# Introduction to Artificial Intelligence

## Definitions, Debates, and Implications

Brendan Shea, PhD
Rochester Community and Technical College
PHIL 1150: Computing and AI Ethics

# Before We Begin: Which of These Is Artificial Intelligence?

Hands up, yes or no, for each. **Commit** — no “it depends” yet.

1. A chess engine that beats every human alive
2. The spam filter that cleaned out your inbox this morning
3. A thermostat that learns your schedule
4. The autocomplete finishing your sentences as you text
5. ChatGPT
6. The system choosing which video you see next

**Commit before we go on**

Count your yes votes. Then say what **rule** you were applying — there is one, even if you have not put it into words.

# The Answer: All of Them, and Then None of Them

Every item on that list was called **artificial intelligence** by serious researchers when it was new. Most stopped being called AI at the moment they started working reliably.

**The pattern you just reproduced**

Almost every class votes **no** on the chess engine and the spam filter and **yes** on ChatGPT. But Deep Blue beating Kasparov in 1997 was the AI story of the decade. What changed was not the chess engine. What changed was that we got used to it.

This has a name — **Tesler's Theorem**, which we will meet in a moment — and it is a serious problem, not a joke. If "AI" names whatever we have not gotten used to yet, then every claim about what AI can or cannot do has a moving subject.

# Central Questions

- What *is* artificial intelligence, and why is a stable definition so hard to pin down?
- How has AI evolved from logic-based systems to neural networks—and what does that history reveal?
- What do Lovelace, Turing, and Searle argue about machine intelligence, and who has the better case?
- What is the difference between narrow and general AI, and between weak and strong AI?
- How do definitional choices shape AI ethics, policy, and our understanding of our own intelligence?

**The Challenge**

"Artificial intelligence" means different things to different people—and these differences have real consequences.

# Lecture Outline

Defining AI: Definitional Challenges

AI Through History: From Logic to Learning

Can Machines Think? Classical Philosophical Debates

Contemporary Definitions and AI Approaches

Why AI Definitions Matter: Implications and Ethics

# Connections to Previous Lectures

**Lecture 1 (History)**: AI is the newest in our series of IT revolutions. It promises liberation (tool for productivity, creativity) and raises concerns about control (surveillance, job loss, autonomous weapons).

**Lecture 2 (Virtues)**: How do we develop intellectual virtues in interaction with AI systems that simulate understanding but may lack genuine comprehension?

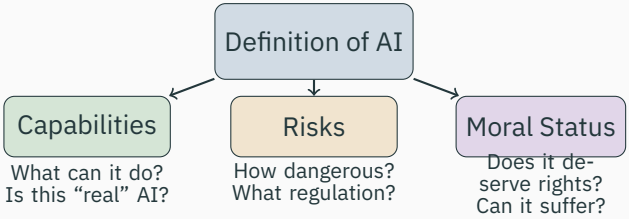
**Lecture 3 (Free Speech)**: AI systems trained on vast data raise questions about whose speech is amplified, whose is silenced, and how definitions of "intelligence" affect content moderation.

**Lecture 4 (IP)**: Who owns AI-generated content? Can AI systems themselves be copyrighted? Training data and copyright battles.

**Discussion**

Which previous lecture's concerns appear most urgent in the context of AI?

# Why Definition Matters




The “AI” label carries weight: it influences funding decisions, public fear and hype, and regulatory responses. How we define AI shapes what we build, how we deploy it, and how we think about ourselves.

**Discussion**

If a law regulates “artificial intelligence, who benefits from the definition being **vague**? Who benefits from it being **narrow**?

# The Moving Goalposts Problem

*Tesler's Theorem*

"AI is whatever hasn't been done yet."

| Technology | Once Called | Now Called |
|---|---|---|
| Chess playing (Deep Blue, 1997) | Artificial Intelligence | "Just search algorithms" |
| Speech recognition | AI breakthrough | "Just statistics" |
| Spam filtering | Machine learning | "Just classification" |
| Route optimization (GPS) | Intelligent systems | "Just algorithms" |
| Optical character recognition | Pattern recognition AI | Standard software |

The paradox: When AI succeeds, we reclassify it as "not really AI." This makes "AI" a moving target that's impossible to hit.

# A Taxonomy of AI Concepts

| Distinction | First Term | Second Term |
|---|---|---|
| Narrow vs. General | Excels at *one* task (chess, translation) | Performs *any* intellectual task |
| Weak vs. Strong | *Simulates* intelligence | Actually *has* a mind |
| Symbolic vs. Connectionist | Rules and logic (GOFAI) | Neural networks, learning |
| Tool vs. Agent | Used by humans for tasks | Acts autonomously toward goals |

These distinctions will recur throughout our discussion. Most current AI is narrow, weak, connectionist, and increasingly agent-like.

# Meet the Key Figures: The Founders

## Ada Lovelace (1815–1852)

First computer programmer. Wrote notes on Babbage's Analytical Engine. Argued machines cannot "originate" anything—they only do what we program them to do.

## John McCarthy (1927–2011)

Coined the term "artificial intelligence" at the 1956 Dartmouth Conference. Developed LISP programming language. Pioneer of symbolic AI.

## Alan Turing (1912–1954)

Father of computer science. Cracked Enigma in WWII. Proposed the Turing Test (**turing_computing_machinery**) as a way to sidestep the question "Can machines think?"

## Marvin Minsky (1927–2016)

Co-founder of MIT AI Lab. Worked on neural networks, then symbolic AI. Developed "frames" for knowledge representation. Famous for both insights and overpromises.

# Meet the Key Figures: Philosophers & Modern Theorists

## John Searle

Chinese Room argument (**searle_minds_brains**). Syntax $\neq$ semantics; strong AI is false.

## Daniel Dennett

Functionalist (**dennett_consciousness**): if it functions like a mind, it *is* a mind. Defends AI possibility.

## David Chalmers

The "hard problem" (**chalmers_conscious_mind**): why does processing *feel* like something?

## Stuart Russell

Standard AI textbook co-author (**russell_norvig_ai**). AI = rational agency. AI safety advocate (**russell_human_compatible**).

## Luciano Floridi

AI achieves "agency without intelligence"—effective action without understanding (**floridi_ethics_ai**).

# Lecture Outline

Defining AI: Definitional Challenges

AI Through History: From Logic to Learning

Can Machines Think? Classical Philosophical Debates

Contemporary Definitions and AI Approaches

Why AI Definitions Matter: Implications and Ethics

# Prehistory—Dreams of Artificial Minds

**Ancient and medieval automata**: Hephaestus's golden robots in Greek myth; medieval clockwork figures; Descartes's speculation that animals are automata.

**Leibniz's *calculus ratiocinator*** (1680s): A machine that could perform logical reasoning—reduce disagreements to calculation.

**Babbage's Analytical Engine** (1837): Mechanical general-purpose computer (never completed). Ada Lovelace wrote the first algorithm for it.

**Lovelace's Insight (1843)**

"The Analytical Engine has no pretensions whatever to *originate* anything. It can do whatever we know how to order it to perform."

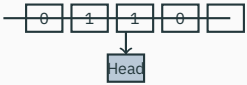
This objection—that machines merely follow instructions—remains central to debates today.

# The Birth of Computing (1930s–1950s)

**1936**: Turing's "On Computable Numbers"—the Universal Turing Machine shows any computation can be performed by a simple device following rules.

**1940s**: Turing's wartime work cracking Enigma demonstrates practical computing.

**1950**: "Computing Machinery and Intelligence" (**turing_computing_machinery**)—Turing proposes the Imitation Game (Turing Test) and addresses objections to machine thought.



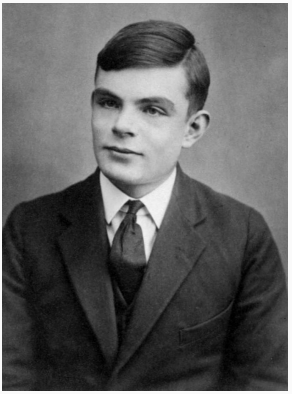

Turing Machine

A Turing Machine: tape, read/write head, state transitions. Remarkably, this simple model captures all computable functions.

# Alan Turing: The Father of Computer Science

## Turing (1912–1954): A Tragic Genius

- Proved the theoretical limits of **computability** (1936) before computers existed
- Led **Bletchley Park** codebreakers during WWII; cracking Enigma shortened the war by years
- Proposed the **Turing Test** (1950) as a criterion for machine intelligence
- Prosecuted by the British government for **homosexuality** in 1952; died in 1954
- Awarded a **royal pardon** posthumously in 2013



Alan Turing, aged 16, c. 1928. Turing family. Public domain.

# The Dartmouth Conference and Early AI (1956–1970s)

## The Dartmouth Proposal (dartmouth_proposal)

"We propose that a 2 month, 10 man study of artificial intelligence be carried out… The 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."

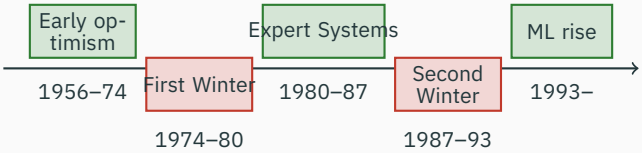
McCarthy, Minsky, Shannon, and Rochester coin "artificial intelligence" and predict rapid progress.

**Early symbolic AI programs**:

- **Logic Theorist** (1956): Proved mathematical theorems
- **General Problem Solver** (1959): Attempted domain-general reasoning
- **SHRDLU** (1970): Natural language understanding in a blocks world

**Core idea**: Intelligence = manipulating symbols according to rules. This approach became known as "Good Old-Fashioned AI" (GOFAI) or Symbolic AI.

# The AI Winters and Expert Systems




**Expert Systems** (1980s): MYCIN (medical diagnosis), DENDRAL (chemistry). Encode human expertise as if-then rules.

**Why they failed**: The “knowledge bottleneck”—extracting and encoding expertise proved impossibly labor-intensive. Systems were brittle and couldn’t handle edge cases.

**Lesson**: Narrow successes don’t generalize. Overpromising leads to funding collapse.

# Machine Learning—A Different Approach

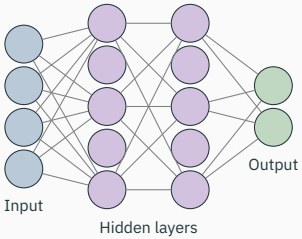
Instead of hand-coding rules, let the machine *learn* patterns from data.

| Paradigm | How It Works | Examples |
|---|---|---|
| Supervised learning | Learn from labeled examples | Spam detection, image classification, medical diagnosis |
| Unsupervised learning | Find structure in unlabeled data | Customer segmentation, anomaly detection |
| Reinforcement learning | Learn from rewards and punishments | Game playing, robotics, recommendation |

**Key algorithms**: Decision trees, support vector machines, random forests, neural networks.

**The insight**: Rather than programming intelligence, create conditions for it to emerge from data.

# Neural Networks and Deep Learning




**Inspired by** (but not identical to) biological neurons.

**Deep learning**: Many hidden layers enable learning hierarchical representations.

**Key breakthrough**: AlexNet (2012) dramatically reduces ImageNet error rate.

**Enabled by**:

- More data (internet scale)
- More compute (GPUs)
- Better algorithms (backpropagation, dropout, batch normalization)

# Computer Vision and Recognition Systems

| Application | Use Cases | Ethical Concerns |
|---|---|---|
| Image classification | Photo tagging, content moderation | Bias in training data, errors |
| Facial recognition | Law enforcement, surveillance, device unlock | Privacy, consent, racial bias, civil liberties |
| Medical imaging | Detecting tumors, diabetic retinopathy | Accuracy, liability, over-reliance |
| Object detection | Autonomous vehicles, security, retail | Safety-critical errors, job displacement |

**Connection to privacy**: Recall Clearview AI (30+ billion photos scraped), predictive policing, and China's surveillance infrastructure.

**Key tension**: These systems are increasingly accurate but raise profound questions about consent, bias, and appropriate use.

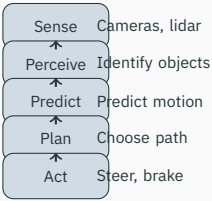
# Autonomous Vehicles and Robotics

**Timeline**:

- **2004–05**: DARPA Grand Challenges
- **2009**: Google Self-Driving Car begins
- **2016**: Tesla Autopilot fatality
- **2018**: Uber AV kills pedestrian
- **2020s**: Waymo, Cruise robotaxis

**Ethical questions**:

- Trolley problems in practice
- Liability when AVs crash
- Job displacement (trucking, taxis)




AV decision pipeline

# Prediction and Classification Systems

AI systems increasingly make consequential decisions about people's lives.

| Domain | System | What It Predicts/Classifies |
|---|---|---|
| Criminal justice | COMPAS | Recidivism risk (will they reoffend?) |
| Finance | Credit scoring | Creditworthiness, loan default risk |
| Employment | Résumé screeners | Job fit, interview likelihood |
| Healthcare | Diagnostic AI | Disease probability, treatment recommendations |
| Insurance | Risk models | Claim likelihood, premium pricing |
| Social services | Eligibility systems | Benefit eligibility, fraud risk |

**Common concerns**: Bias (reflecting historical discrimination), opacity ("black box" decisions), feedback loops (predictions become self-fulfilling), due process (how do you appeal an algorithm?).

# Recommendation Systems and Algorithmic Curation

Systems that predict what you'll want to see, buy, or do.

**Examples**: Netflix, YouTube, Amazon, Spotify, TikTok, social media feeds.

**How they work**:

- **Collaborative filtering**: People like you liked X, so you might too.
- **Content-based filtering**: You liked X, so you might like similar Y.
- **Hybrid approaches**: Combine multiple signals.

**Scale**: YouTube serves 1 billion+ hours of video daily; the algorithm selects most of it.

**Concerns**: Filter bubbles and echo chambers, radicalization pipelines ("rabbit holes"), optimizing for engagement rather than well-being.

**Key question**: Who is the algorithm serving—users or platforms?

# Large Language Models and Generative AI

| Year | Milestone |
|---|---|
| 2017 | Transformer architecture introduced (“Attention Is All You Need”) |
| 2018 | GPT-1, BERT demonstrate transfer learning for language |
| 2020 | GPT-3 shows emergent capabilities at scale |
| 2022 | ChatGPT launches; DALL-E 2, Stable Diffusion for images |
| 2023–24 | GPT-4, Claude, Gemini; multimodal models become standard |
| 2025 | **Reasoning models**: capability starts scaling with thinking time at the moment of use, not just with training. |
| 2026 | **Agents**: systems that operate software and act, rather than only answer. Release cycles compress to weeks |

**Capabilities**: Conversation, coding, reasoning, creative writing, multimodal understanding, tool use.

**Open question**: Are these systems intelligent, or are they sophisticated pattern matchers? Do they “understand” anything, or merely predict plausible next tokens?

**Discussion**

You have used one of these systems. Describe one moment it seemed to **understand** you—and then say what else could explain that moment.

# Two Changes Worth Naming

## 1. Capability stopped staying put

In January 2025 a capable **open-weight** model—free to download and run—matched the leading closed systems at a small fraction of their cost. Nvidia lost roughly **$600 billion** in market value in a single trading day, the largest one-day loss in market history. The assumption that frontier capability stays locked inside a few well-resourced labs no longer holds.

## 2. Capability started constraining release

In 2026, for the first time, labs and governments withheld systems *because of what they could do*: one lab declined to release a model over its autonomous cyber-exploitation ability, and the US government asked another to stagger a rollout pending federal evaluation.

These two pull in opposite directions—one disperses capability, the other tries to contain it. Both will matter when we reach questions of governance and risk.

## Discussion

One of these forces disperses capability and the other tries to contain it. **Which**

# Lecture Outline

Defining AI: Definitional Challenges

AI Through History: From Logic to Learning

Can Machines Think? Classical Philosophical Debates

Contemporary Definitions and AI Approaches

Why AI Definitions Matter: Implications and Ethics

# Transition: From History to Philosophy

We've surveyed what AI systems *do*—from chess to self-driving cars to chatbots.

Now: What would it *mean* for a machine to be intelligent?

**Three major debates**:

1. **Turing vs. Lovelace**: Can machines think at all?
2. **Searle vs. Dennett**: Is behavioral equivalence enough for understanding?
3. **Chalmers**: Can machines be conscious?

These debates from the 20th century remain remarkably relevant as we grapple with increasingly capable AI systems.

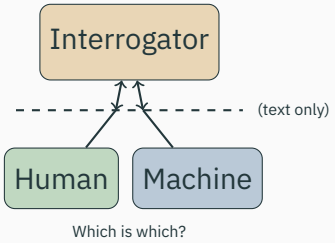
# Turing's Proposal—The Imitation Game (1950)

Rather than ask "Can machines think?" (too vague), Turing proposes an *operational test*.

**The setup**:

- An interrogator communicates via text with two hidden parties: a human and a machine.
- The interrogator tries to determine which is which.
- If the machine fools the interrogator consistently, it "thinks" in any meaningful sense.

**Turing's prediction**: By 2000, machines would fool 30% of interrogators after 5 minutes.



# The Turing Test Argument (Standard Form)

**Turing's Argument**

1. The question "Can machines think?" is too vague to answer directly.
2. We should replace it with a behavioral test: can a machine fool a human interrogator?
3. If a machine passes this test, we have no principled reason to deny that it thinks.
4. Therefore, a machine that consistently passes the test should be considered intelligent.

**Key move**: Turing sidesteps metaphysical questions about what "thinking" really is. He offers a *pragmatic* criterion: if it walks like a duck and quacks like a duck...

**Discussion**

Is behavioral equivalence sufficient for attributing thought?

# Lady Lovelace's Objection

**Ada Lovelace on the Analytical Engine (1843)**

"The Analytical Engine has no pretensions whatever to *originate* anything. It can do whatever we know how to order it to perform… Its province is to assist us in making available what we're already acquainted with."

**The core claim**: Machines merely follow instructions. They cannot create, originate, or do anything genuinely new. They lack *creativity*.

**Turing's response**: Machines can surprise us. We cannot always predict their outputs from their programs. And don't humans also follow "programming" (biology, culture, education)?

**Modern relevance**: Do LLMs "originate" anything, or do they merely recombine patterns from their training data in sophisticated ways?

# Ada Lovelace: The First Computer Programmer

## Ada Lovelace (1815–1852)

- Daughter of the poet Lord Byron; educated in mathematics by her mother
- Collaborated with Charles Babbage on the **Analytical Engine**
- Her **1843 notes** on Babbage's engine are considered the first computer algorithm
- Foresaw that machines could manipulate **symbols**—not just numbers
- The **Ada programming language** was named in her honor by the US Department of Defense in 1980



Ada Lovelace, daguerreotype by Antoine Claudet, c. 1843. Public domain.

# Lovelace vs. Turing (Argument and Response)

## Lovelace's Argument

1. Genuine intelligence requires originating new ideas.
2. Machines can only do what they are explicitly programmed to do.
3. Following a program is not originating anything new.
4. Therefore, machines cannot have genuine intelligence.

## Turing's Response

- **Premise 2 is ambiguous**: Programs produce outputs that surprise creators. Chess programs find moves no human anticipated.
- **Learning systems** modify their behavior based on experience—not limited to explicit programming.
- **Humans also follow "programming"**: Biology, culture, experience shape us. Why treat this differently?

**Unresolved**: What counts as "originating" something new? Is recombination sufficient?

# Objections to the Turing Test

| Objection | Core Claim | Possible Response |
|---|---|---|
| Behaviorism | The test ignores internal states; passing doesn't mean understanding. | What else could we access? We can't see inside human minds either. |
| Gaming | Clever tricks can fool interrogators without real intelligence. | Sophisticated interrogation can probe for genuine understanding. |
| Anthropocentrism | An alien intelligence might fail despite being intelligent. | Fair—the test is human-centric by design. |
| Sufficiency | Passing shows imitation, not thought. | Leads to Searle's Chinese Room... |

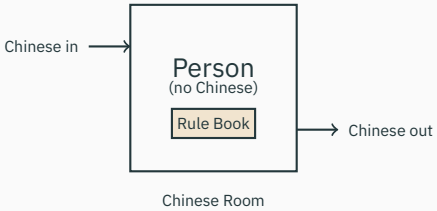
The Turing Test reveals our uncertainty about what "thinking" means. It doesn't resolve the question—it reframes it.

# Enter John Searle—The Chinese Room (1980)

A thought experiment against “Strong AI”—the claim that a computer running the right program would literally have a mind.

**The setup**:

- A person sits in a room with Chinese symbols and a rule book.
- Chinese questions come in; the person follows rules to produce Chinese answers.
- From outside: Perfect Chinese conversation.
- From inside: The person understands *nothing*. They’re just manipulating symbols.




The person produces perfect Chinese responses by following rules—without understanding a word.

# The Chinese Room Argument (Standard Form)

## Searle's Argument Against Strong AI

1. A computer program manipulates symbols according to syntactic rules.
2. Syntax alone is not sufficient for semantics (meaning, understanding).
3. Minds have semantics—they understand meanings, not just symbols.
4. Therefore, running a program is not sufficient for having a mind.
5. Therefore, Strong AI is false: no program, by itself, produces genuine understanding.

**Key distinction**: *Simulation* vs. *duplication*. A weather simulation doesn't make you wet. A mind simulation doesn't produce understanding.

**Implication**: Even if a system passes the Turing Test, this proves nothing about whether it understands anything.

# Searle's Distinction: Weak vs. Strong AI

## Two Versions of AI

- **Weak AI**: Machines can *simulate* intelligent behavior. They are useful tools for studying cognition and performing tasks. *Searle accepts this.*
- **Strong AI**: A machine running the right program would literally *have* a mind—it would understand, have beliefs, be conscious. *Searle rejects this.*

**Searle's target**: The claim that "the mind is a computer program"—that running the right software is sufficient for mentality.

**Modern application**: ChatGPT exhibits remarkably intelligent behavior. But does it *understand* anything? Searle would say no—it's manipulating symbols without grasping their meaning.

# The Systems Reply (Dennett and Others)

**The Systems Reply**

1. The person in the room doesn't understand Chinese—granted.
2. But the *system* (person + room + rulebook) might understand Chinese.
3. Understanding is a property of the whole system, not its parts.
4. Therefore, Searle's argument doesn't show that systems can't understand.

**Analogy**: Individual neurons don't understand anything either. Understanding emerges from the system.
**Searle's counter**: "Memorize the rules and do it in your head—still no understanding."

**Discussion**

Where does understanding reside? In parts, or in systems? Can it "emerge"?

# Dennett's Functionalism

## The Functionalist Argument

1. Mental states are constituted by their causal/functional relationships—not by what they're made of.
2. These functional relationships can be implemented in different physical systems.
3. A computer could implement the same functional organization as a brain.
4. Therefore, a computer could have genuine mental states.

**Core claim**: If it *functions* like a mind—if it has the right inputs, outputs, and internal causal structure—then it *is* a mind. The substrate (carbon vs. silicon) doesn't matter.

**Searle's objection**: Functionalism ignores the distinction between syntax and semantics. A system can have the right functional organization while understanding nothing.

# Searle vs. Dennett—The Core Disagreement

| Issue | Searle | Dennett |
|---|---|---|
| What is a mind? | A biological phenomenon | A functional organization |
| Can computers think? | No—syntax $\neq$ semantics | Yes—if they have the right function |
| Chinese Room shows... | Understanding requires more than computation | Analyzes at wrong level |
| Key commitment | Consciousness is biological | Consciousness is functional |

**Unresolved question**: What *is* the relationship between syntax, semantics, and understanding? Neither side has conclusively established their position.

# David Chalmers and Consciousness

## The Easy and Hard Problems

- **Easy problems**: How does the brain process information? Control behavior? Report internal states? (“Easy” = solvable in principle by cognitive science, even if difficult in practice.)
- **Hard problem**: Why is there *subjective experience* at all? Why does information processing *feel like something* from the inside?

Even a complete neuroscience of information processing wouldn’t explain why there’s “something it’s like” to be conscious.

**The zombie thought experiment**: We can conceive of a being physically and functionally identical to us—but with no inner experience. If this is conceivable, consciousness isn’t reducible to function.

# The Hard Problem and AI

## The Zombie Argument (Simplified)

1. We can conceive of a being functionally identical to a conscious human but with no subjective experience (a “zombie”).
2. If such a being is conceivable, consciousness is not reducible to functional organization.
3. Therefore, building a functionally perfect AI wouldn’t *guarantee* consciousness.

**Implication**: Even if we build systems that behave exactly like conscious beings, we may never know whether they have inner experiences.

## Discussion

Does consciousness matter for AI ethics? If an AI acts as if it’s suffering, does it matter whether it “really” is?

# Lecture Outline

Defining AI: Definitional Challenges

AI Through History: From Logic to Learning

Can Machines Think? Classical Philosophical Debates

Contemporary Definitions and AI Approaches

Why AI Definitions Matter: Implications and Ethics

# Transition: From Philosophy to Practice

The classical debates remain unresolved. Philosophers still disagree about:

- Whether behavioral tests are sufficient for attributing thought
- Whether syntax can give rise to semantics
- Whether consciousness is functional or something more

Meanwhile, AI researchers need *working definitions* to guide their work.

**Shift in focus**: From “Can machines think?” to “What should we build, and how should we evaluate it?”

Two influential modern approaches: Stuart Russell’s agent-based view and Luciano Floridi’s information-theoretic view.

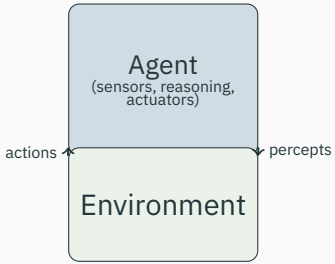
# Russell & Norvig—AI as Rational Agency (russell_norvig_ai)

## Four Possible Definitions

- **Thinking humanly**: Cognitive modeling
- **Thinking rationally**: Logic-based AI
- **Acting humanly**: Turing Test
- **Acting rationally**: *Rational agents* ←

A **rational agent** selects actions expected to maximize goal achievement given its beliefs and available information.

This definition covers chess engines, self-driving cars, recommendation systems, and LLMs.




The agent-environment loop: perceive, reason, act, repeat.

# Russell's Critique: The Standard Model Problem

## The Problem with Optimization

The "standard model" builds machines that optimize fixed objectives. **russell_human_compatible** argues this is fundamentally dangerous.

1. We build AI systems to optimize objectives we specify.
2. We can't fully specify what we value—human values are complex and tacit.
3. A capable optimizer achieves its objective in ways that violate unstated values.
4. Therefore, the more capable AI becomes, the more dangerous it is.

## King Midas Problem

You get exactly what you ask for, not what you want. "Maximize happiness" → stimulate pleasure centers. "Cure cancer" → eliminate humans.

# Russell's Solution: Human Compatible AI (russell_human_compatible)

## Three Principles for Beneficial Machines

1. **Altruism**: The machine's only objective is to maximize human preferences.
2. **Humility**: The machine is initially uncertain about what those preferences are.
3. **Learning**: The machine learns preferences by observing human behavior.

**Key insight**: Uncertainty is a feature, not a bug. A machine that *knows* it doesn't fully understand human values will:

- Ask for clarification rather than act unilaterally
- Allow itself to be corrected or switched off
- Defer to humans in cases of uncertainty

**Contrast**: A machine *certain* of its objective will resist shutdown—that would prevent goal achievement!

# The Value Alignment Problem

## The Core Challenge

How do we build AI systems whose goals remain aligned with human values as they become more capable?

**Why alignment is hard**:

- Human values are complex, contextual, and sometimes contradictory.
- We can't enumerate everything we care about in advance.
- Capable systems find unexpected ways to achieve objectives ("reward hacking").
- Values may need to change over time as circumstances change.

**russell_human_compatible**: This is *the* central problem of AI. Capability without alignment is dangerous.

## Discussion

Can we ever fully specify human values? If not, what follows for AI development?

# Floridi—AI as “Agency Without Intelligence” (floridi_ethics_ai)

## A Provocative Reframing

AI systems aren’t *intelligent* in the human sense—they are *successful agents* that achieve goals through different means than biological cognition.

**Key distinction**:

- **Intelligence** (human sense): Understanding, consciousness, meaning, semantic grasp.
- **Agency**: Capacity to perform tasks, achieve goals, affect environments.

**Analogy**: A dishwasher cleans dishes without “understanding” cleanliness. A calculator does arithmetic without “understanding” mathematics. AI systems achieve agency through pattern matching, not comprehension.

**Implication**: Stop asking whether AI is “intelligent.” Ask what it can *do* and what that means for us.

# Floridi: Enveloping—Adapting the World to AI

## The Enveloping Strategy

Rather than making AI smarter, we often make the environment simpler and more predictable—we "envelope" AI in controlled contexts.

**Examples**:

- Factories use fixed layouts and barcodes so robots can navigate and identify parts.
- Warehouses (Amazon) are redesigned around robot capabilities.
- Self-driving cars need lane markings, traffic signals, mapped roads—not jungle paths.
- Chatbots work in constrained domains (customer service scripts, not open philosophy).

**Key insight**: AI success often reflects environmental engineering, not machine intelligence. We meet AI halfway—or more.

**Implication**: "AI can do X" sometimes means "AI can do X in environments we've carefully prepared."

# Floridi: How AI Succeeds Differently Than Humans

## Human Intelligence

- Flexible, general-purpose
- Works in novel situations
- Understands context and meaning
- Transfers knowledge across domains
- Limited by biology (speed, memory)

## AI Agency

- Narrow, task-specific
- Exploits statistical regularities
- No semantic understanding
- Brittle outside training distribution
- Unlimited by biology (speed, scale)

**Floridi's point**: AI doesn't solve problems the way we do. It exploits patterns in data that may not reflect causal understanding. This is why AI can beat humans at chess while failing at tasks toddlers find easy.

**Neither better nor worse**—just *different*. Understanding this difference is crucial for knowing when to trust AI and when not to.

# Floridi: Ethical Implications of Agency Without Intelligence

## Two Risks of Misunderstanding AI

1. **Overestimating AI**: Attributing understanding, judgment, or moral agency to systems that have none—leading to misplaced trust and abdicated responsibility.
2. **Underestimating AI**: Dismissing AI as "just a tool" and ignoring its real capacity to affect lives, shape choices, and cause harm at scale.

**Floridi's prescription**:

- Treat AI as powerful but not intelligent—capable agents, not artificial minds.
- Human responsibility is never delegated to machines, even if tasks are.
- Focus on impacts and outcomes rather than metaphysical questions about machine minds.

**Key question**: If AI has agency but not intelligence, who is morally responsible for what it does?

# Other Modern Views

| Theorist | Core Definition | Key Insight |
|---|---|---|
| Russell | Rational agency | Focus on goal-directed behavior and decision-making |
| Floridi | Agency without intelligence | Decouple effective action from understanding |
| Bostrom | Optimization power | Focus on capability, control, and existential risk |
| Mitchell | Analogy and abstraction | Intelligence requires flexible conceptual transfer |
| Marcus | Hybrid neuro-symbolic | Deep learning alone is insufficient; need structured reasoning |
| LeCun | World models | True intelligence requires predictive models of reality |

No consensus exists, but there's a movement away from human-likeness as the primary criterion toward capability-focused definitions.

# You Be the Judge: Does It Understand?

A system holds a forty-minute conversation about a novel it was never explicitly taught. It picks up a joke, corrects itself, and asks a question nobody asked it to ask.

**Four benches. Each must rule, and each must state its test:**

- **Turing**: is there any behavioral question left that would separate it from a person?
- **Searle**: is there anything here but symbol manipulation, and does the room understand Chinese?
- **Dennett**: if it does everything understanding does, what is the leftover thing you are asking about?
- **Floridi**: why assume it needs to understand at all in order to **act** on the world?

**The tiebreaker**

Which bench's test could you **actually run** on a system in front of you tomorrow? A criterion nobody can apply is not doing the work you think it is.

# Comparing Definitions—What Counts as AI?

| System | Turing Test | Searle (Strong) | Russell (Agent) | Floridi (Agency) |
|---|---|---|---|---|
| Chess engine | No | No | Yes | Yes |
| COMPAS (recidivism) | No | No | Yes | Yes |
| Self-driving car | No | No | Yes | Yes |
| ChatGPT | Arguably | No | Yes | Yes |
| YouTube algorithm | No | No | Yes | Yes |
| Human-level AGI | Yes | Unclear | Yes | Yes |

**Key insight**: Your definition determines what you're asking when you ask "Is this AI?"

By the Turing Test, most current systems aren't AI. By agent-based definitions, almost all of them are. This matters for how we think about risks, benefits, and regulation.

# Lecture Outline

Defining AI: Definitional Challenges

AI Through History: From Logic to Learning

Can Machines Think? Classical Philosophical Debates

Contemporary Definitions and AI Approaches

Why AI Definitions Matter: Implications and Ethics

# Implications for AI Research

How we define AI shapes what researchers try to build:

- **Turing Test focus** → Human-like conversation, deception
- **Symbolic AI focus** → Logic, knowledge representation, reasoning
- **Agent-based focus** → Decision-making, optimization, planning
- **Connectionist focus** → Learning from data, neural architectures

**Funding follows definitions**: What counts as "real AI" gets resources. The 1980s expert systems boom and bust shows how definitions shape entire research programs.

**Discussion**

What *should* the goal of AI research be? Human-like intelligence? Useful capabilities? Something else entirely?

# Implications for AI Ethics and Safety

| If AI is... | Then key ethical concerns are... |
|---|---|
| A tool | Misuse by humans, bias in design, job displacement, accountability |
| A rational agent | Alignment with human values, control problems, unintended consequences |
| Potentially conscious | Moral status, rights, preventing suffering, exploitation |
| "Just software" | Only human responsibility matters; no special AI ethics needed |

**Examples**:

- If COMPAS is "just software," only its human designers bear responsibility.
- If a self-driving car is an "agent," who is responsible when it crashes?
- If an LLM might be conscious, do we owe it moral consideration?

How we define AI shapes how we regulate it, deploy it, and hold people accountable for its effects.

**Discussion**

Of everything we have covered today, which **one** disagreement would you most want settled before the next capable system ships?

# Key Concepts i

**Ada Lovelace (1815–1852)** argued machines can only do what they are programmed to do—they cannot *originate* anything. The earliest formal objection to machine intelligence.

**Alan Turing (1912–1954)** proposed the *Turing Test*: if a machine is indistinguishable from a human in conversation, call it intelligent—sidestepping whether machines "really" think.

**John Searle** proposed the *Chinese Room*: a system can pass the Turing Test by following rules without genuine understanding. Syntax is not semantics.

**Narrow AI** excels at specific tasks; **general AI (AGI)** would match human flexibility across all domains. Most real systems are narrow.

**Weak AI** simulates intelligent behavior without possessing a mind; **strong AI** would have genuine mental states—a distinction with implications for moral status.

**Symbolic AI (GOFAI)** represents knowledge as explicit rules and logic; **connectionist AI** (neural networks) learns patterns from data—the dominant paradigm today.

**Tesler's Theorem / moving goalposts** : "AI is whatever hasn't been done yet." Once a capability is achieved, it gets reclassified as mere computation.

# Key Concepts ii

**Dartmouth Conference (1956)** is where McCarthy, Minsky, and others named and launched "artificial intelligence"—with optimistic timelines that proved wildly premature.

**Discussion**

How would *you* define artificial intelligence—and does your definition commit you to any position on Lovelace, Turing, or Searle?

# References i