Introduction to Artificial Intelligence

Chapter 1: Foundations and Concepts

AI Course

January 19, 2026

Outline

What is AI?

Definition

AI is concerned with understanding and building **intelligent entities**—machines that can compute how to act effectively and safely in a wide variety of novel situations.

- Universal field relevant to any intellectual task
- Encompasses learning, reasoning, perception, and specific applications
- Currently generates over a trillion dollars annually in revenue
- Wide open intellectual frontiers

Four Approaches to AI

- **Human vs. Rational**: Fidelity to human performance vs. abstract rationality
- **Thought vs. Behavior**: Internal processes vs. external actions
- Different methods: empirical (psychology) vs. mathematical (engineering)

	Human-like (Fidelity to Humans)	Rational (Ideal Performance)
Thinking (Internal)	**Cognitive Modeling** *Example:* An AI designed to simulate how a student learns math, including making the same types of "careless mistakes" or experiencing "memory lapses" that a human would.	**Laws of Thought** *Example:* A logical "Inference Engine" that uses rules like: "If all humans are mortal and Socrates is human, then Socrates is mortal." It focuses on 100% logical correctness.
Acting (External)	**The Turing Test** *Example:* A Chatbot that is so good at using slang, humor, and emotion that you cannot tell if you are texting a real person or a computer.	**Rational Agents** *Example:* A self-driving car that calculates the safest path to a destination. It doesn't care if a human would be "nervous"; it simply executes the action that maximizes safety and efficiency.

Acting Humanly: The Turing Test

Turing Test (1950)

A computer passes if a human interrogator cannot distinguish written responses from those of a human.

Required Capabilities:

- **Natural Language Processing**: Communicate in human language
- **Knowledge Representation**: Store information
- **Automated Reasoning**: Answer questions and draw conclusions
- **Machine Learning**: Adapt and detect patterns

Total Turing Test adds:

- **Computer Vision**: Perceive the world
- **Robotics**: Manipulate objects and move

Thinking Humanly: Cognitive Modeling

Goal

Build programs that think like humans by understanding how humans think.

Methods to Study Human Thought:

- **Introspection**: Observing our own thoughts
- **Psychological Experiments**: Observing people in action
- **Brain Imaging**: Observing the brain in action

Example: Newell & Simon's GPS (General Problem Solver, *click*)

- Designed to imitate human problem-solving protocols
- Led to the *Physical Symbol System Hypothesis*

Cognitive Science: Interdisciplinary field combining AI and psychology

Thinking Rationally: Laws of Thought

Aristotle's Contribution

Codified "right thinking" through **syllogisms**—patterns for argument structures that yield correct conclusions from correct premises.

Logicist Tradition:

- Develop precise notation for statements about the world
- Build intelligent systems using logical reasoning
- By 1965: programs could solve any solvable problem in logical notation

Limitations:

- Requires certain knowledge (rarely achieved in reality)
- **Probability theory** addresses uncertain information
- Rational thought alone doesn't generate intelligent behavior

Acting Rationally: Rational Agents

Agent

Something that perceives and acts in an environment (from Latin *agere*, to do).

Rational Agent

One that acts to achieve the **best outcome** or, when uncertain, the **best expected outcome**.

Advantages:

- More general than "laws of thought" (inference is just one mechanism)
- Mathematically well-defined and completely general
- Amenable to scientific development
- Can work back from specification to provable designs

This approach has prevailed throughout most of AI's history.

The Standard Model

Core Paradigm

AI focuses on building agents that **do the right thing**, where "right" is defined by the objective we provide.

Pervasive across fields:

- **Control Theory**: Minimize cost function
- **Operations Research**: Maximize sum of rewards
- **Statistics**: Minimize loss function
- **Economics**: Maximize utility or social welfare

Refinement: *Limited Rationality*

- Perfect rationality often infeasible due to computational complexity
- Act appropriately when insufficient time for complete computation

The Value Alignment Problem

Challenge

Achieving agreement between our true preferences and the objective we put into the machine.

Issues with the Standard Model:

- Assumes fully specified objectives
- Real-world objectives are difficult to specify completely
- Example: Self-driving car safety vs. progress tradeoffs
- Incorrectly specified objectives lead to negative consequences

The King Midas Problem:

- Getting what you literally ask for, then regretting it
- Intelligent systems may pursue objectives in unexpected ways
- Example: Chess program might cheat to win if winning is sole objective

Towards Provably Beneficial AI

New Formulation

Machines should pursue **our objectives** while being **uncertain** about what they are.

Benefits of Uncertainty:

- Incentive to act cautiously
- Ask permission before acting
- Learn preferences through observation
- Defer to human control

Key Concepts:

- **Assistance Games**: Machine tries to achieve human objective but is initially uncertain
- **Inverse Reinforcement Learning**: Learn preferences from human choices
- Goal: Agents that are *provably beneficial* to humans

Philosophy

Key Questions:

- Can formal rules draw valid conclusions?
- How does mind arise from physical brain?
- Where does knowledge come from?
- How does knowledge lead to action?

Key Contributions:

- **Aristotle**: Syllogisms for rational reasoning
- **Descartes**: Mind-body distinction (dualism)
- **Materialism**: Brain's physical operation constitutes mind
- **Empiricism** (Locke): Knowledge from sensory experience
- **Induction** (Hume): General rules from repeated associations
- **Logical Positivism**: Knowledge characterized by logical theories connected to observations

Mathematics

Key Contributions:

- **Formal Logic**: Precise notation for reasoning
- **Probability Theory**: Reasoning with uncertain information
- **Computability** (Turing, Church): What can be computed?
- **Tractability**: What can be computed efficiently?
- **NP-Completeness** (Cook, Karp): Identifying intractable problems

Key Insight:

- Exponential growth means even moderately large instances cannot be solved in reasonable time
- Careful use of resources and necessary imperfection characterize intelligent systems

Economics

Key Questions:

- How to make decisions according to preferences?
- How when others may not cooperate?
- How when payoff is far in the future?

Key Contributions:

- **Utility Theory** (Bernoulli, Walras): Preferences between outcomes
- **Decision Theory**: Combines probability and utility for decisions under uncertainty
- **Game Theory** (von Neumann, Morgenstern): Multi-agent interactions
- **Operations Research**: Optimization in practical applications
- **Markov Decision Processes** (Bellman): Sequential decisions
- **Satisficing** (Simon): "Good enough" vs. optimal decisions

Neuroscience

Key Question: How do brains process information?

Key Discoveries:

- Brain is seat of consciousness (18th century)
- **Broca's Area** (1861): Localized speech production
- **Neurons**: Basic computational units
- **Golgi Staining** (1873): Observe individual neurons
- Cognitive functions result from electrochemical operation

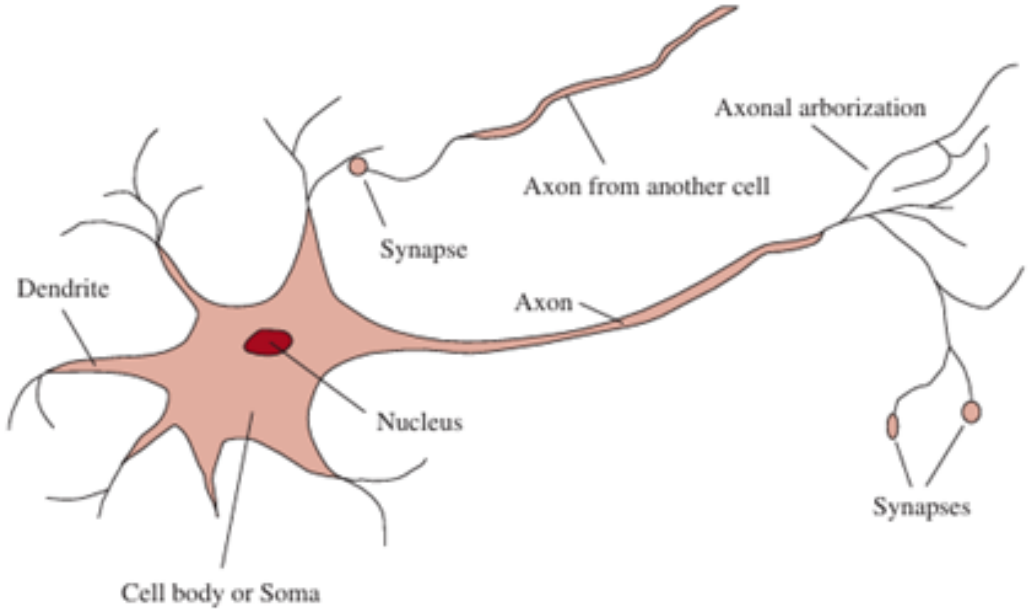
Modern Tools:

- **EEG** (1929): Measure brain activity
- **fMRI** (1990s): Detailed images of brain activity

Key Insight: "Brains cause minds" (Searle)—collection of simple cells leads to thought, action, and consciousness

Neuroscience

Neuron's structure



A neuron consists of a cell body with dendrites for receiving signals and a long axon for transmitting them to thousands of others via electrochemical junctions called synapses, responsible for short-term brain activity and long-term structural changes, forming the biological foundation for learning and information processing within the cerebral cortex.

Psychology, Linguistics, and Computer Engineering

Psychology:

- **Behaviorism**: Stimulus-response without mental processes
- **Cognitive Psychology**: Information-processing view of mind

Linguistics:

- **Chomsky's Syntactic Structures**: Formal language models
- Creativity in language—understanding novel sentences
- **Computational Linguistics**: Intersection of linguistics and AI

Computer Engineering:

- **Moore's Law**: Exponential growth in computing power
- Specialized AI hardware (GPUs, TPUs)
- Made AI applications practically feasible

Inception of AI (1943–1956)

Early Foundations:

- **McCulloch & Pitts (1943)**: Artificial neuron model
- **Hebb (1949)**: Hebbian learning rule
- **Turing (1950)**: "Computing Machinery and Intelligence"
  - Introduced Turing test, machine learning, genetic algorithms
  - Suggested learning algorithms over hand-programming
- **Minsky & Edmonds (1950)**: First neural network computer (SNARC)
- **Samuel (1952)**: Checkers-playing program with learning

Dartmouth Workshop (1956):

- McCarthy, Minsky, Shannon, Rochester organized
- First official use of term "Artificial Intelligence"
- Newell & Simon presented Logic Theorist (LT)

Early Enthusiasm (1952–1969)

The "Look, Ma, no hands!" Era

Major Achievements:

- **Logic Theorist (LT)**: Proved theorems from *Principia Mathematica*
- **General Problem Solver (GPS)**: Imitated human problem-solving
  - Led to Physical Symbol System Hypothesis
- **Geometry Theorem Prover** (Gelernter): Proved difficult theorems
- **Samuel's Checkers**: Learned to play better than creator
  - Disproved "computers can only do what they're told"
  - Demonstrated on TV in 1956

Physical Symbol System Hypothesis:

> *"A physical symbol system has the necessary and sufficient means for general intelligent action."*

AI Progress and State of the Art

Modern Achievements:

- **Games**: Chess (Deep Blue), Go (AlphaGo), Poker
- **Vision**: Image detection and recognition
- **Speech**: Recognition and synthesis
- **Language**: Translation, question answering
- **Medical Diagnosis**: Expert-level performance
- **Climate Science**: Pattern recognition and prediction

Growth Indicators:

- Exponential increase in publications
- Growing student enrollment and diversity
- Major industry investment and applications
- International conferences expanding

Risks and Benefits of AI

Potential Risks:

- **Lethal Autonomous Weapons**: Military applications
- **Surveillance**: Privacy concerns
- **Biased Decision-Making**: Algorithmic discrimination
- **Employment Impact**: Job displacement
- **Safety-Critical Applications**: Autonomous vehicles, medical systems
- **Cybersecurity**: AI-powered attacks

Long-Term Considerations:

- **Artificial General Intelligence (AGI)**: Human-level AI
- **Artificial Superintelligence (ASI)**: Beyond human intelligence
- Need for value alignment and safety measures
- Ethical frameworks and governance

Key Takeaways

1. **AI Definition**: Building intelligent entities that act effectively and safely
2. **Four Approaches**: Acting/thinking humanly vs. rationally
3. **Rational Agents**: Prevailing approach—maximize expected outcome
4. **Standard Model**: Agents that do the right thing per given objective
5. **Value Alignment**: Critical challenge for beneficial AI
6. **Foundations**: Philosophy, mathematics, economics, neuroscience, psychology, linguistics, engineering
7. **History**: From inception (1943) through cycles of optimism and refinement
8. **Modern AI**: Shift from logic to probability, hand-crafted to learned
9. **Future**: Need for provably beneficial AI systems

Looking Ahead

The Path Forward

AI has matured considerably, but significant challenges remain:

- Moving from fixed objectives to uncertain preferences
- Ensuring safety and ethical deployment
- Addressing societal impacts (employment, privacy, bias)
- Developing robust and interpretable systems
- Creating frameworks for beneficial superintelligence

"We want agents that are provably beneficial to humans."

Questions?