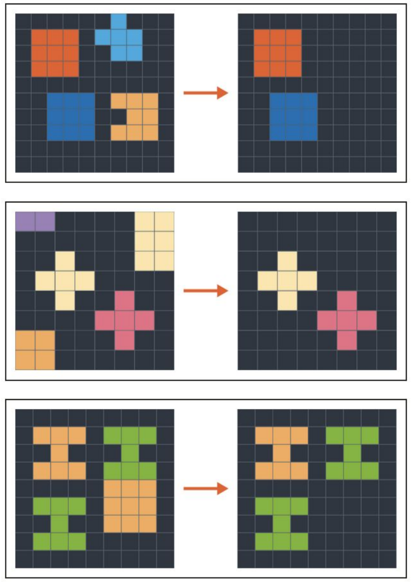


An abstract-thinking test that defeats machines

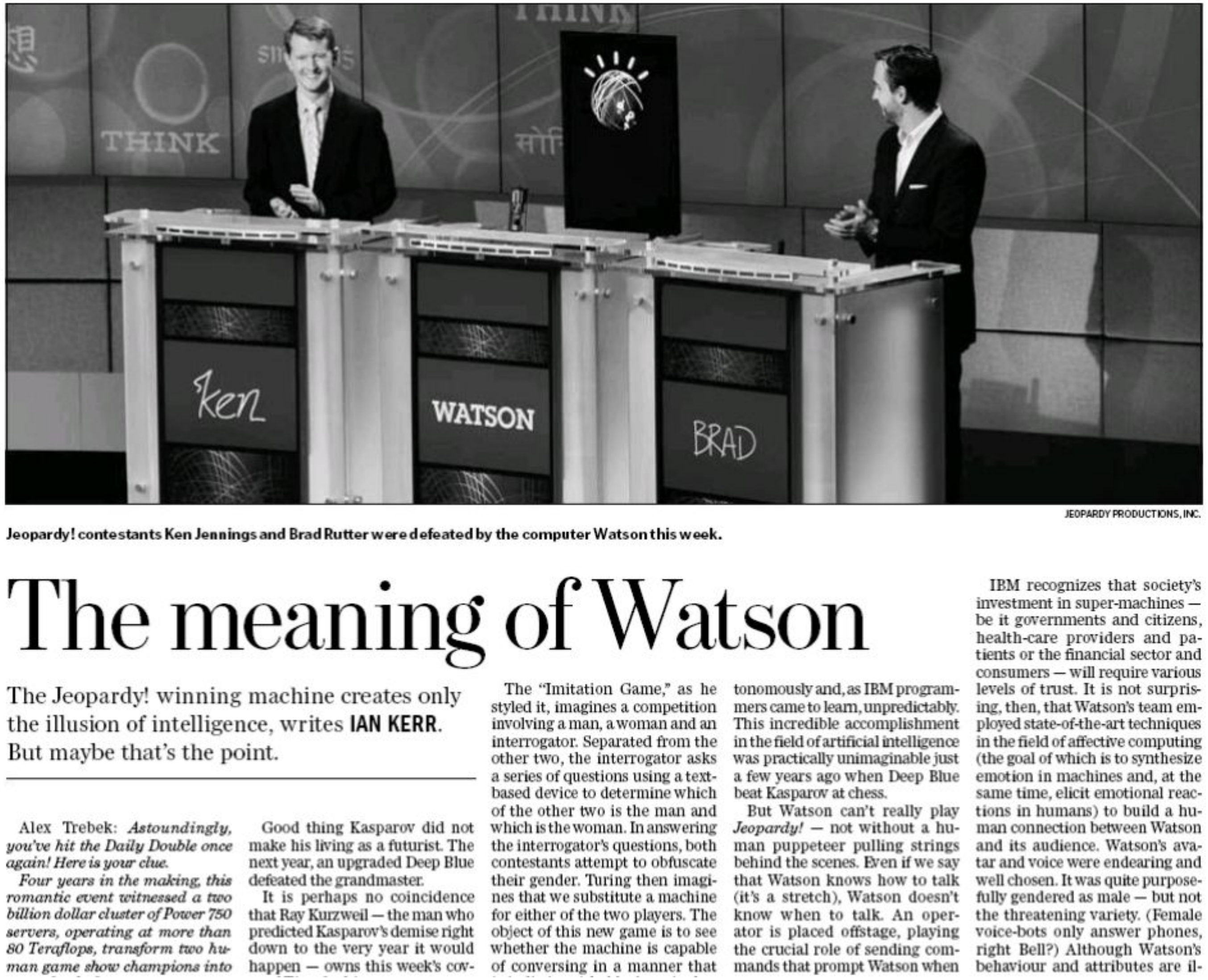
Artificial-intelligence systems have so far been unable to achieve human-level performance on the ConceptARC test. This logic puzzle asks solvers to show how grid patterns will change after the solver has seen multiple demonstrations of an underlying abstract concept. Here are two sample tasks based on the same underlying concept. Can you solve them?

Task A

Demonstration:



Group 1 - AI Milestones


THINK
Ken
WATSON
BRAD

JEOPARDY PRODUCTIONS, INC.

Jeopardy! contestants Ken Jennings and Brad Rutter were defeated by the computer Watson this week.

The meaning of Watson

The Jeopardy! winning machine creates only the illusion of intelligence, writes **IAN KERR**. But maybe that's the point.

Alex Trebek: *Astoundingly, you've hit the Daily Double once again! Here is your clue.*

Four years in the making, this romantic event witnessed a two billion dollar cluster of Power 750 servers, operating at more than 80 Teraflops, transform two human game show champions into

Good thing Kasparov did not make his living as a futurist. The next year, an upgraded Deep Blue defeated the grandmaster.

It is perhaps no coincidence that Ray Kurzweil — the man who predicted Kasparov's demise right down to the very year it would happen — owns this week's cov-

The "Imitation Game," as he styled it, imagines a competition involving a man, a woman and an interrogator. Separated from the other two, the interrogator asks a series of questions using a text-based device to determine which of the other two is the man and which is the woman. In answering the interrogator's questions, both contestants attempt to obfuscate their gender. Turing then imagines that we substitute a machine for either of the two players. The object of this new game is to see whether the machine is capable of conversing in a manner that

tonomously and, as IBM programmers came to learn, unpredictably. This incredible accomplishment in the field of artificial intelligence was practically unimaginable just a few years ago when Deep Blue beat Kasparov at chess.

But Watson can't really play *Jeopardy!* — not without a human puppeteer pulling strings behind the scenes. Even if we say that Watson knows how to talk (it's a stretch), Watson doesn't know when to talk. An operator is placed offstage, playing the crucial role of sending commands that prompt Watson when

IBM recognizes that society's investment in super-machines — be it governments and citizens, health-care providers and patients or the financial sector and consumers — will require various levels of trust. It is not surprising, then, that Watson's team employed state-of-the-art techniques in the field of affective computing (the goal of which is to synthesize emotion in machines and, at the same time, elicit emotional reactions in humans) to build a human connection between Watson and its audience. Watson's avatar and voice were endearing and well chosen. It was quite purposefully gendered as male — but not the threatening variety. (Female voice-bots only answer phones, right Bell?) Although Watson's behaviour and attributes are il-

IBM Watson & The Jeopardy! Challenge 2011

- IBM supercomputer named after the first CEO competed against two of the most successful players
- handled diverse question types across multiple domains
- processes complex information and masters factual questions
- limited by inability to handle dynamic questions or real-time conversation
- considered more impressive than Deep Blue's chess victory

Groundbreaking Victory in Natural Language Processing

Google DeepMind's AlphaGo 2016

- Defeated champion Lee Seedol in 4 out of 5 matches
- higher complexity in Go compared to chess
- however, while AlphaGo is a great Go player, it cannot understand a description of the rules in written English
- Highlights distinction between specialized vs. general intelligence

Historical Roots

- AI runs deep into the past, and has always had philosophy in its veins
- both fields are concerned with understanding intelligence, reasoning, and the nature of thought
- philosophy has provided the foundational ideas, questions, and frameworks that continue to shape AI
- philosophers have shaped the theoretical foundations of probability and its interpretation
- computer science emerged from logic and probability theory

Evolution of AI Thinking

Beyond Just Modern Computers

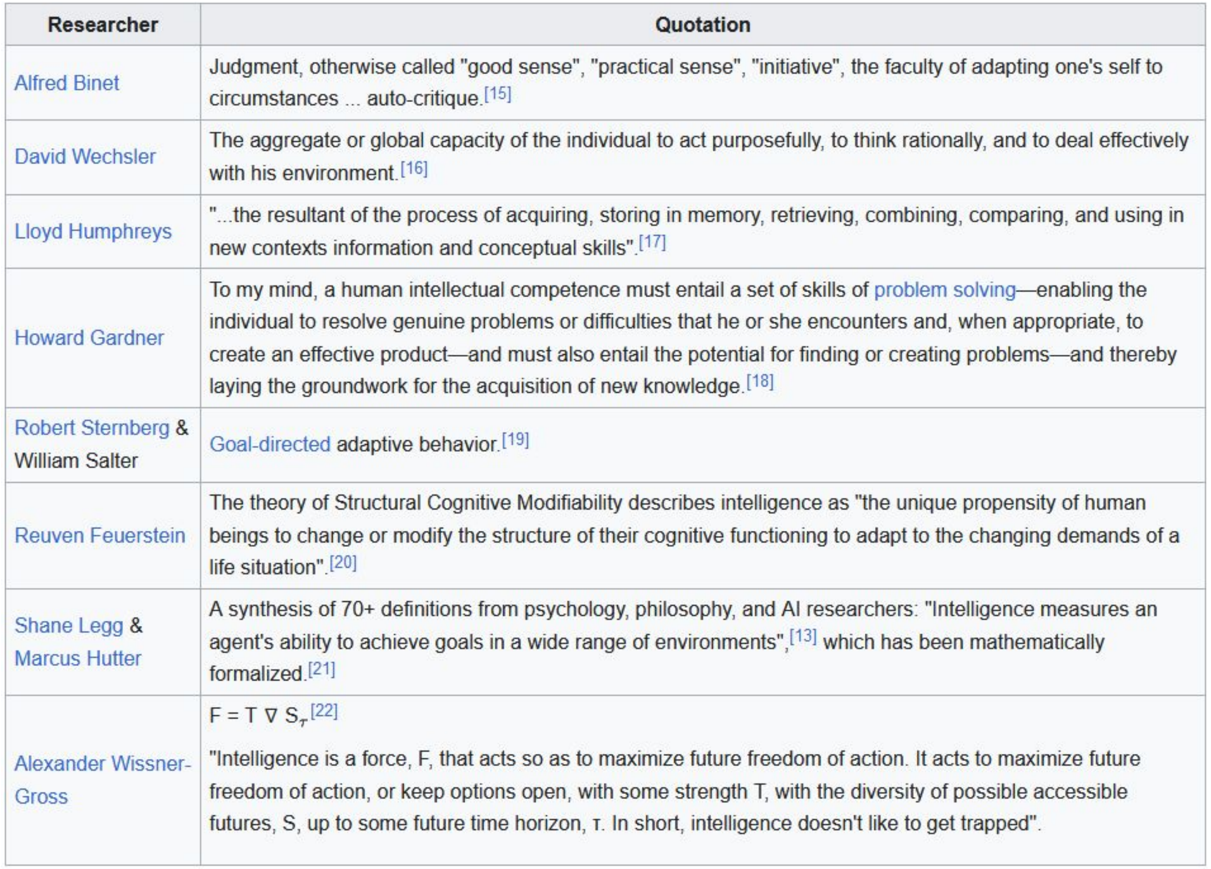
- AI's development preceded digital computers
- rooted in fundamental questions about reasoning and proof
- early logicians proposed all thinking could be expressed in formal logic
- contemporary AI combines ancient logical foundations with modern computational power

What *exactly* is AI?

AI lacks a universally accepted definition

Artificial: “*Made or produced by people rather then occurring naturally*”

Intelligence → Latin word <u>intellegere</u>: “*the acquirement, processing and storage of information*”



Researcher	Quotation
Alfred Binet	Judgment, otherwise called "good sense", "practical sense", "initiative", the faculty of adapting one's self to circumstances ... auto-critique.[15]
David Wechsler	The aggregate or global capacity of the individual to act purposefully, to think rationally, and to deal effectively with his environment.[16]
Lloyd Humphreys	"...the resultant of the process of acquiring, storing in memory, retrieving, combining, comparing, and using in new contexts information and conceptual skills".[17]
Howard Gardner	To my mind, a human intellectual competence must entail a set of skills of problem solving—enabling the individual to resolve genuine problems or difficulties that he or she encounters and, when appropriate, to create an effective product—and must also entail the potential for finding or creating problems—and thereby laying the groundwork for the acquisition of new knowledge.[18]
Robert Sternberg & William Salter	Goal-directed adaptive behavior.[19]
Reuven Feuerstein	The theory of Structural Cognitive Modifiability describes intelligence as "the unique propensity of human beings to change or modify the structure of their cognitive functioning to adapt to the changing demands of a life situation".[20]
Shane Legg & Marcus Hutter	A synthesis of 70+ definitions from psychology, philosophy, and AI researchers: "Intelligence measures an agent's ability to achieve goals in a wide range of environments",[13] which has been mathematically formalized.[21]
Alexander Wissner-Gross	$F = T \nabla S_\tau$.[22] "Intelligence is a force, F, that acts so as to maximize future freedom of action. It acts to maximize future freedom of action, or keep options open, with some strength T, with the diversity of possible accessible futures, S, up to some future time horizon, τ. In short, intelligence doesn't like to get trapped".

https://en.wikipedia.org/wiki/Intelligence

What *exactly* is AI?

Cambridge dictionary: “*the use or study of computer systems or machines that have some of the qualities that the human brain has, such as the ability to interpret and produce language in a way that seems human, recognize or create images, solve problems, and learn from data supplied to them*”

AI defined by goals

Russell and Norvig framework → AI: the Modern Approach (1995)

Two type of AI goals based on two dimensions:

1. Thinking vs. Acting

2. Human vs. Rational

The Four Categories:

1. **Thinking like Humans**: Replicating human thought processes.
2. **Thinking Rationally**: Logical reasoning and problem-solving.
3. **Acting like Humans**: Mimicking human behavior (e.g., Turing Test).
4. **Acting Rationally**: Intelligent agents that act based on rational principles.

Modern AI and the 4 categories

1. Systems That Think Like Humans

   Examples: Chatbots, sentiment analysis, creative AI (art/music).

   Movie: Her (2013)

2. Systems That Think Rationally

   Examples: Expert systems, theorem provers, planning tools.

   Movie: I, Robot (2004)

3. Systems That Act Like Humans

   Examples: Humanoid robots, virtual assistants, gaming AI.

   Movie: Ex Machina (2014)

4. Systems That Act Rationally

   Examples: Self-driving cars, recommendation systems, trading bots.

   Movie: The Matrix (1999)

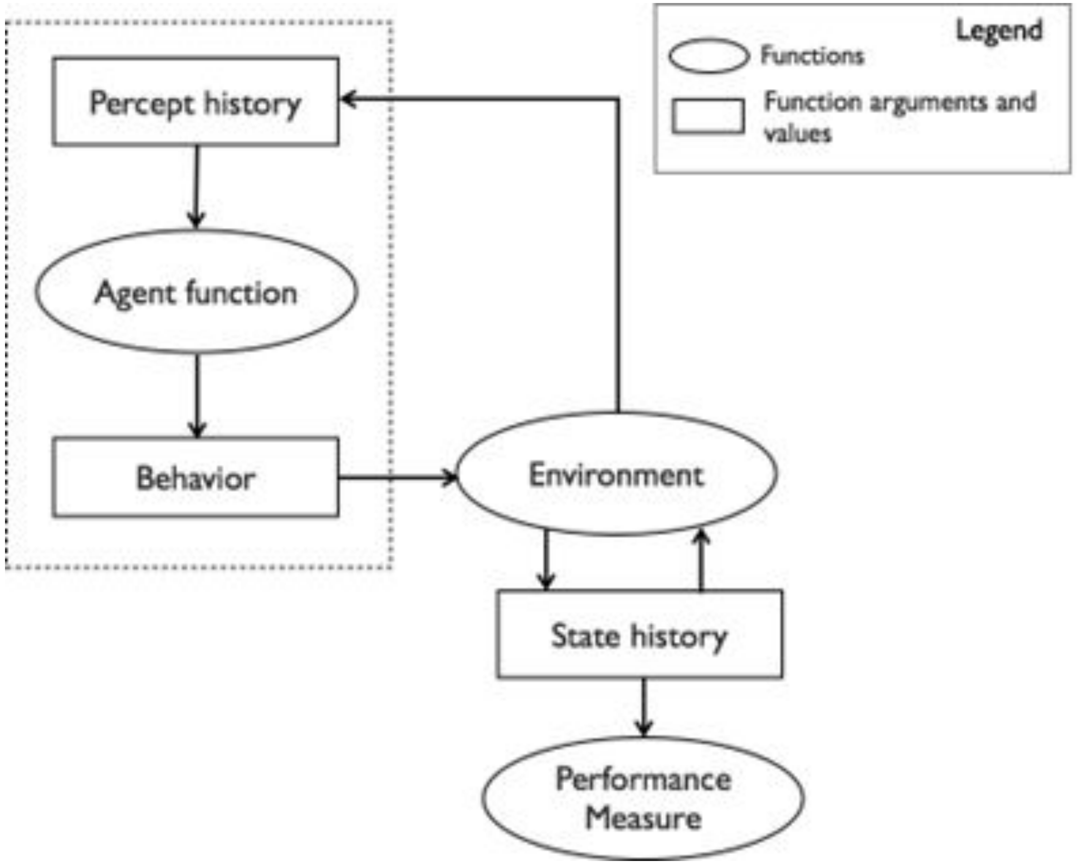
Conceptual model of intelligent agents

Russells work

Percept History → Agent Function → Behavior

Behavior → Environment → State History

State History → Performance Measure



Perfect rationality in AI

- A perfectly rational agent is defined as one that maximizes expected utility (V) in its environment (E).
- **Limitation :** Building a perfectly rational agent is often impractical due to computational limitations.

  For example : writing an algorithm that guarantees an invincible chess strategy is infeasible because the number of possible game states is extremely large, making it impossible to evaluate all possibilities within a reasonable amount of time.

- Instead, AI focuses on building *calculatively rational* agents

Calculative Rationality in AI

- Aim for optimal solutions within a reasonable duration. If the program is *executed infinitely fast*, it would result in perfectly rational behavior.

- **Chess Example**: calculatively rational agents search for good moves but limit their searches to ensure the game is played within reasonable timeframes. This means they find good (but not necessarily perfect) moves in limited time.

Russell’s Bounded Optimality

- **Machine Constraints**: AI systems face limitations in computational power, memory, and processing capabilities. Efficient performance is particularly challenging on limited hardware, such as embedded systems or mobile devices, where both memory and processing power are restricted.

- **Russell’s Bounded Optimality:** These agents are designed to be as optimal as possible within real-world machine and time constraints.

General Intelligence

- Designing agents for general intelligence is currently beyond reach. Most AI systems are designed for specific tasks, not for general intelligence.
- **Example**: In chess, the environment (E) is defined by the game’s rules and the board's state, whereas in Jeopardy!, E encompasses language comprehension and trivia knowledge across numerous domains. A generally intelligent agent would need a unified model capable of operating across all types of environments – an extremely challenging, currently unfeasible task.

Group 2 - Chapter 3

Group 2 - Chapter 3 - Approaches to AI (Nicolas)

3.1) The Intelligent Agent Continuum

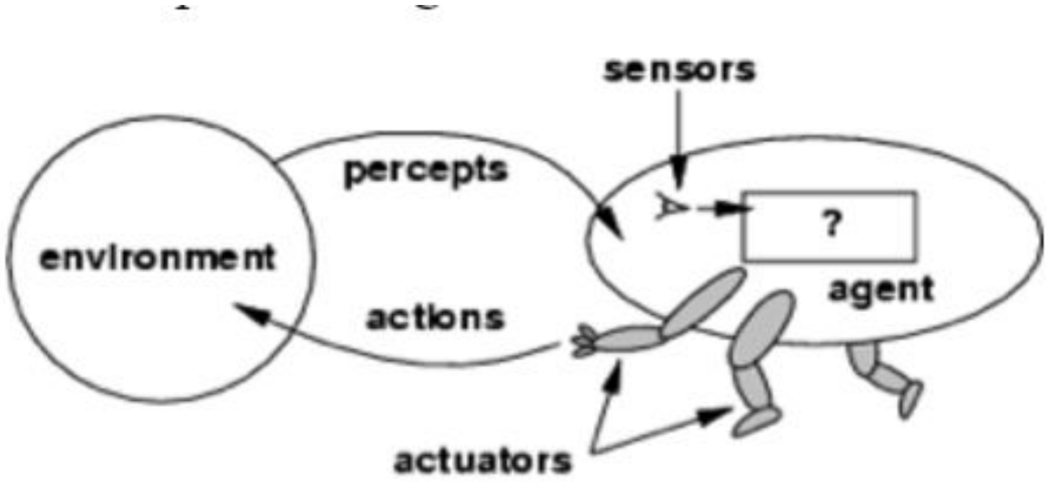
3.2) Logic-Based AI

3.3) Non-Logicist AI

3.4) AI Beyond the Clash of Paradigms

3.1) The intelligent Agent Continuum (Nicolas)

- Concept: Agents represent functions mapping percept sequences to actions



Bringsjord, Selmer and Naveen Sundar Govindarajulu, "Artificial Intelligence", *The Stanford Encyclopedia of Philosophy* (Fall 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2024/entries/artificial-intelligence/>

- Agent Evolution

➢ **Simple reflex**: react to direct stimuli

➢ **Model-based:** internal models to predict and act beyond immediate perception

➢ **Advanced**: Incorporate reasoning, learning, and communication

3.1) The intelligent Agent Continuum (Nicolas)

Unsolved Challenges in AI

- **Learning by Reading:**

➢ Current models focus on function-based learning (e.g., mapping inputs to outputs).

➢ Reading as a knowledge acquisition process remains a struggling challenge for AI compared to humans.

➢ Example: Understanding "classroom" from pixels or actions like "HavingACupOfTea" requires high-level abstraction.

- **Consciousness and Creativity:**

➢ Subjective experiences and creativity are not addressed in mainstream AI research.

3.1) The intelligent Agent Continuum (Nicolas)

Philosophical and Practical Implications

- **Central Questions:**

➢ Can AI replicate the vast knowledge acquisition capabilities of humans (e.g., through reading)?

➢ What about subjective consciousness and creativity?

- **Engineering Focus**: AI emphasizes function over human-like processes (e.g., natural language understanding without full linguistic mastery).

3.1) The intelligent Agent Continuum (Nicolas)

Textbook Framework and Limitations

- **Common Themes:**

➢ Progression from simple to complex agents is a shared structure across AI textbooks.

➢ Focuses on engineering intelligence rather than mirroring human cognition.

- **Gaps in Coverage**

➢ Phenomenal consciousness, creativity, and advanced abstraction largely ignored.

3.2) Logic-Based AI: Some Surgical Points (Nicolas)

Introduction to Reasoning in AI

- **Monotonic vs. Nonmonotonic Logic:**

➢ Monotonic logic: Adding new information doesn't invalidate prior inferences.

- **Nonmonotonic logic:** New information can invalidate previous inferences.

➢ Example: Tweety is a bird → Tweety can fly; Tweety is a penguin → Tweety cannot fly.

- **Application:** Captures real-world reasoning where beliefs can change.

3.2) Logic-Based AI: Some Surgical Points (Nicolas)

Logicist AI Framework

- Concept: Use logical systems (e.g., first-order logic) to build intelligent agents.

- Agent’s Life Cycle:
  - ➢ Sense: Perceives the environment.
  - ➢ Adjust: Updates the knowledge base using reasoning techniques.
  - ➢ Act: Performs actions based on goals.

- Reasoning Modes: Deductive, inductive, abductive.

3.2) Logic-Based AI: Some Surgical Points (Nicolas)

Challenges and Future Directions

- **Scalability:**

Reasoning speed needs to match real-world demands (e.g., robotics).

- **Broad Cognition:**

Extending logical reasoning to tasks like motor control and perception.

- **Research Opportunities:**

Advances in description logics for specialized domains (e.g., biomedicine).

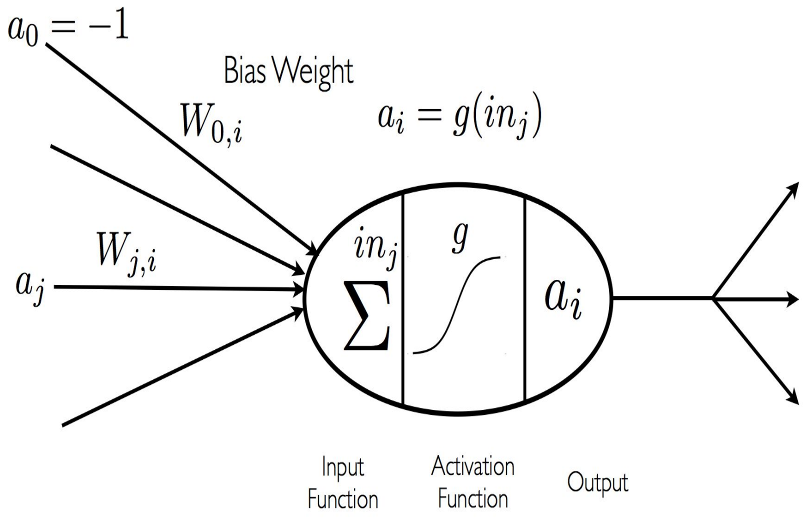
Further development in encoding and hybrid reasoning systems.

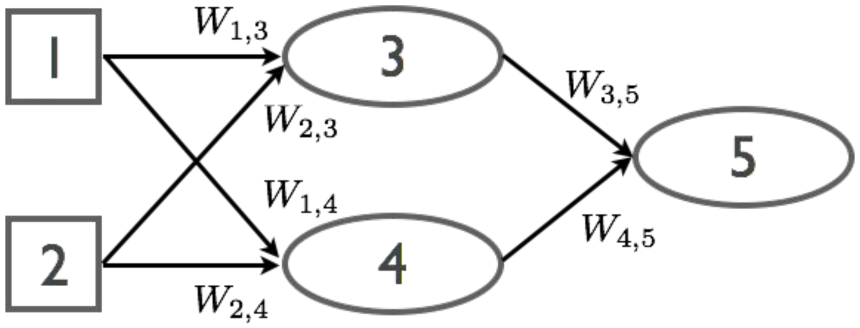
3.3) Non Logicist AI (Moataz)

Non-logicist AI is an approach that diverges from logicist AI by avoiding reliance on formal logical systems or declarative propositions to represent knowledge. Instead, it utilizes alternative formalisms that fall into two main categories:

1. Symbolic but Non-Logicist Approaches: These include methods like semantic networks, conceptual dependency schemes (Schank, 1972), and frame-based systems. These approaches organize knowledge in structured yet non-logical ways to improve readability and usability, but they lack the rigor of formal logic.
2. Neurocomputational Approaches: These rely on artificial neural networks (ANNs), which are inspired by biological neural systems. ANNs are composed of:
    - Nodes (neurons): Represent individual processing units.
    - Weighted connections (dendrites): Links between nodes, with each having a numeric weight.
    - Layers: Organized into input, hidden, and output layers, enabling hierarchical data processing.

Neural networks, while theoretically capable of universal computation, are primarily used for building learning systems. Their architecture allows them to learn from data by adjusting weights based on input, enabling them to recognize patterns and generalize across tasks. This flexibility makes ANNs particularly suited for tasks that cannot be easily structured using traditional logic.



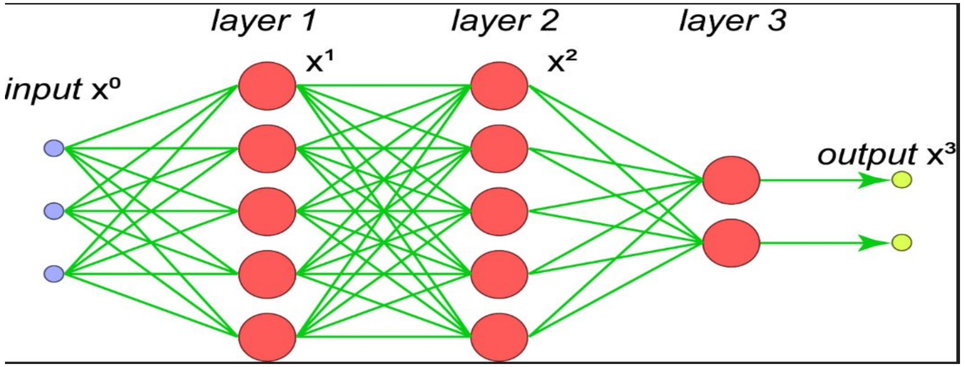


3.3) Non Logicist AI (Moataz)

Artificial Neural Networks (ANNs)

1. Structure and Components:
   - Neurons: Basic computational units that aggregate inputs and produce outputs based on an activation function.
   - Connections: Weighted links between neurons simulate synapses in the brain.
   - Layers: Include input, hidden, and output layers, with hidden layers enabling abstraction of data.
2. Computational Process:
   - Input Function: Aggregates inputs (sum of weighted activations).
   - Activation Function: Processes the aggregated input to determine the neuron's output.
   - Common activation functions:Step Function: and Sigmoid/Tanh Functions: Nonlinear outputs
3. Training:
   - Weights are updated iteratively using algorithms like backpropagation.
   - Networks learn from labeled data (supervised learning) or unlabeled data (unsupervised pre-training).

4. Early Challenges:
   - Training deep, multi-layered networks was computationally expensive and inefficient.
   - Limited to small-scale problems due to hardware constraints.



3.3) Non Logicist AI (Moataz)

Updated Summary: Non-Logicist AI with Today's Technology

Non-logicist AI, propelled by deep learning and hybrid methods, now addresses challenges previously unattainable by early AI paradigms. It continues to redefine artificial intelligence, bridging the gap between data-driven adaptability and structured reasoning to solve real-world problems at scale.

Modern Innovations in Neural Networks:

- Hardware Advances: GPUs and TPUs accelerated computations, enabling large-scale training.
- Deep Learning: Architectures like convolutional neural networks (CNNs) for vision and transformers (e.g., GPT models) for language tasks enable state-of-the-art performance.
- Pre-Training Techniques: Unsupervised pre-training allows layers to progressively learn from raw data (e.g., detecting edges, combining edges into features like eyes or faces).

Comparison with Early Non-Logicist AI

- Early systems struggled with training inefficiencies and lacked scalability.
- Today's systems leverage distributed computing and optimized algorithms to handle massive datasets and complex tasks.
- Neural networks now adapt dynamically, across tasks and domains, enabled by advances in transfer learning and multi-modal architectures.

3.3) Non Logicist AI (Moataz)

Updated Summary: Non-Logicist AI with Today’s Technology

Applications of Today’s Non-Logicist AI

1. Handwriting Recognition:
   - Benchmark problem solved effectively by neural networks trained on labeled datasets (e.g., MNIST database).
2. Image Recognition:
   - Layers process data hierarchically, enabling feature detection and object classification.
3. Natural Language Processing:
   - Neural networks underpin systems like translation tools and conversational AI. systems like ChatGPT handle tasks from summarization to creative writing.
4. Real-World Tasks:
   - Autonomous driving uses neural networks for vision, prediction, and decision-making.
5. Generative AI:

   - Models generate new data—images, text, or music , Systems like DALL-E and ChatGPT exemplify the ability of neural networks to create realistic images and coherent text.

6. Hybrid Problem-Solving:

   - Neurosymbolic systems integrate structured knowledge with adaptive learning for fraud detection and scientific discovery., Used in medical diagnosis, autonomous systems, and knowledge graph reasoning.

3.4) AI Beyond Clash Paradigms (Moataz)

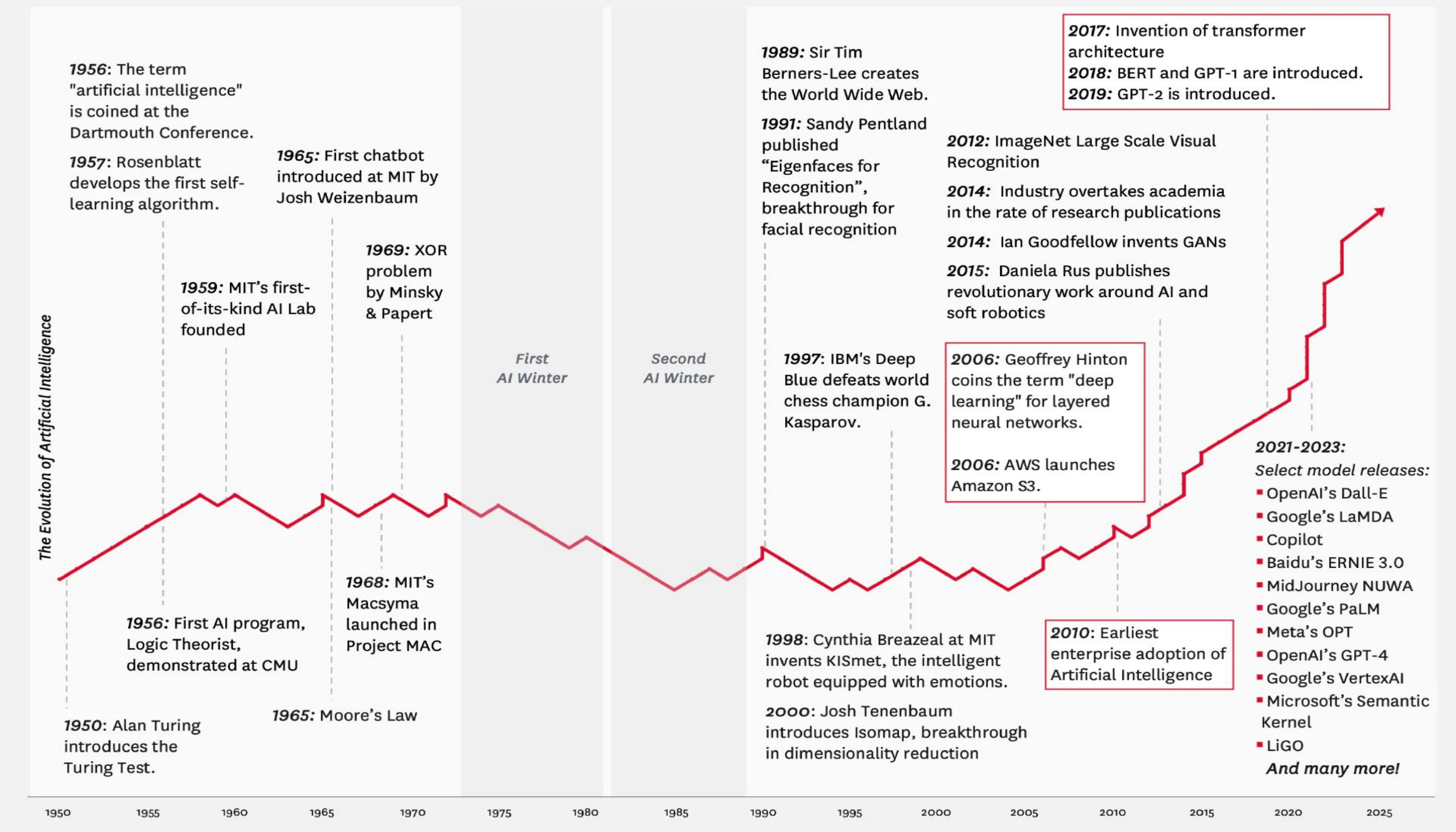
Updated Summary: AI Beyond the Clash of Paradigms with Today's Technology

Modern AI increasingly relies on hybrid systems that integrate logicist, probabilistic, and neurocomputational paradigms to address complex, real-world challenges. Today's technology enhances these systems with unprecedented scale, speed, and adaptability.

Key Examples of Modern Hybrid Systems

1. IBM Watson and Advanced Q&A Systems:
   - Evolved versions integrate large language models (LLMs) like GPT-4 for deeper natural language understanding.
   - Combine symbolic reasoning with machine learning for applications in legal analysis, healthcare, and enterprise knowledge systems.
2. DeepMind's AlphaZero and Generalized Game AI:
   - Successors to AlphaGo extend hybrid paradigms using reinforcement learning and neural networks, achieving superhuman performance in complex games like chess, Go, and StarCraft.
   - Combine planning algorithms with real-time learning for dynamic decision-making.
3. OpenAI's GPT-4 and Codex:
   - Blend symbolic reasoning (code and structure) with neural models for language and programming tasks.
   - Enable real-time problem-solving, such as generating software code or assisting in scientific research.

4 The Explosive Growth of AI



1950: Alan Turing introduces the Turing Test.

1956: The term "artificial intelligence" is coined at the Dartmouth Conference.

1956: First AI program, Logic Theorist, demonstrated at CMU

1957: Rosenblatt develops the first self-learning algorithm.

1959: MIT's first-of-its-kind AI Lab founded

1965: First chatbot introduced at MIT by Josh Weizenbaum

1965: Moore's Law

1968: MIT's Macsyma launched in Project MAC

1969: XOR problem by Minsky & Papert

1989: Sir Tim Berners-Lee creates the World Wide Web.

1991: Sandy Pentland published "Eigenfaces for Recognition", breakthrough for facial recognition

1997: IBM's Deep Blue defeats world chess champion G. Kasparov.

1998: Cynthia Breazeal at MIT invents KISmet, the intelligent robot equipped with emotions.

2000: Josh Tenenbaum introduces Isomap, breakthrough in dimensionality reduction

2006: Geoffrey Hinton coins the term "deep learning" for layered neural networks.

2006: AWS launches Amazon S3.

2010: Earliest enterprise adoption of Artificial Intelligence

2012: ImageNet Large Scale Visual Recognition

2014: Industry overtakes academia in the rate of research publications

2014: Ian Goodfellow invents GANs

2015: Daniela Rus publishes revolutionary work around AI and soft robotics

2017: Invention of transformer architecture
2018: BERT and GPT-1 are introduced.
2019: GPT-2 is introduced.

2021-2023:
Select model releases:
- OpenAI's Dall-E
- Google's LaMDA
- Copilot
- Baidu's ERNIE 3.0
- MidJourney NUWA
- Google's PaLM
- Meta's OPT
- OpenAI's GPT-4
- Google's VertexAI
- Microsoft's Semantic Kernel
- LiGO

And many more!

https://glasswing.vc/blog/thinking-corner/the-history-of-artificial-intelligence/

4.1 Bloom in ML

A huge part of AI’s growth is due to invention of new algorithms of **machine learning (systems that improve performance based on examples or experience.)**.

- **Supervised Learning:**
  - Learns a function using labeled data pairs.
  - Goal is to minimize error between the predicted function and true function.
  - Dominates ML applications, such as labeling images or detecting spam.
- **Unsupervised Learning:**
  - Discovers patterns in raw data without labeled outputs.
  - Examples: Data mining and Google's PageRank algorithm.
- **Reinforcement Learning:**
  - Machines learn through trial and error, using feedback (rewards/punishments).
  - Common in building game-playing agents (e.g., learning optimal chess moves).

https://glasswing.vc/blog/thinking-corner/the-history-of-artificial-intelligence/

4.1 Bloom in ML

Machine-learning algorithms are more and more used in all stages of the scientific process.

- CERN’s particle accelerators: filter and analyze petabytes of data to identify meaningful events, such as Higgs Boson discoveries.
- Genomics: analyze genomic sequences, predict gene structures, and identify genetic variants.
- Astronomy: process images from telescopes to classify galaxies based on shape, size, and other features.

Big data: applying techniques derived from AI to large volumes of data

- Explosion in data that does not have any explicit semantics attached to it and is not easily machine-processable (images, text, video (as opposed to carefully curated data in a knowledge- or data-base)).

https://glasswing.vc/blog/thinking-corner/the-history-of-artificial-intelligence/

4.2 - The Resurgence of Neurocomputational Techniques

Central dogma of AI: **“What the brain does may be thought of at some level as a kind of computation”**

Aspect	**Symbolic AI**	**Connectionist AI**
Inspiration	Logical reasoning	Brain's neural structure
Approach	Rule-based	Data-driven
Strengths	Interpretability, reasoning	Pattern recognition, adaptability
Limitations	Inflexible, hard to scale	Black-box nature, data-intensive
Applications	Knowledge bases, expert systems	Image recognition, speech processing

Both approaches are complementary, and modern AI often integrates symbolic reasoning with connectionist techniques for more robust and versatile solutions

https://glasswing.vc/blog/thinking-corner/the-history-of-artificial-intelligence/

4.2 - The Resurgence of Neurocomputational Techniques

Feature vector representation function is a transformation of the input into a format that filters out irrelevant information in the input and keep only information useful for the task

- **Feature Engineering:** Previously, human experts manually crafted representations to transform raw data into useful inputs for learning, a labor-intensive process known as a "black art."
- **Deep Learning Breakthrough:** Deep neural networks, with multiple hidden layers, automatically learn feature representations, reducing the need for manual engineering.
    - Example: Recognizing facial features while disregarding irrelevant factors like lighting.

https://glasswing.vc/blog/thinking-corner/the-history-of-artificial-intelligence/

4.2 - The Resurgence of Neurocomputational Techniques

Deep learning can safely be regarded as the study of models that either involve a **greater amount of composition of learned functions or learned concepts** than traditional machine learning does. (Bengio et al. 2015, Chapter 1)

Recent innovations have made deep learning more efficient and feasible, enabling its widespread application. State-of-the-art results in:

- image recognition (given an image containing various objects, label the objects from a given set of labels)
- speech recognition (from audio input, generate a textual representation)
- the analysis of data from particle accelerators

https://glasswing.vc/blog/thinking-corner/the-history-of-artificial-intelligence/

4.2 - The Resurgence of Neurocomputational Techniques

Two significant challenges of deep learning:

Minor:

- **Still requires human expertise for architectural design and hyperparameter tuning.** This process is often guided by intuition and experience, lacking systematic methodologies.

Major:

- **Vulnerability to Adversarial Inputs** (intentionally altered data that appear normal to humans but cause neural networks to make incorrect predictions).
  - Can lead to high-confidence misclassifications, undermining the reliability of AI systems (a slight modification to an image can cause a model to misidentify objects).
  - The persistence of adversarial inputs across different models and datasets raises concerns about deploying AI in safety-critical applications

https://glasswing.vc/blog/thinking-corner/the-history-of-artificial-intelligence/

4.3 - The Resurgence of Probabilistic Techniques

- 2nd element to the explosive growth of AI:
  Non-neurocomputational probabilistic methods

- AI uses a different than the standard approach stemming from a logic and technical philosophy

- *Random Variables:*
  Fundamental proposition, allowing initial uncertainty for an agent

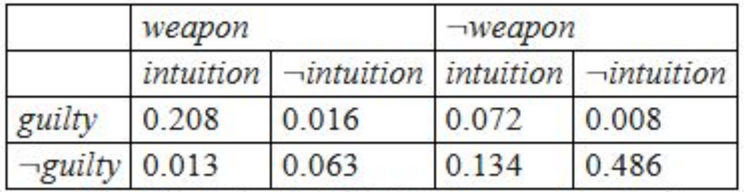
4.3 - The Resurgence of Probabilistic Techniques

- The Kalmogorov Axioms apply and yet base the probabilistic approach in logic

- ***Bayesian networks*** as a solution for the 2 problems:
  1. Processing large quantities of data
  2. Only propositional expressivity

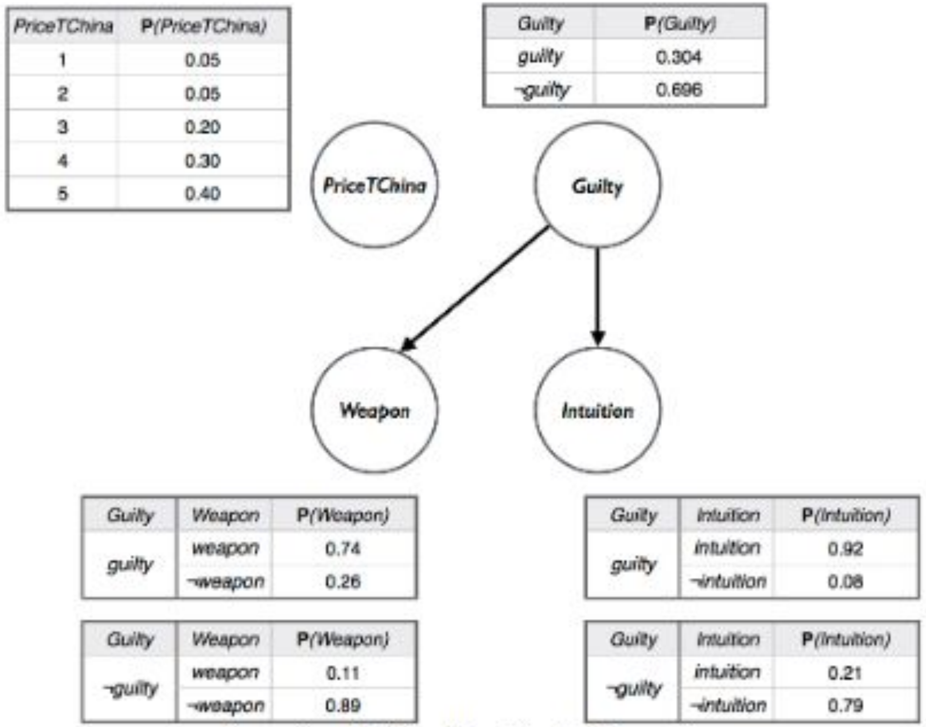
- One application of this model could be to ask for the probability of an hypothesis H_x being true

4.3 - The Resurgence of Probabilistic Techniques



	weapon		*¬weapon*	
	intuition	*¬intuition*	*intuition*	*¬intuition*
guilty	0.208	0.016	0.072	0.008
¬guilty	0.013	0.063	0.134	0.486

Joint Distribution for $\mathbf{P}(Guilty, Weapon, Intuition)$



PriceTChina	P(PriceTChina)
1	0.05
2	0.05
3	0.20
4	0.30
5	0.40

Guilty	P(Guilty)
guilty	0.304
¬guilty	0.696

Guilty	Weapon	P(Weapon)
guilty	weapon	0.74
	¬weapon	0.26

Guilty	Intuition	P(Intuition)
guilty	intuition	0.92
	¬intuition	0.08

Guilty	Weapon	P(Weapon)
¬guilty	weapon	0.11
	¬weapon	0.89

Guilty	Intuition	P(Intuition)
¬guilty	intuition	0.21
	¬intuition	0.79

Bayesian Net for the Murder Example

Plugging in the values from the Bayes net into the equation gives us:

$$P\left(guilty, \neg weapon, \neg intuition, PriceTChina = 5\right)$$
$$= 0.304 \times 0.26 \times 0.08 \times 0.40$$
$$= 0.0025$$

Group 2 - Chapter 5 - AI in the Wild

Welcome to ChatGPT

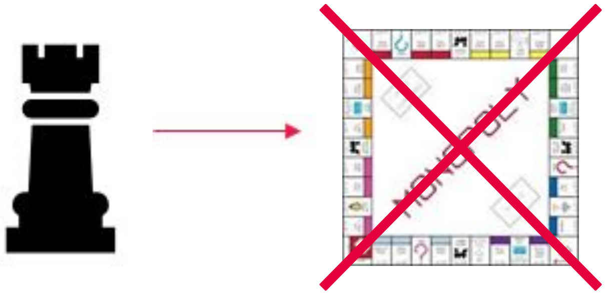
your OpenAI accoun

Hi

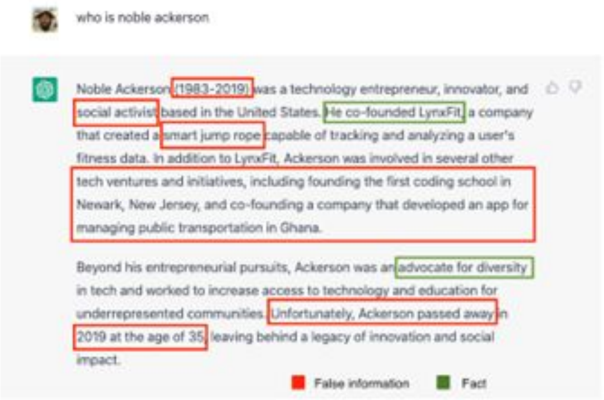
Hello! How can I help you today? Is there something you need help with or would like to learn more about? I'm here to assist you with any ques you may have.

Source: Absolutely AI (2023); Eldagsen (2023); GE Healthcare (2019); Passingham and Laughlin (2023); Sullivan (2023)

Group 2 - Chapter 5 - Limits of AI



Group 2 - Chapter 5 - Limits of AI

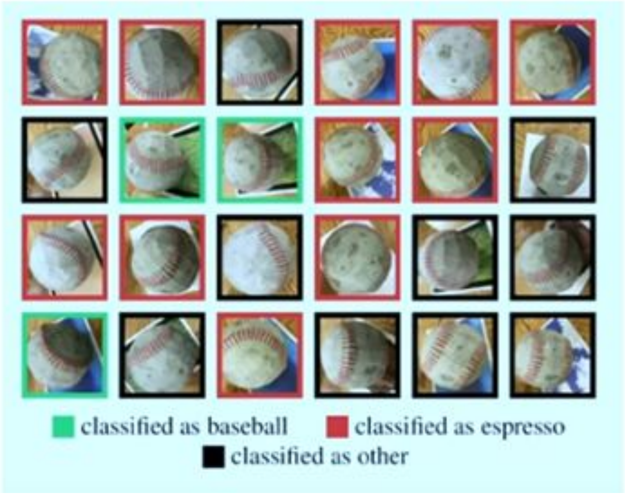


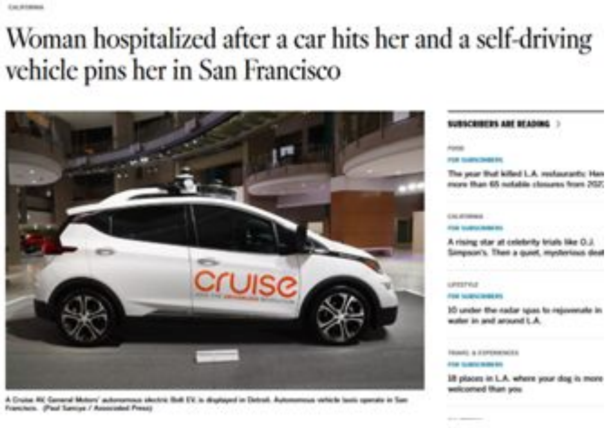
who is noble ackerson

Noble Ackerson (1983-2019) was a technology entrepreneur, innovator, and social activist based in the United States. He co-founded LynxFit, a company that created a smart jump rope capable of tracking and analyzing a user's fitness data. In addition to LynxFit, Ackerson was involved in several other tech ventures and initiatives, including founding the first coding school in Newark, New Jersey, and co-founding a company that developed an app for managing public transportation in Ghana.

Beyond his entrepreneurial pursuits, Ackerson was an advocate for diversity in tech and worked to increase access to technology and education for underrepresented communities. Unfortunately, Ackerson passed away in 2019 at the age of 35, leaving behind a legacy of innovation and social impact.

🟥 False information 🟩 Fact





CALIFORNIA

Woman hospitalized after a car hits her and a self-driving vehicle pins her in San Francisco

A Cruise AV, General Motors' autonomous electric Bolt EV, is displayed in Detroit. Autonomous vehicle taxis operate in San Francisco. (Paul Sancya / Associated Press)

SUBSCRIBERS ARE READING >

FOOD

FOR SUBSCRIBERS

The year that killed L.A. restaurants: Here are more than 65 notable closures from 202

CALIFORNIA

FOR SUBSCRIBERS

A rising star at celebrity trials like O.J. Simpson's. Then a quiet, mysterious deat

LIFESTYLE

FOR SUBSCRIBERS

10 under the radar spas to rejuvenate in water in and around L.A.

TRAVEL & EXPERIENCES

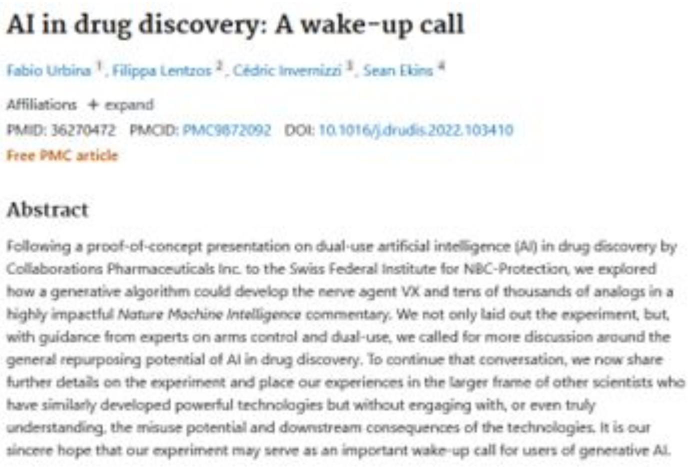
FOR SUBSCRIBERS

18 places in L.A. where your dog is more welcomed than you

Source: Ackerson (2023); Betzendahl (2021); Sancya (2023)

Group 2 - Chapter 5 - Dark Side of AI

Dual use problem



AI in drug discovery: A wake-up call

Fabio Urbina [1], Filippa Lentzos [2], Cédric Invernizzi [3], Sean Ekins [4]

Affiliations + expand

PMID: 36270472 PMCID: PMC9872092 DOI: 10.1016/j.drudis.2022.103410

Free PMC article

Abstract

Following a proof-of-concept presentation on dual-use artificial intelligence (AI) in drug discovery by Collaborations Pharmaceuticals Inc. to the Swiss Federal Institute for NBC-Protection, we explored how a generative algorithm could develop the nerve agent VX and tens of thousands of analogs in a highly impactful *Nature Machine Intelligence* commentary. We not only laid out the experiment, but, with guidance from experts on arms control and dual-use, we called for more discussion around the general repurposing potential of AI in drug discovery. To continue that conversation, we now share further details on the experiment and place our experiences in the larger frame of other scientists who have similarly developed powerful technologies but without engaging with, or even truly understanding, the misuse potential and downstream consequences of the technologies. It is our sincere hope that our experiment may serve as an important wake-up call for users of generative AI.

Source: art_is_2_inspire (2023); Urbina et al. (2022)

Chapter 6: Moral AI

Introduction to Moral AI

Computer Ethics vs. Robot Ethics

- Computer Ethics focus on human behaviour in tech contexts
    - How do we humans behave when we use computers/internet/AI
- Robot ethics (Moral AI, machine ethics): Focus on autonomous machines making moral decisions.
    - morality of machine itself: Machine makes a decision with moral weight and choosing between live and death

Challenges in Moral AI

- Robots making ethically significant decisions (e.g., lethal actions).
- The need for ethical reasoning in robots with decision-making power.
    - Challenges for developing a system that ensures the robot acts ethically by its own decision. (no human interaction)

Approaches to Building Moral Machines

Ethical AI Spectrum

- Includes both lethal and non-lethal situations (e.g., lying machines).
    - also ethically questionable actions (Trust, Honesty, Social impact of robots)

Framework for Moral Machines

- Machines need a moral code to guide decision-making.
    - to align with human moral standards
- Moral code can be input directly or inferred from basic laws.
- Challenge:
    - Make the Robot understand and follow the moral code in unpredictable situations

Deontic Logics in Moral AI

Deontic Logic for Ethical Reasoning

- Formal logic concerned with moral codes (obligations, permissions, prohibitions).
  - what actions are required, allowed, forbidden according to Moral Code
- Used to ensure robots follow a moral code in all circumstances.
  - MUST not harm human being, but MAY CHOOSE to help someone

Formal-Logic Verification

- Robots’ actions can be verified before real-world deployment using deontic logic.
- Ensures robots behave ethically in infinite scenarios. (as many scenarios as you are able to test)

The Future of Moral AI

Early Work in Ethical Machines

- Example: Pereira and Saptawijaya's work on contractualism.
    a. Moral Theory states: actions are right if they are justifiable to others based on social contract.

Growing Importance of Robot Ethics

- As robots gain autonomy, ethical considerations will increase in importance.
- Ethical standards for robots may differ from those applied to humans.

Chapter 7: Philosophical AI

Introduction to Philosophical AI

- What is Philosophical AI?

Philosophical AI applies ideas and methods from philosophy to AI development. It’s not about debating AI’s impact or ethics (that’s Philosophy of AI), but about solving philosophical problems in ways that can be implemented in AI systems.

- Key Feature: Philosophical reasoning leads to technical implementations.

For example, solving a paradox through logic and creating a program that can handle such paradoxes in practice.

- Distinction:

Philosophy of AI: Explores the implications of AI (e.g., can machines think? Is AI ethical?).

Philosophical AI: Focuses on using philosophy to shape AI systems.

Daniel Dennett’s View on AI and Philosophy

- Who is Daniel Dennett?

A philosopher known for exploring consciousness, cognition, and the relationship between AI and philosophy.

- His Claim: AI is philosophy in action, particularly cognitive psychology, as it seeks to answer the question:

- How is knowledge possible?

Top-Down Approach: AI should not just mimic biological systems (like neural networks do) but instead design abstract algorithms that replicate high-level cognitive processes. Top-Down means starting with abstract, overarching principles and applying them, rather than building from smaller, simpler components.

Critique of Dennett’s View

- Objection to Dennett’s Claim:

Critics argue that by focusing only on computational mechanisms, AI limits itself to a specific type of intelligence.

- Mechanistic Limitations:
- AI assumes intelligence can be reduced to mechanisms (like algorithms and computations).
- This excludes broader perspectives, like those from vitalists (who believe life involves more than physical processes) or dualists (who see the mind as separate from the body).
- Dennett’s Defense:
- AI’s mechanistic approach aligns with psychology and relies on Church’s Thesis.

What is Church’s Thesis?

Church’s Thesis (also known as the Church-Turing Thesis) is a foundational concept in the theory of computation. It is named after mathematicians Alonzo Church and Alan Turing, who independently proposed it in the 1930s. The thesis makes a claim about the nature of computability, specifically the kinds of functions that can be computed.

Church’s Thesis Explained:
Church’s Thesis asserts that a function is effectively calculable (or computable) if and only if it can be computed by a Turing machine. In other words, if a problem can be solved by an algorithm or mechanical procedure, it can be solved by a Turing machine.

Key Concepts:

1. Turing Machine:

A theoretical machine introduced by Alan Turing in 1936 that manipulates symbols on an infinite tape according to a set of rules. Turing machines are used to model what it means to compute a function or solve a problem. The power of a Turing machine lies in its simplicity and ability to simulate any computation that can be done by any other mechanical or algorithmic process.

2. Effective Computability:

A function or problem is considered effectively computable if it can be solved by an algorithm that can be executed step-by-step by a human or machine. This type of computation corresponds to the idea of solving a problem with a clear, finite procedure.

3. Computability and Functions:

Church’s Thesis suggests that any function that can be computed by a well-defined algorithm or mechanical process can be computed by a Turing machine. In other words, the set of computable functions is the same as the set of functions that can be computed by a Turing machine.

The Scope of AI vs. Philosophy

- Philosophy’s Breadth:

Philosophy and psychology explore intelligence broadly, including non-mechanistic or non-computational aspects.

- AI’s Narrow Focus:

AI is about creating computational artifacts (programs, systems) to model or replicate intelligence.

- Risk of Misidentification:
- Calling AI “philosophy” might confine philosophy to computational perspectives, ignoring broader insights.

AI’s Practical Goals vs. Abstract Philosophy

- AI’s Main Goal:

Build systems that are useful, efficient, and profitable.

Examples: Recommendation systems, chatbots, self-driving cars.

- Philosophical AI’s Goal:

Integrate philosophical principles into AI to solve abstract or conceptual problems.

Example: Programming a system to handle ethical dilemmas using philosophical reasoning.

Case Study: The OSCAR Project, which merges philosophy and AI.

John Pollock’s OSCAR Project

- What is OSCAR?

A reasoning system developed by philosopher John Pollock. It addresses problems like decision-making and reasoning using both philosophical principles and AI techniques.

- Why Important?

Shows how philosophy can directly inform AI development.

Published in leading AI journals, proving its technical and practical relevance.

Takeaway: Philosophical thinking can enhance AI by providing deeper and conceptual frameworks for solving problems.

Broader Connections Between AI and Philosophy

1. Epistemic Logic: A branch of philosophy exploring knowledge and belief; applied in AI for reasoning systems.
2. Inductive Learning: Philosophical studies on how we learn new ideas help improve AI methods like Inductive Logic Programming, which creates rules based on data.

Philosophical Nature of AI Textbooks

- Textbooks like AIMA (Artificial Intelligence: A Modern Approach) discuss the philosophical underpinnings of AI.
- They explore questions like: What is intelligence? How can it be built?
- Integration of Philosophy: AI education incorporates philosophical questions to guide technical approaches.

Conclusion

- Philosophical AI: A powerful blend of philosophy and technical engineering that enhances AI’s capabilities and depth.
- Distinction: AI is not philosophy, but philosophical concepts are integral to AI’s development.
- Future Directions: Continuing to explore the philosophical dimensions of intelligence will broaden AI’s scope and potential applications.

Chapter 8.1 Strong vs. Weak AI

Weak AI

- **Goal:** Build machines that **simulate** human behavior and pass the **Total Turing Test (TTT)**:
  - Beyond linguistic indistinguishability.
  - Mimics also other human behaviors
- Weak AI is generally seen as feasible achieve

Strong AI

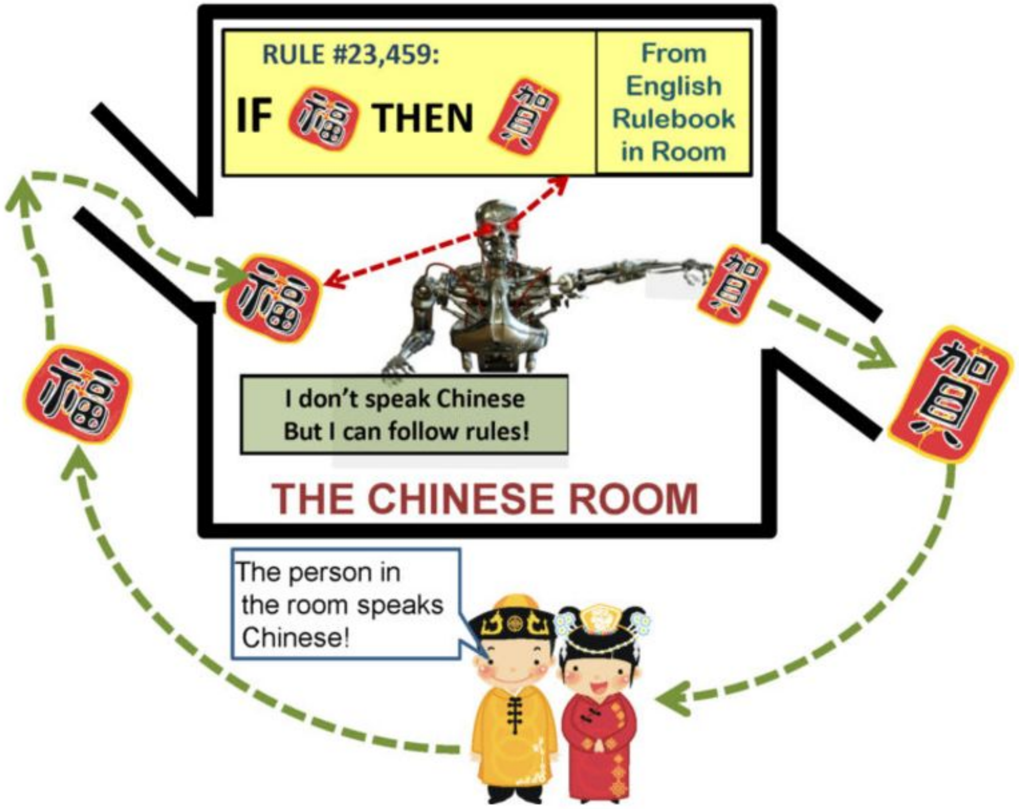
- **Goal:** Create artificial persons with human-like **mental powers**, including **consciousness**.
- Machines would truly **think, feel, and possess consciousness**, similar to humans.
- Strong AI strives for **true human-like intelligence** but faces philosophical skepticism.
- Heavily debated; critics (e.g., Searle) challenge whether machines can truly **think or feel**.

Chapter 8.2 The Chinese Room Argument

Thought Experiment Introduced by **John Searle (1980)** to challenge "Strong AI."

- Searle is in a room with a rulebook for Chinese.
- Native Chinese speakers pass questions (input) written in Chinese into the room.
- Searle uses the rulebook to generate answers (output) in Chinese.
- Searle manipulates symbols without understanding Chinese.
- Conclusion: Computers, like Searle in the room, process information syntactically without true semantic understanding.

Machines can simulate thought but cannot possess genuine understanding or consciousness.



https://www.learningmachines101.com/lm101-006-interpret-turing-test-results/

Chapter 8.2 The Chinese Room Argument

Criticism from AI Researchers:

- Many AI experts dismiss CRA, calling it unrealistic or irrelevant to real-world AI development.
- Focus is on building machines that act intelligent, not on proving they "understand."

Searle’s Response:

- Even advanced AI systems only simulate understanding.
- He argues they manipulate symbols without real meaning or consciousness.

Why CRA Still Matters:

- Sparks debates on whether machines can truly **think** or just mimic thought.
- Raises concerns about future AI, including ethical issues and risks.
- Relevant for guiding AI research and policymaking.

Chapter 8.3 The Gödelian argument

2nd argument against “strong” AI (= machines with all human mental powers)

J.R. Lucas (1964):

No machine can ever reach human-level intelligence according Gödel’s first incompleteness theorem.

Gödel (1906-1978; a logician, mathematician and philosopher)

Any consistent and powerful enough set of axioms and rules cannot prove or disprove the statements that can be written down using symbols:

- some statements are true but unprovable
- some statements are false but undecidable

Source: The New Yorker, 2021; https://www.newyorker.com/

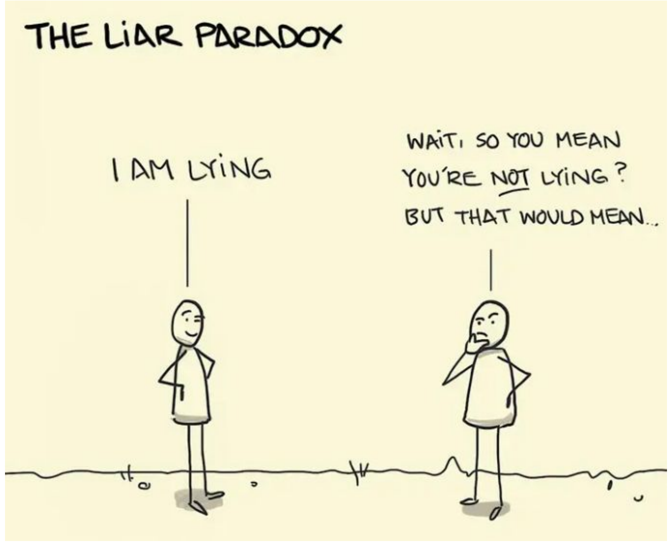
Chapter 8.3 The Gödelian argument

Consistent set of axioms and rules ≠ contradiction

Contradiction: a statement that is both true and false at the same time

“This statement is false”:

- if it is true, then it is false
- if if it is false, then it is true



Source: https://sketchplanations.com/

“This statement cannot be proved using the axioms and rules” (contradiction)

Gödel’s Incompleteness Theorems for Dummies! | by Almas K. | Medium

Chapter 8.3 The Gödelian argument

Gödel’s theorem means: no matter how powerful and AI system is, there will always be some truths that it cannot know

1. AI systems can never be truly perfect or infallible (capable of making mistakes)
2. AI systems will always need to be updated with new information
3. There are things AI systems can never understand

Limitations of Gödel’s theorems: they only apply to:

- formal axiomatic systems (do not apply to all kinds of knowledge; e.g. acquired through experience or intuition)
- systems that are consistent
- systems that are capable of modelling basic arithmetic (do not apply to more complex models; e.g. systems that can model the real world)

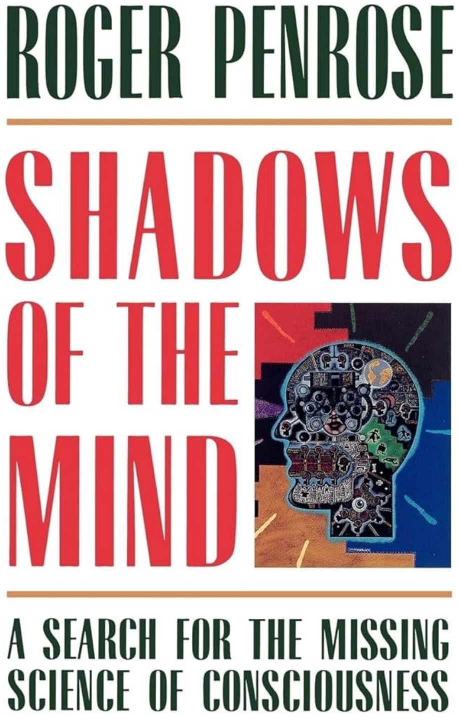
Gödel’s Incompleteness Theorem and the Limits of AI | by Matt Fleetwood | Medium

Chapter 8.3 The Gödelian argument

J.R. Lucas (1964): “No machine can ever reach human-level intelligence according Gödel’s first incompleteness theorem”. Although it has not proved to be compelling, initiated a debate.

Roger Penrose (physicist) defender of the argument:

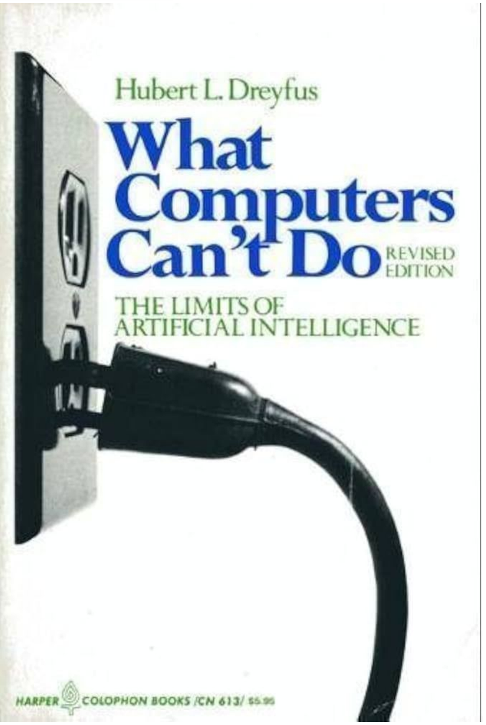
- The Emperor’s New Mind (1989)
- Shadows of the Mind (1994)



The four fundamental positions on AI (extracted from Bringsjord and Xiao, 2000)

1. All thinking is computation (feelings of conscious awareness evoked by appropriate computations)
2. Awareness is a feature of the brain physical action
3. Appropriate physical action of the brain evokes awareness (cannot be properly simulated computationally)
4. Awareness cannot be explained by physical, computatiational, or any other scientific terms

Chapter 8.4 Additional topics and readings


Hubert L. Dreyfus
What Computers Can't Do
REVISED EDITION
THE LIMITS OF ARTIFICIAL INTELLIGENCE
HARPER COLOPHON BOOKS /CN 613/ $5.95

3rd attack to “strong” AI: the Dreyfusian attack (Dreyfus)

“Human expertise is not based on the explicit, disembodied, mechanical manipulation of symbolic information”

- Intelligence cannot be reduced to following rules
- Context is critical to facilitating learning (environment “shows up” for an individual as one progresses through their learning curve: novice vs. expert)
- Mind and body work together to cultivate learning and intelligence
- A living body has interests and needs (nothing is neutral in the act of learning and mastering skills)

To replicate intelligence with AI, we need:

1. replicating intelligence and consciousness
2. programming:
   a. the way things show up for us
   b. a model of our bodies
   c. our distinctive values and motivations

Fjelland, 2020. Why general artificial intelligence will not be realized ? *Humanities & Social Sciences* Communications, 7:10. https://doi.org/10.1057/s41599-020-0494-4