

# AI in Context: From Human to Machine Intelligence

---

[chris.wiggins@columbia.edu](mailto:chris.wiggins@columbia.edu)

Fall 2025 - Lecture 1 of 4

# Course Introduction

---

## Module goals

- Historical context/understanding of intelligence (human/AI)
- Modern goals/definition/evaluation of artificial intelligence
- Critical examination of dynamics of "define and design"
- Implications for society and technology

## Core Questions

1. What does it mean to define (and measure) "intelligence"?
2. How do historical patterns repeat in modern AI evaluation?
3. What are the ethical implications of reducing complex capabilities to numbers?
4. Is defining + designing intelligence what we *should* be doing?

# Today's Agenda

- Module overview and lecture series structure
- Course philosophy and responsibilities
- Instructor background
- The power of names in AI
- (Why history matters for understanding present)
- History: The measurement of intelligence: Spearman to benchmarks
- Critical perspectives and the path forward
- Hands-on notebook exploration

# Intelligence module: Overview (1)

## Lecture 1: The Birth of Quantifying Intelligence (Today!)

- Spearman's g-factor (1904) and its legacy
- PCA and dimensionality reduction
- AI benchmarks and leaderboard culture
- Critical analysis of quantification

## Intelligence module: Overview (2)

### Lecture 2: AI as How We Think We Think (1900s-1950s)

- Cognitive revolution: behaviorism to computation
- Turing's imitation game
- Cybernetics and feedback loops
- Early neural networks

# Intelligence module: Overview (3)

## Lecture 3: AI 1.0 - The Era of Rules (1960s-1980s)

- Expert systems and knowledge encoding
- Logic and reasoning (GOFAI)
- The AI winters
- Knowledge representation

# Intelligence module: Overview (4)

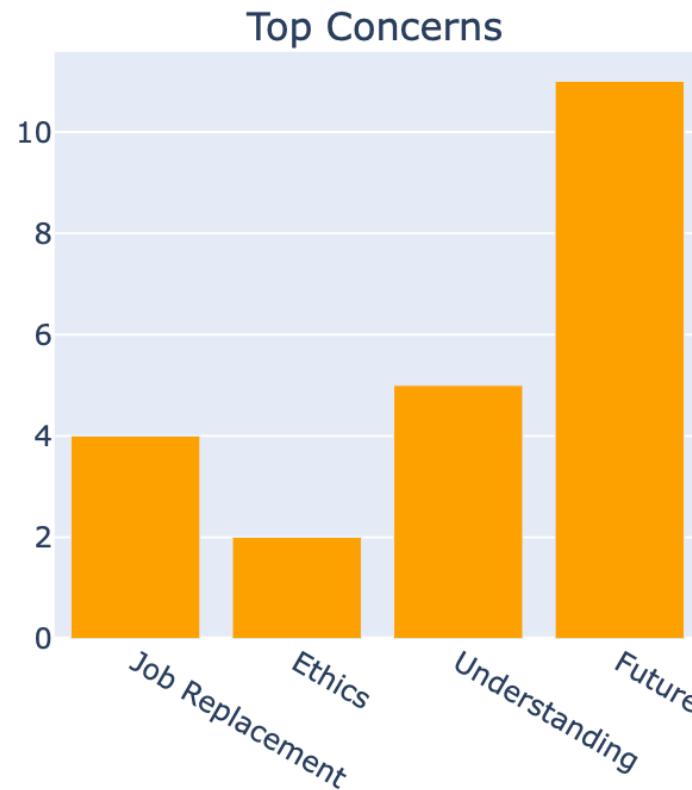
## Lecture 4: AI 2.0 - The Data Revolution (1990s-2025)

- Statistical turn: from rules to patterns
- Deep learning revolution
- Large language models
- Future directions: AGI and alignment

# Background & Perspective

---

## What You Want to Learn



# Background & Perspective

---

## What You Want to Learn

Based on pre-course surveys:

- Understanding AI capabilities and limitations
- Historical context for current developments
- Hands-on experience with AI tools
- Critical thinking about societal impact

# Instructor Background

## Education & Experience

- Applied mathematics PhD (machine learning focus)
- 20+ years in data science and AI research
- Chief Data Scientist industry experience

## Example works

- *How Data Happened* (2023): Technical + social history (cf. Ch 4 on Intelligence)
- *Data Science in Context* (2021): Foundations + opportunities
- Papers on ML fairness, interpretability, applications

## Principle 1: Kranzberg's First Law (1985):

*"Technology is neither good nor bad; nor is it neutral... The same technology can have quite different results when introduced into different contexts or under different circumstances."*

**Key Insight:** AI's impact depends on:

- Who builds it
- Who uses it
- What problems it addresses
- Which values it embodies

## Principle 2: Capabilities have politics

Phillip Rogaway (2015):

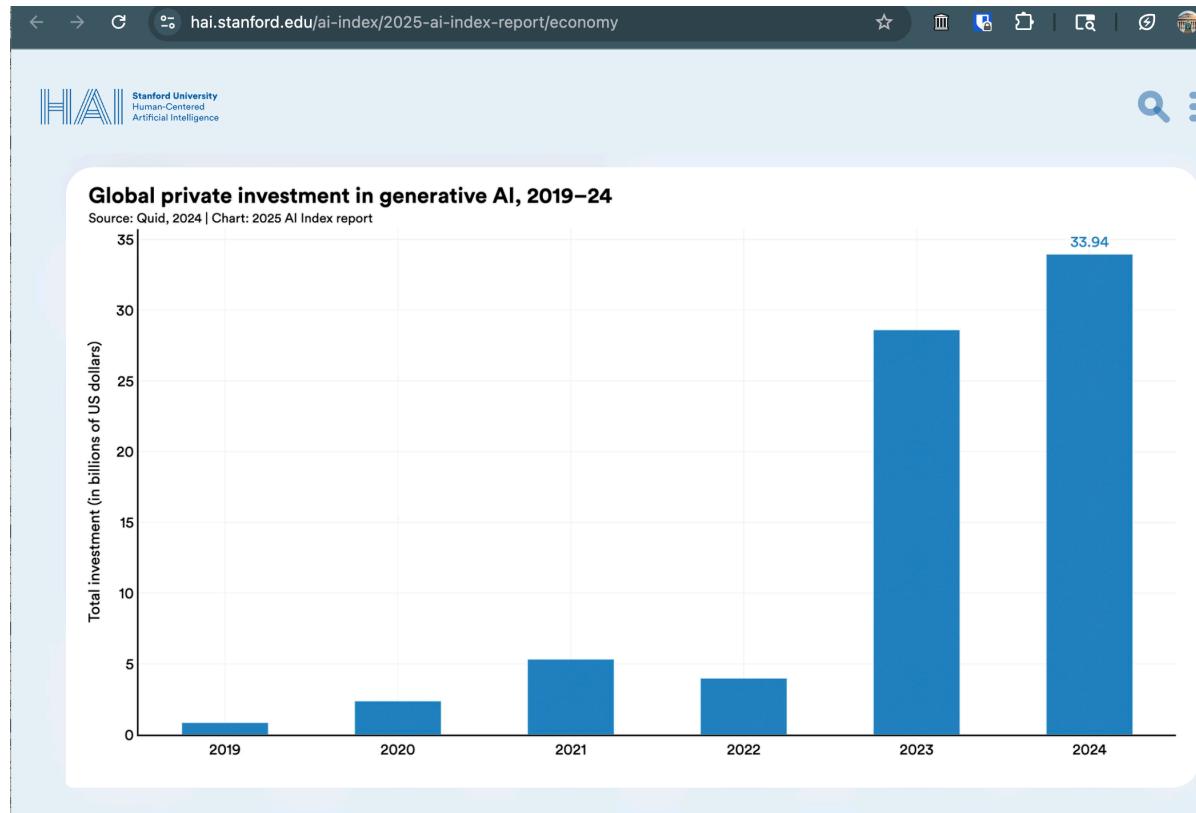
*"Cryptography rearranges power: it configures who can do what, from what. This makes cryptography an inherently political tool, and it confers on the field an intrinsically moral dimension."*

Replace "cryptography" with "AI" — the statement remains true.

# **What we talk about when we talk about AI...**

---

# ...and how did we get here?



## McCarthy's Coinage (1955)

AI, **definition** (70 yrs ago):

*"Every aspect of learning or intelligence can be precisely described so a machine can simulate it."*

*— Dartmouth Conference Proposal on AI*

## An aspiration, not a method (ch 7, HDH)

1. “Automatic Computers” (programming languages)
2. “How Can a Computer be Programmed to Use a Language” (natural language processing)
3. “Neuron Nets” (neural nets and deep learning)
4. “Theory of the Size of a Calculation” (computational complexity)
5. “Self-improvement” (machine learning)
6. “Abstractions” (feature engineering)
7. “Randomness and Creativity” (Monte Carlo methods including stochastic learning).

The term “artificial intelligence,” in 1955, was an aspiration rather than a commitment to one method. AI, in this broad

# AI...found a way (diverging methods)

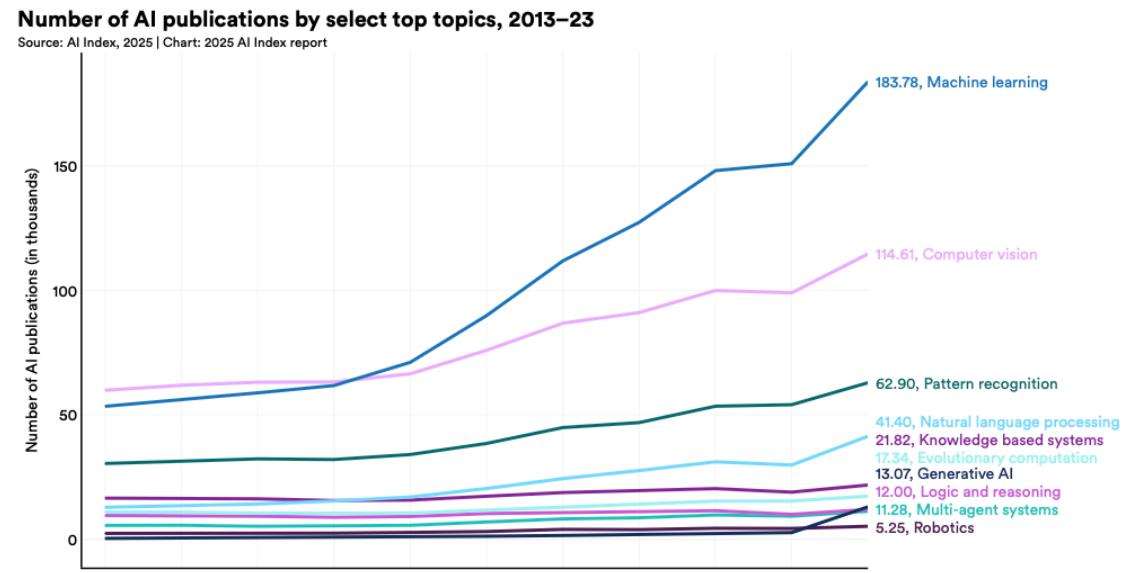


Figure 1.1.10<sup>a</sup>

# JCM and the 1955 Coining of AI



## Later Admission:

*"I invented the term artificial intelligence because... we were trying to get money for a summer study."*

— John McCarthy, Lighthill Debate (1973)

## **Q1/post-1955: Why "AI" Succeeded**

The term succeeded because it:

- **Captured imagination**
- **Secured funding**
- **Set expectations**
- **Created identity**
  - for field
  - esp. among Homo Sapiens
  - JCM reflected on this definition in [2007](#)

## Alternative Histories

What if we'd chosen different terms?

- "Augmented Reasoning"
- "Computational Cognition"
- "Synthetic Problem-Solving"
- "Machine Understanding"

Would the field have developed differently?

## **Q2/pre-1955: What did "intelligence" even mean to JCM et al?**

*"Every aspect of learning or intelligence can be precisely described so a machine can simulate it."*

To answer *that* we will use history...

# Why History Matters

---

## Pattern Recognition

### 1. Common Patterns

- Hype cycles repeat (1960s, 1980s, 2010s, 2020s)
- Technical barriers recur (data, compute, algorithms)
- Social concerns persist (jobs, bias, control)

# Making the Present Strange

## 2. Alternative Perspectives

- Current approaches aren't inevitable
- Alternative paths were/are possible
- Today's "obvious" was yesterday's "impossible"

# Disrupting Determinism

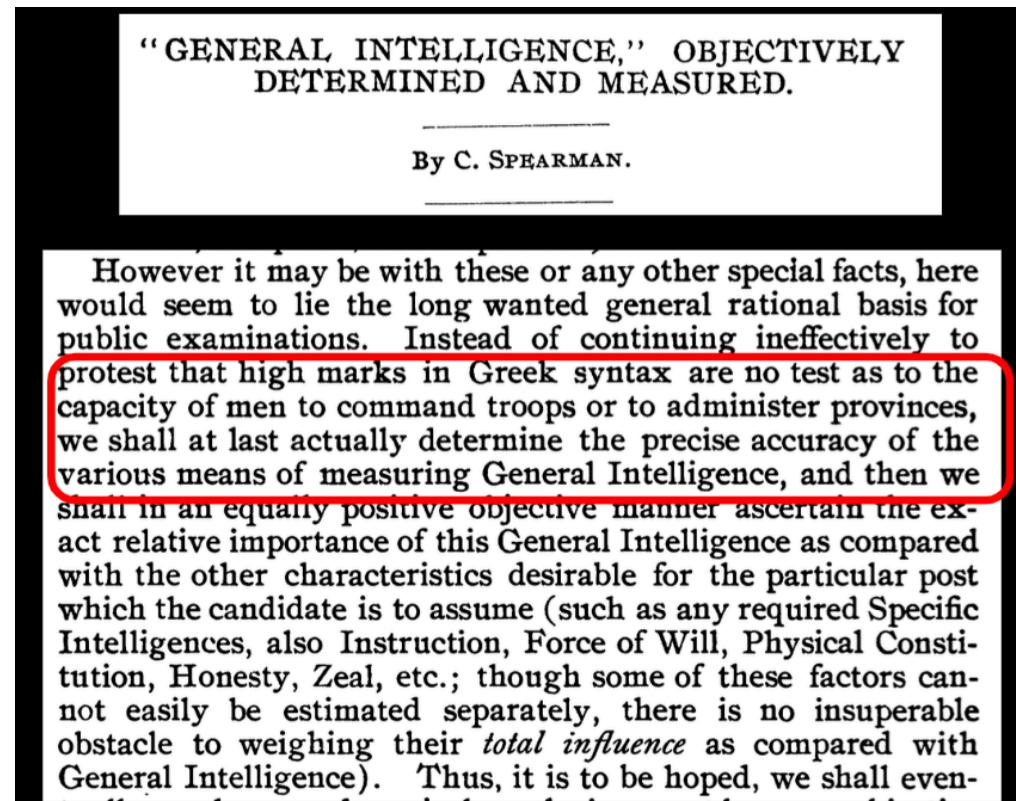
## 3. Agency and Choice

- Technology doesn't develop in straight lines
- Social choices shape technical directions
- Multiple futures remain possible

# The Measurement of Intelligence

---

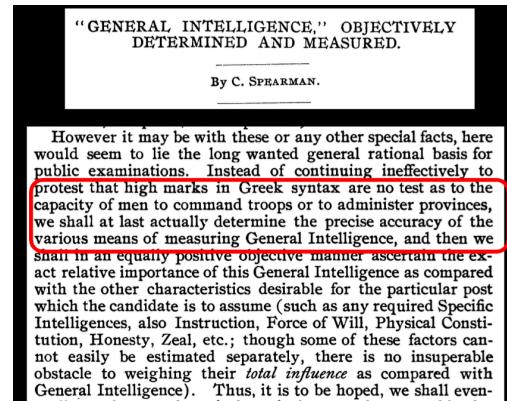
## Spearman's Revolution (1904)



# The Measurement of Intelligence

---

## Spearman's Revolution (1904)



"'General Intelligence,' Objectively Determined and Measured"

- Published in American Journal of Psychology
- Introduced the **g-factor** concept
- Novel statistical techniques (factor analysis)
- Claimed "objective" measurement

# From many grades, 1 number

*High Class Preparatory School for Boys.*

A. Original Data.

Age.	Discriminative Threshold.			Place in School ( <i>before modification to eliminate Age</i> ).			
	Pitch From Ex.Ser. IV.	Light 1:200	Weight 1:200	Classics	French	English	Mathem.
I0 9	50	10	4	16	19	10	7
I2 4	3	10	6	5	6	6	5
II I	10	10	9	13	11	11	13
I0 II	> 60	10	9	22	23	22	22
I3 7	4	12	5	1	1	1	2
I2 6	2	10	10	4	2	2	1
I0 4	4	10	11	12	14	13	18
9 5	20	10	11	23	22	23	23
I2 0	11	10	12	8	8	15	15
I0 2	11	12	11	3	5	4	4
II 2	24	14	10	7	7	7	6

# From many "attributes", 1 number

*Village School, 24 Oldest Children.*

## A. Original Data.

Sex.	Age.		Discriminative Threshold.			Intellectual Rank.	
	Years	Months	Pitch 1/3 v. d.	Light 1:200	Weight 1:200	Common Sense out of School. (A)	Cleverness in School. (B)
f	II	6	8	4	4	6	5
m	I2	II	15	3	4	11	7
f	I2	8	14	6	4	16	10
f	I3	8	13	4	9	1	1
m	II	4	5	14	7	3	2
f	II	II	25	7	4	10	14
f	II	3	10	19	8	8	19
f	I3	1	10	12	10	2	4
m	I2	5	18	11	9	5	6
m	I2	7	14	30	7	21	22
f	I2	8	60	3	10	12	9
f	I2	10	22	12	12	12	12

# Spearman's Core Ideas

## Four Key Insights

- 1. Low-D Manifold:** Cognitive tasks correlate
- 2. Hidden Factor:** Single underlying intelligence
- 3. Mathematical Extraction:** Statistics reveal factors
- 4. Practical Application:** Objective rankings

# The Statistical Innovation

## Spearman's Process:

Many test scores → Correlation matrix →  
Factor analysis → Single g-factor

# Spearman's findings

Activity.	Correlation with Gen. Intell.	Ratio of the common factor to the specific factor.	
Classics,	0.99	99	to 1
Common Sense,	0.98	96	4
Pitch Dis.,	0.94	89	11
French,	0.92	84	16
Cleverness, <sup>3</sup>	0.90	81	19
English,	0.90	81	19
Mathematics, <sup>4</sup>	0.86	74	26
Pitch Dis. among the uncultured, <sup>5</sup>	0.72	52	48
Music,	0.70	49	51
Light Dis., <sup>5</sup>	0.57	32	68
Weight Dis., <sup>5</sup>	0.44	19	81

## Spearman's findings

*"In the above Hierarchy one of the most noticeable features is the high position of languages; to myself, at any rate, it was no small surprise to find Classics and even French placed unequivocally above English."*

## Spearman's findings

*"Instead of continuing ineffectively to protest that high marks in Greek syntax are no test as to the capacity of men to command troops or to administer provinces, we shall at last actually determine the precise accuracy of the various means of measuring General Intelligence, and then we shall in an equally positive objective manner ascertain the exact relative importance of this General Intelligence as compared with the other character" (1904, p. 277).*

- how does this rearrange power?

## Spearman's findings

*Conclusion. On the whole, then, we reach the profoundly important conclusion that there really exists a something that we may provisionally term "General Sensory Discrimination" and similarly a "general Intelligence," and further the functional correspondence between these two is not appreciably less than absolute.*

- how does this rearrange power?

# Hands-On: Notebook Exploration

---

## Today's Interactive Notebook

### What You'll Explore:

- 2D/3D PCA visualizations with real data
- Synthetic experiments showing dimensionality reduction
- Student grade analysis revealing hidden patterns
- AI benchmark evolution over time

Link: [Colab Notebook](#)

## Modern Equivalent (PCA)

PCT today..

High-dimensional data → Covariance matrix →  
Eigendecomposition → Principal components

Same mathematical intuition, 120 years apart!

# Post-Spearman, pre-JCM Intelligence: "IQ"

## Intelligence testing and eugenics [ edit ]

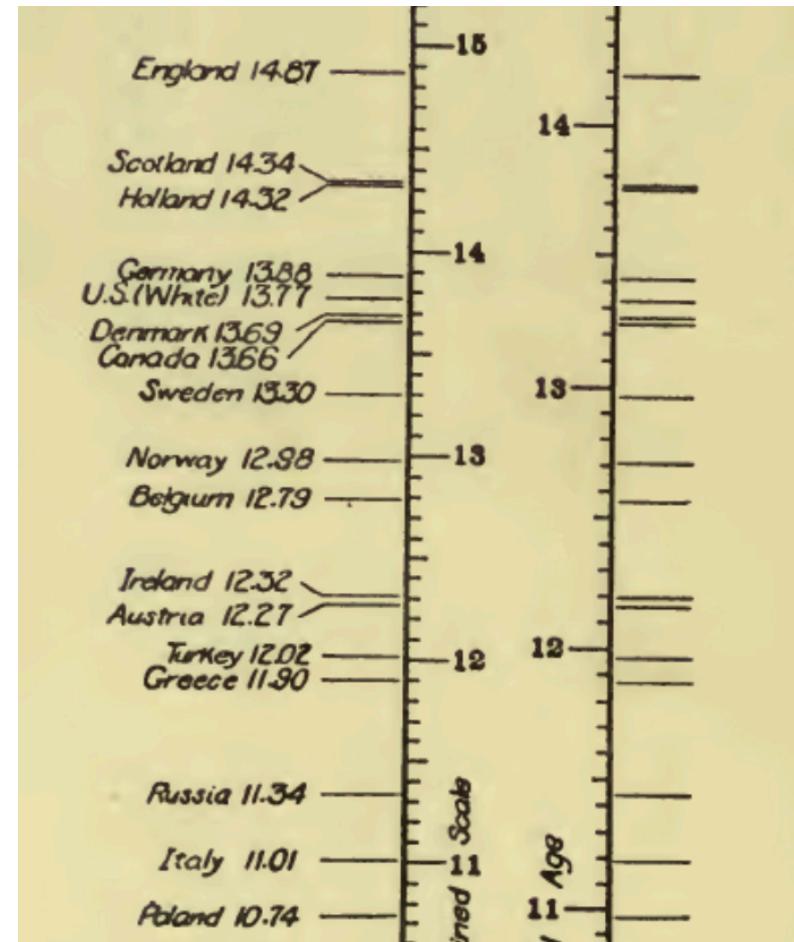
---

In 1917, Yerkes served as president of the [American Psychological Association](#) (APA). Under his influence, the APA began several programs devoted to the war effort in [World War I](#). As chairman of the Committee on the [Psychological Examination](#) of Recruits, he developed the [Army Alpha](#) and [Army Beta](#) Intelligence Tests, the first nonverbal group tests, which were given to over 1 million United States soldiers during the war.

Although Yerkes claimed that the tests measured native intelligence, and not education or training, this claim is difficult to sustain in the face of the questions themselves. Question 18 of Alpha Test 8 reads: "Velvet Joe appears in advertisements of ... (tooth powder)(dry goods) ([tobacco](#))([soap](#))."<sup>[6]</sup>

Reduction to 1 number, 1 ordering

# Post-Spearman, pre-JCM Intelligence: "IQ"



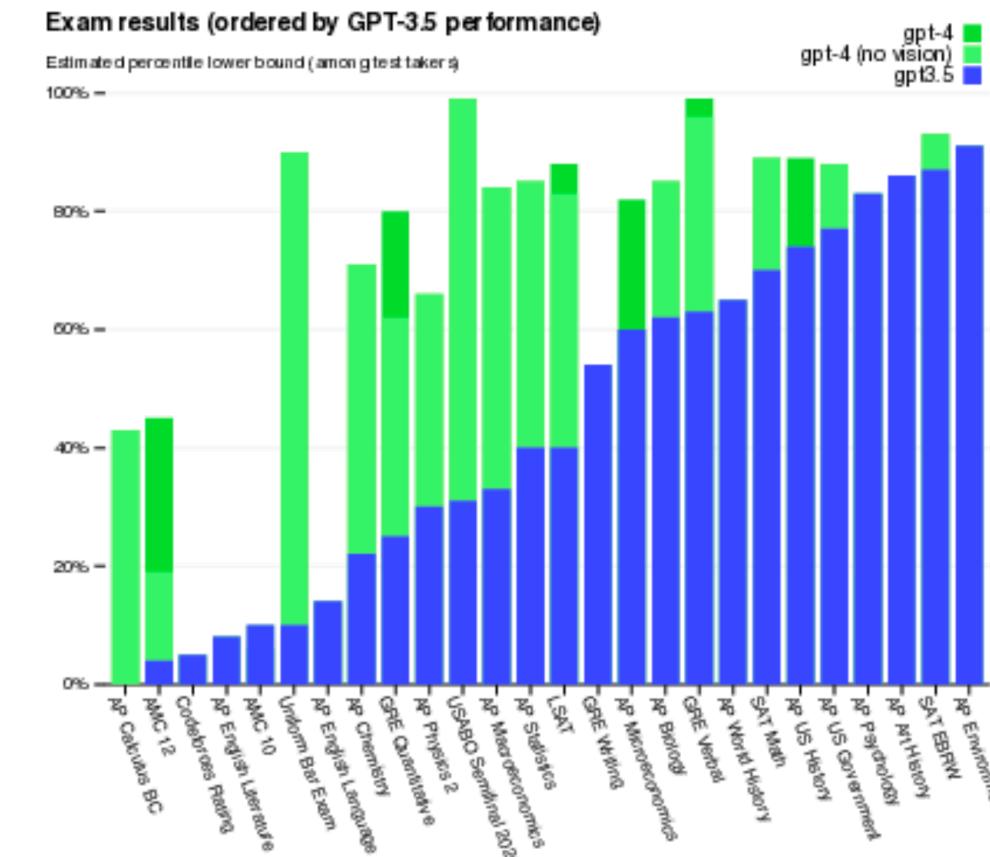
Reduction to 1 number, 1 ordering

## **But Wait... Is This Obvious for AI?**

If human intelligence is a smooth, low-dimensional manifold...

Shouldn't AI intelligence work the same way?

# Should AI... take "tests"?



see also 2023 "Sparks of Artificial General Intelligence: Early experiments with GPT-4"

# Moravec's Paradox (1988)

## The Historical Foundation



[https://en.wikipedia.org/wiki/Moravec%27s\\_paradox](https://en.wikipedia.org/wiki/Moravec%27s_paradox)

As Moravec writes:

Encoded in the large, highly evolved sensory and motor portions of the human brain is a billion years of experience about the nature of the world and how to survive in it. The deliberate process we call reasoning is, I believe, the thinnest veneer of human thought, effective only because it is supported by this much older and much more powerful, though usually unconscious, sensorimotor knowledge. We are all prodigious olympians in perceptual and motor areas, so good that we make the difficult look easy. Abstract thought, though, is a new trick, perhaps less than 100 thousand years old. We have not yet mastered it. It is not all that intrinsically difficult; it just seems so when we do it.<sup>[8]</sup>

# Moravec's Key Insight

As Moravec writes:

*"Encoded in the large, highly evolved sensory and motor portions of the human brain is a billion years of experience about the nature of the world and how to survive in it. The deliberate process we call reasoning is, I believe, the thinnest veneer of human thought, effective only because it is supported by this much older and much more powerful, though usually unconscious, sensorimotor knowledge."*

**Hard for humans = Easy for AI**

**Easy for humans = Hard for AI**

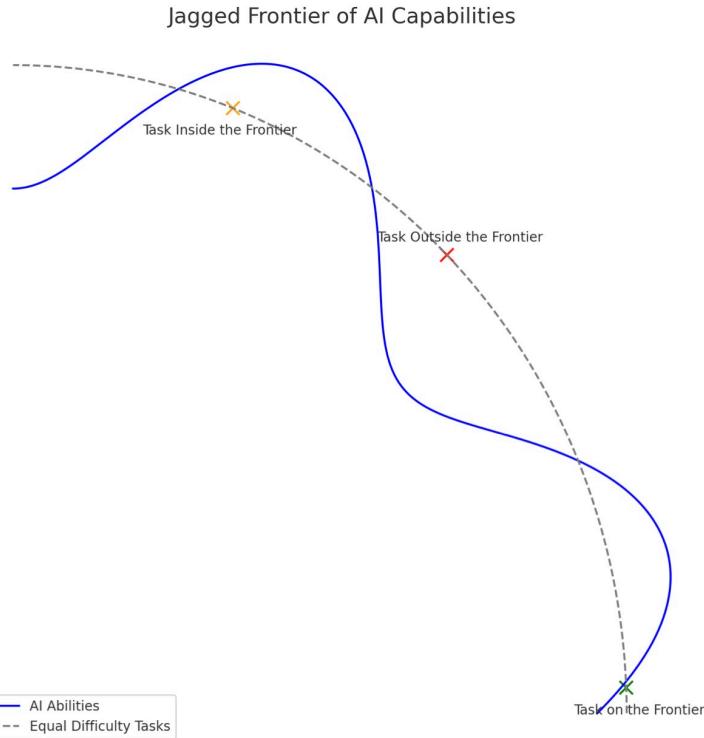
# The Jagged Frontier of AI (2023)

Modern Validation of Moravec's Paradox

## Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality\*

Fabrizio Dell'Acqua<sup>1</sup>, Edward McFowland III<sup>1</sup>, Ethan Mollick<sup>2</sup>, Hila Lifshitz-Assaf<sup>1,3</sup>, Katherine C. Kellogg<sup>4</sup>, Saran Rajendran<sup>5</sup>, Lisa Krayer<sup>5</sup>, François Candelier<sup>5</sup>, and Karim R. Lakhani<sup>1</sup>

# Visualizing the Jagged Frontier



**The frontier is irregular and unpredictable** - some "simple" tasks are outside AI capabilities while "complex" ones are inside.

# Karpathy's "Jagged Intelligence"

The tweet is from Andrej Karpathy (@karpathy). It features a profile picture of a colorful abstract painting. The text reads:

**Jagged Intelligence**

The word I came up with to describe the (strange, unintuitive) fact that state of the art LLMs can both perform extremely impressive tasks (e.g. solve complex math problems) while simultaneously struggle with some very dumb problems.

E.g. example from two days ago - which number is bigger, 9.11 or 9.9?  
Wrong.  
[x.com/karpathy/status/...](https://x.com/karpathy/status/...)

Below the tweet is a screenshot of a AI interface showing a conversation:

A Is 9.11 bigger than 9.9?

Yes, 9.11 is bigger than 9.9.

Copy ⌂ Retry ⌂

**Perfect example:** AI can solve complex math but fails at "Is 9.11 bigger than 9.9?"

# Jagged vs Human Intelligence

## Human Intelligence (Spearman's model)

- Smooth, correlated abilities
- If you can do A, you can probably do simpler task B
- Single g-factor explains performance

## AI Intelligence (Jagged model)

- Unpredictable, non-correlated abilities
- Excelling at A tells us nothing about ability to do B
- No single factor explains capabilities

## **Why This Breaks Everything**

**Spearman's Fundamental Assumption:** Abilities correlate smoothly

**AI Reality:** The frontier is invisible and jagged

**Implication:** Traditional psychometric approaches fail for AI

**This challenges our entire framework for measuring intelligence**

# From Human g-factor to AI Benchmarks

---

## The Striking Parallel (Part 1)

1904: Human Intelligence	2025: AI Intelligence
Multiple test scores	Multiple benchmark scores
Reduce to single g-factor	Reduce to leaderboard rank
IQ tests emerge	MMLU, HumanEval emerge

## The Striking Parallel (Part 2)

1904: Human Intelligence	2025: AI Intelligence
Testing industry	AI evaluation industry
College admissions (SAT)	Company hiring (benchmarks)
Cultural bias debates	Dataset bias debates

## The Striking Parallel (Part 3)

1904: Human Intelligence	2025: AI Intelligence
Teaching to the test	Training to benchmarks
Standardization pressure	Benchmark optimization
"Scientific" ranking	"Objective" leaderboards

# Language Understanding Benchmarks

## Current State-of-the-Art (2025)

- MMLU: 92% (GPT-4o)
- HellaSwag: 95% (Claude 3.5)
- TruthfulQA: 68% (Gemini Ultra)

# Reasoning & Code Benchmarks

## Reasoning & Math

- GSM8K: 94% (GPT-4)
- MATH: 76% (Claude 3.5)
- GPQA: 59% (Gemini Ultra)

# Code & Multimodal Benchmarks

## Code Generation

- HumanEval: 92% (Claude 3.5 Sonnet)
- MBPP: 89% (GPT-4)
- SWE-bench: 24% (GPT-4 + tools)

## Multimodal

- MMBench: 83% (GPT-4V)

# "Leaderboard" meme persists



The screenshot shows a table of AI models and their performance metrics. The columns are labeled: Model, Average, IFEval, BBH, MATH, and LVi. The table lists 11 models, each with a small icon and a link to the model's page. The models are sorted by Average score.

T	Model	Average	IFEval	BBH	MATH	LVi
◆	<a href="#">dnhkng/RYS-XLarge</a>	44.75	79.96	58.77	38.97	
◆	<a href="#">MaziyarPanahi/calme-2.1-rys-78b</a>	44.14	81.36	59.47	36.4	
◆	<a href="#">MaziyarPanahi/calme-2.2-rys-78b</a>	43.92	79.86	59.27	37.92	
◆	<a href="#">MaziyarPanahi/calme-2.1-qwen2-72b</a>	43.61	81.63	57.33	36.03	
◆	<a href="#">MaziyarPanahi/calme-2.2-qwen2-72b</a>	43.4	80.08	56.8	41.16	
◆	<a href="#">dfuxman/Qwen2-72B-Oppo-v0.1</a>	43.32	78.8	57.41	35.42	
◆	<a href="#">Qwen/Qwen2-72B-Instruct</a>	42.49	79.89	57.48	35.12	
◆	<a href="#">abacusai/Dracarys-72B-Instruct</a>	42.37	78.56	56.94	33.61	
◆	<a href="#">VAGOsolutions/Llama-3.1-SauerkrautLM-70B-Instruct</a>	42.24	86.56	57.24	29.91	
◆	<a href="#">alpindale/magnum-72b-v1</a>	42.17	76.06	57.65	35.27	
◆	<a href="#">meta_llama/Meta-Llama-3.1-70B-Instruct</a>	41.74	86.69	55.93	28.02	
◆	<a href="#">dnhkng/RYS-Llama3.1-Large</a>	41.6	84.92	55.41	28.4	

Risks: (i) gaming the metric (Flynn effect) (ii) train/test corruption (iii) jagged

# Performance increasing

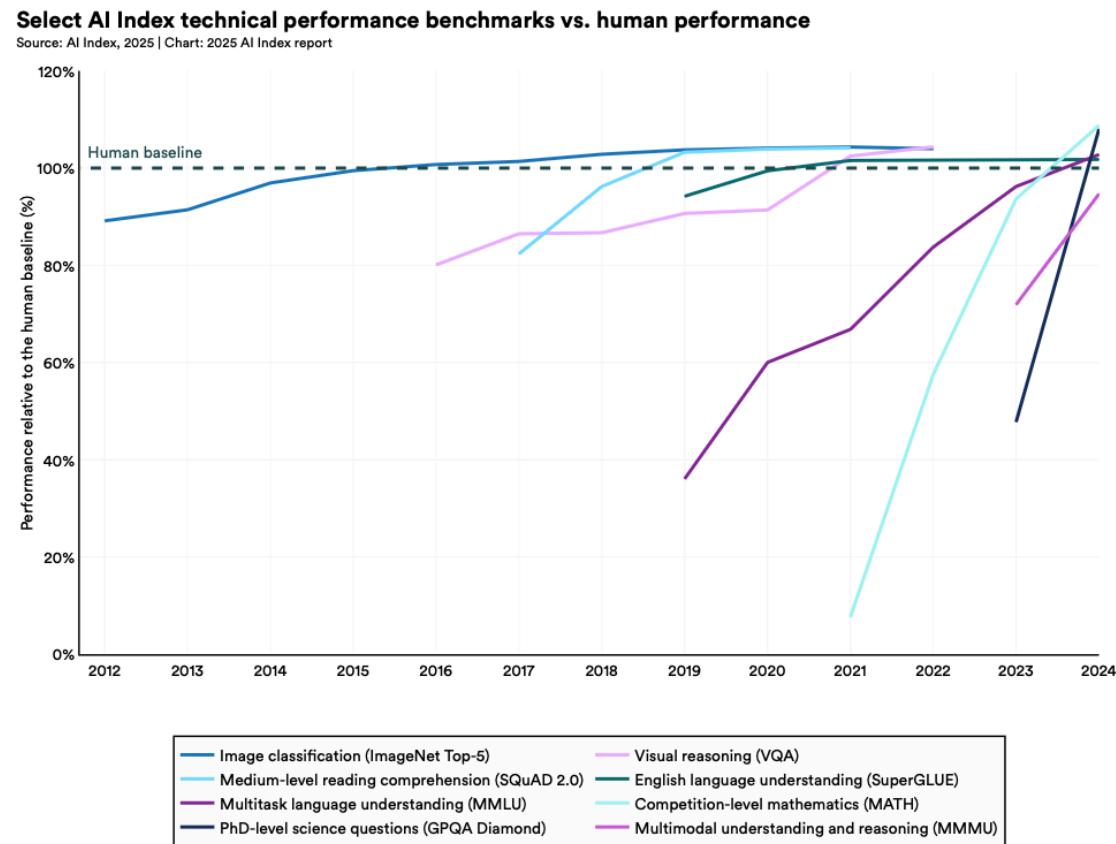


Figure 2.1.33<sup>2</sup>

# PCA: The Mathematical Magic

---

## What PCA Does

### The Process

**Input:** High-dimensional correlated data

**Process:** Find directions of maximum variance

**Output:** Orthogonal components ranked by importance

**Goal:** Dimensionality reduction with minimal information loss

# PCA for Student Grades

## Example Application:

- Input: Math, Science, History, Art scores
- PC1: Often "general academic ability"
- PC2: Often "STEM vs Humanities"
- Insight: Few components explain most variation

# PCA for AI Benchmarks

## Hypothetical Application:

- Input: MMLU, HumanEval, GSM8K scores
- PC1: "General capability"?
- PC2: "Language vs Code"?
- Question: Meaningful or reductionist?

# The Irony

## Historical Continuity

**We use 120-year-old psychometric techniques to evaluate cutting-edge AI!**

- Same mathematical foundations
- Same reductionist impulses
- Same social implications
- Different technological context

# Critical Perspectives

---

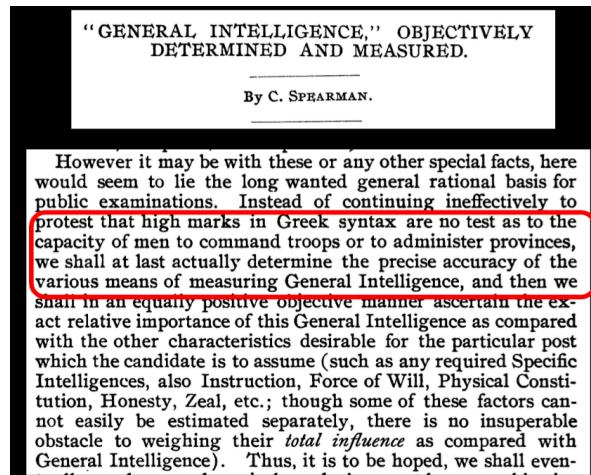
## Stephen Jay Gould on Reification

The Warning:

*"We give the word 'intelligence' to this complex set of capabilities. This symbol is then reified, achieving dubious status as a unitary thing."*

— *The Mismeasure of Man* (1981)

# Technical Issues with AI Benchmarks



## Current Problems

- Data contamination (test in training)
- Distribution shift (lab vs. real world)
- Spurious correlations
- Gaming and overfitting

# Assignment & Resources

---

## This Week's Assignment

Critical Reflection (500 words, due next week)

Choose ONE prompt:

1. **Historical Parallel:** How does IQ testing inform AI evaluation?
2. **Design Challenge:** Create a new AI benchmark avoiding pitfalls
3. **Philosophical Question:** Is "general intelligence" misguided?
4. **Practical Application:** How might PCA be misused in AI applications?

# Looking Ahead

---

## Next Week Preview

### Lecture 2: "AI as How We Think We Think"

- The computational metaphor of mind
- Turing's imitation game implications
- Early neural networks

## Questions?