The Future of Operations Analysis in an Era of Artificial Intelligence: Work, Automation, and Expertise

David L. Alderson, Ph.D.

Professor and Chair, Operations Research Department

Naval Postgraduate School

MORS AI Workshop, April 20-23, 2026
Carnegie Mellon University, Pittsburgh

Disclaimers:
The views expressed are solely those of the author and do not reflect the official position of the U.S. Navy or DOD.

History Highlights

1909 Founded at U.S. Naval Academy

1951 Moved to Monterey, CA
Operations Analysis curriculum established

Operations Research Department:

- 43 faculty (16 tenure-track, 8 military)
- Resident and DL curricula in Operations Analysis, Logistics, Defense Systems Analysis
- Home of Data Science, Logistics at NPS

University statistics:

~500 faculty
~1400 resident students (incl. 170 international)
~700 Distance Learning students
10,000+ Exec Ed / Professional Dev students
Approx. $100M in sponsored research funding

IMMEDIATE IMPACT | FUTURE ADVANTAGE | ENDURING LEADERSHIP

NPS-NVIDIA PARTNERSHIP INITIATIVE

Innovative Education & Training

NPS AI FACTORY
DGX GB300 AI SUPERCOMPUTER

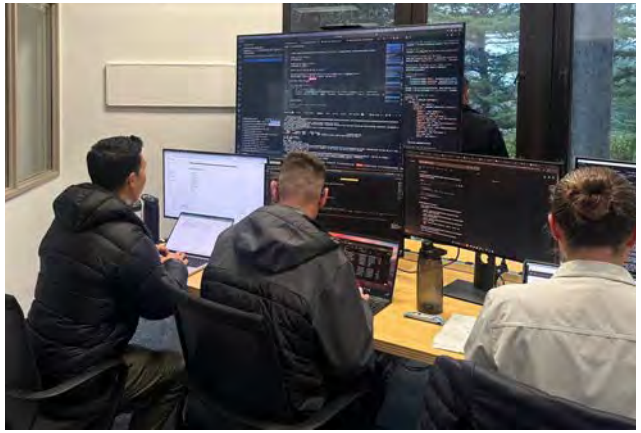
The DGX GB300 provides Navy, Marine Corps, and Special Operations forces access to a level of AI compute power that was previously available only in large commercial or national-scale facilities. This capability allows military teams to train and operate large, domain-specific models that can support maritime operations, expeditionary warfare, and special operations missions, including AI-enabled planning, ISR fusion, autonomous behaviors, and operational decision support.

The system's extreme compute density and memory capacity enable workloads that would otherwise be infeasible, such as training long-context reasoning models, running multi-agent simulations at scale, and integrating diverse operational data streams into a single analytic environment. This power is critical for exploring complex, highly-contested scenarios where speed, scale, and fidelity directly affect the operational relevance of the output or application.

By placing this level of capability within an academic and military research environment, the DGX GB300 allows NPS and its partners to stress, evaluate, and mature advanced AI concepts at operational scale. This ensures that AI systems intended for naval, expeditionary, and special operations use are developed with realism, engineering rigor, and a clear path toward transition and adoption.

System: Rack-scale AI factory (Grace Blackwell Ultra)
Compute: 72× Blackwell Ultra GPUs
+ 36× Grace CPUs (2,592 Arm cores)
Performance: Up to 1.44 EFLOPS FP4 (sparse)
1.08 EFLOPS (dense)
Memory: ~20 TB HBM; ~37 TB total fast system memory
Interconnect: NVLink fabric + 800 Gb/s InfiniBand (ConnectX-8)
Storage (NPS): ~0.5 PB high-performance shared AI storage

NVIDIA AI Tech Center at NPS



POINTS OF CONTACT:
Shared mailbox: AI@nps.edu
CAPT Mike Owen, USN
MAJ Neal MacDonald, USA

The NPS-OR Department is Hiring!

Non-tenure-track Professor of the Practice

Responsibilities include:

- **Teaching graduate-level courses** in operations research, naval and joint operations analysis, and systems analysis.
- **Designing and delivering executive seminars** focused on organizational risk and decision-making.
- **Liaising with U.S. Navy** Fleet-level commands, OPNAV staff, naval warfare development centers, Joint organizations, and defense industry partners.
- **Assisting faculty and students** with sponsored research programs in future force design, naval tactics, warfare analysis, and maritime decision aids for tactical and operational command and control.
- **Developing curricula, courses, executive seminars, research proposals, and thesis topics** that bring current and future Navy challenges to NPS students, faculty, and staff.
- **Assisting graduate-student researchers** pursuing master's or PhD degrees.

Eligibility: U.S. Citizens only

Check out our OR Hiring Webpage for the official announcement and additional details!

https://www.nps.edu/web/or/hiring

We are living in a Gold Rush!

The AI Gold Rush

Shane Greenstein, *Harvard Business School, Boston, MA, 02163, USA*

Large language models (LLMs) have overrun commercial markets, more like a tsunami than the normal technical wave of interest. The topic is everywhere—news stories, blogs, podcasts, start-up investments, analyst reports, hackathons, and government announcements. A virtual frenzy surrounds it.

If you possess a technical background, you might find this frenzy puzzling. The technological roots of LLMs go back many years. However, today's experience looks like more than the continuation of a preexisting trend. Something in the zeitgeist changed recently, making entrepreneurs and financiers rethink and shift the direction of investment. You might be tempted to call this an AI gold rush. If you are old enough to recall them, you might compare this frenzy to the gold rushes during the dot-com or PC booms.

information spreading—happened within a month of each other.

What about impatient economic actors? Gold miners overran Sutter's Mill that spring. By the following year's snowmelt, the area was packed with miners who believed they needed to stake a claim as soon as possible. Many panned for gold in the rivers. They were called "49ers."

Sadly, though the western slopes of the Sierra Nevada range contain some of the best deposits of gold in the world, most of those deposits do not have rivers passing over them. Most of those who panned for gold did not recover much and soon gave up.

Old-fashioned digging into the earth, which followed years after the 49ers, yielded much better rewards. This activity was expensive and slow, requiring the financing

IEEE Micro November/December 2023, pp. 126-128

Reminiscent of California Gold Rush of 1849

- **Hopes** among senior executives in business and government: *prospecting for a golden future where using AI will allow making decisions quickly, in the face of uncertainty, and with assurances of favorable outcomes.*
- **Fears** among information workers: *anyone not using AI or LLMs in their daily work is at risk of losing their livelihood, because they are left behind and/or out of a job.*

Fueling a massive rush of people and companies seeking new capabilities (and fortune) in developing and deploying AI-enabled technologies.

This is the not the first AI Gold Rush!

Two Heads Are Better than One: The Collaboration between AI and OR

HERBERT A. SIMON

Departments of Psychology and Computer Science
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

Originally delivered as the plenary address at the TIMS/ORSA Joint National Meeting, Miami, Florida, October 1986.

Published 1987. Interfaces 17(4):8-15. https://doi.org/10.1287/inte.17.4.8

Computer scientist, economist, cognitive scientist

Famous for "bounded rationality", "satisficing"

Turing Award (1975)

Nobel Prize in Economics (1978)

"Management science and operations research are a part of the great effort, often styled the Second Industrial Revolution, that is striving to understand and enhance intelligence. ***Joining hands with AI, management science and operations research can aspire to tackle every kind of problem-solving and decision-making task the human mind confronts****."*

This is the not the first AI Gold Rush!

The first AI Gold Rush (1980s) was similarly characterized by a desire for automation via algorithms.

It didn’t work out as intended.

Not because the algorithms weren’t capable.

But because of integration challenges that were not overcome.

Reification Fallacy

- AI is not a single thing
- Actually, a set of technologies (algorithms) of varying capability
- Each with strengths and weaknesses

- Treating AI as a “thing” we tend to ascribe
  - All of the strengths (utopians)
  - Or weaknesses (dystopians)
- As is often the case, the truth is somewhere in the middle

Magazine @1962
Warhol time capsule
Warhol museum 10/25

Courtesy
David
Woods

Today's Agenda

- What do [we] want from AI?
- What is the *work* of Military Operations Analysis?
- Automation
- Expertise
- What about LLMs?
- Challenges and Opportunities

Magazine @1962
Warhol time capsule
Warhol museum 10/25

Courtesy
David
Woods

What we want from AI

1. Speed

AI adoption is believed to accelerate decision-making and action.

- Industry: Faster, Better, Cheaper (FBC) Pressures
- Military: Observe – Orient – Decide – Act (OODA) Loop

> ***"future wars will be won by those who can 'see, understand, decide, and act faster,'*** *underscoring the necessity of integrating artificial intelligence (AI), electronic warfare, and space operations into military strategies, and 'finding ways to combine AI tools and human decision making that deliver a decision advantage.'"*
>
> Dec 10, 2024 SGL by Admiral Paparo; https://nps.edu/-/indopacom-commander-discusses-challenges

> *"**AI is not entirely new, but advancements in computing power and big data are transforming how we think about processes** — not just acquisition, but our daily operations… AI can significantly enhance the Joint Staff's ability to integrate and analyze global military operations, ultimately **enabling better, faster decisions**."* – the Joint Staff AI Lead (April 24, 2025)
>
> https://www.war.gov/News/News-Stories/Article/Article/4165279/defense-officials-outline-ais-strategic-role-in-national-security/

What we want from AI

1. Speed

AI adoption is believed to accelerate decision-making and action.

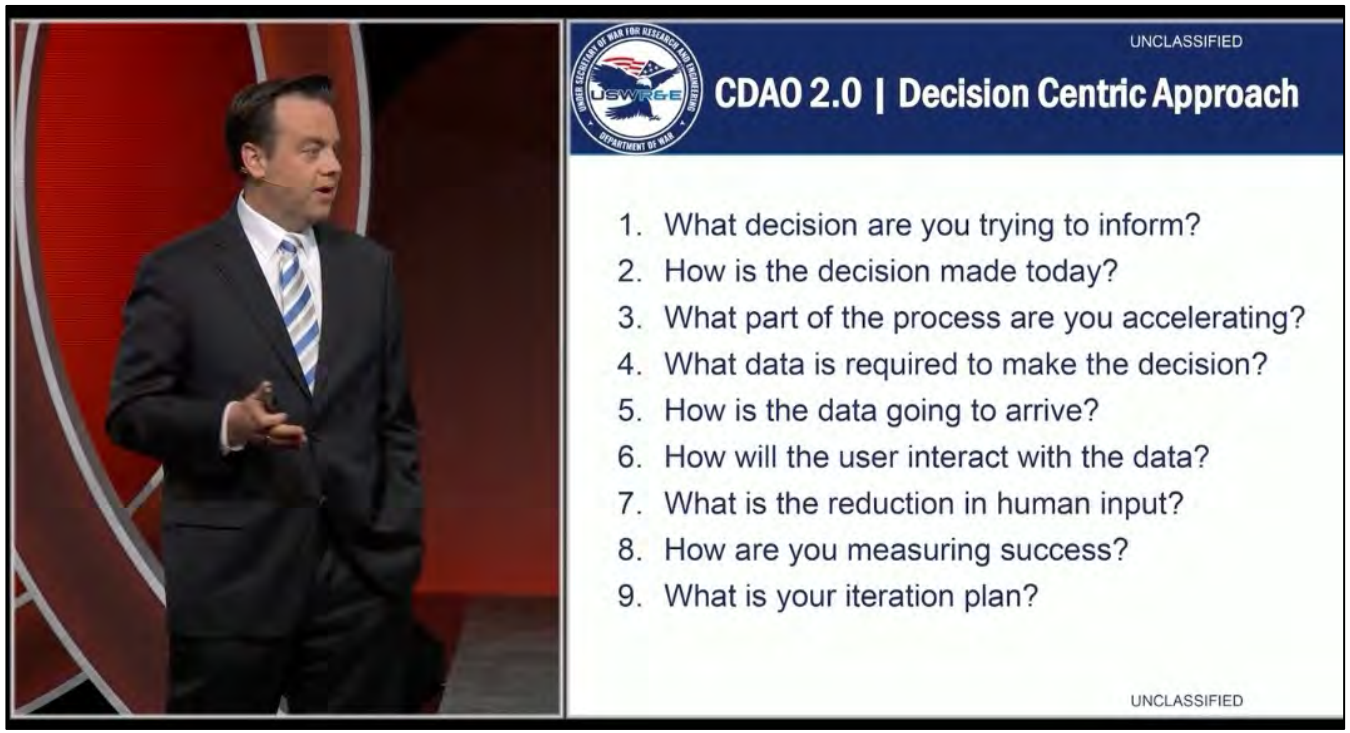
Navy program turns ships into continuous data pipelines for AI development

Applied Intuition has delivered the first Data Edge Collection Kit to the Navy, which will allow the service to constantly collect AI-ready data from the operational environment.

BY MIKAYLA EASLEY • MARCH 19, 2026

Listen to this article 3:54 Learn more.

https://defensescoop.com/2026/03/19/navy-deck-data-pipelines-ai-development-applied-intuition/



https://www.dvidshub.net/video/1000573/cdao-discusses-capabilities-and-cjadc2-aipcon

03.13.2026 Cameron Stanley, Chief Digital and Artificial Intelligence Officer of the Department of War, shares how the DoW is driving enterprise-wide adoption of data, analytics, and AI to generate decision advantage—and what it takes to move cutting-edge technology from the lab to the warfighter at speed.

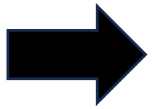
What we want from AI

1. Speed

AI adoption is believed to accelerate decision-making and action.

- Industry: Faster, Better, Cheaper (FBC) Pressures
- Military: Observe – Orient – Decide – Act (OODA) Loop

Too much data (often siloed)
Too fast (timeline compression)



A signal-to-noise crisis !

"How can we quickly learn from automatically collected data?"

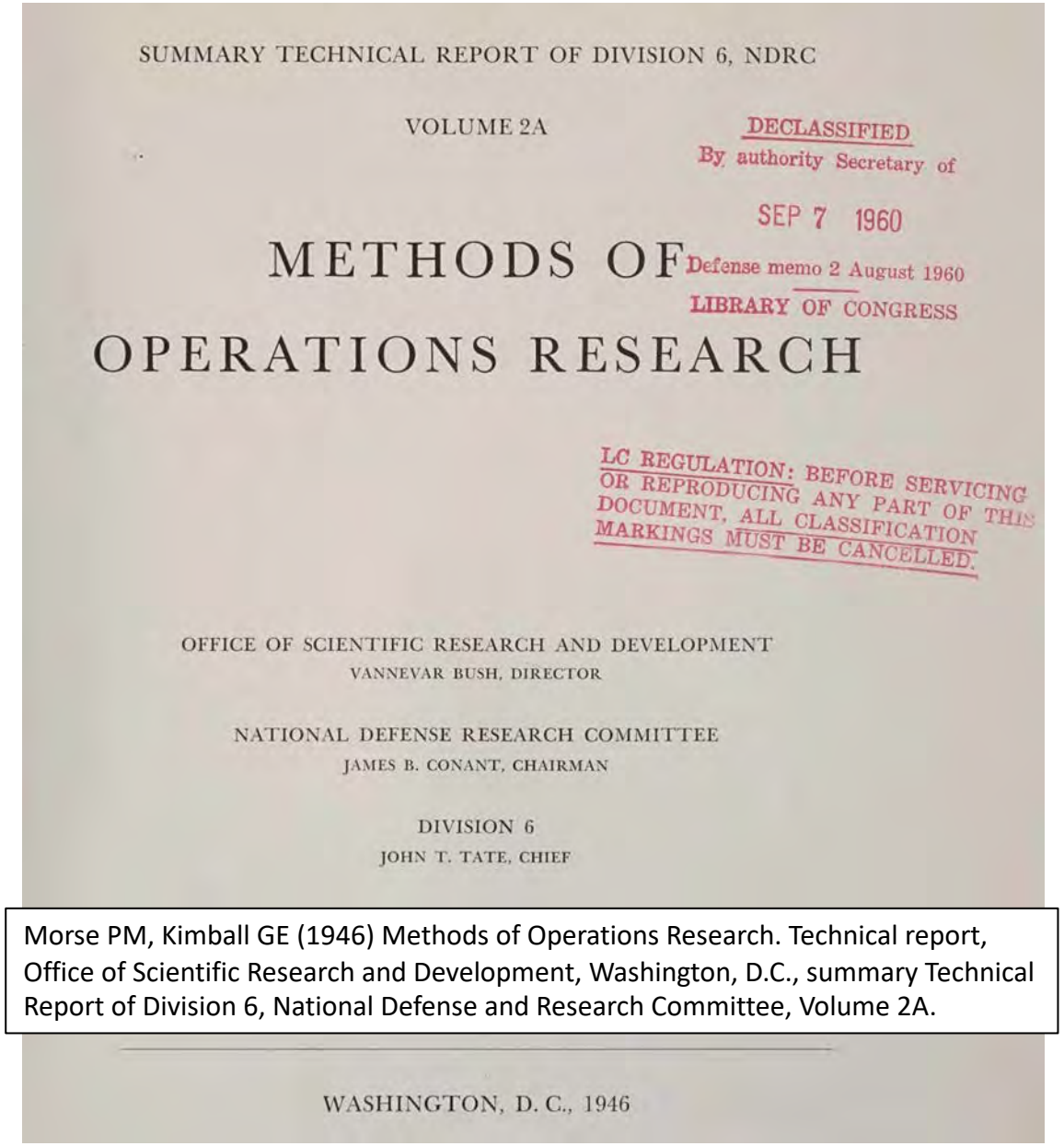
2. Expertise

"Generalists are the best folks to use GenAI so they can do specialist things"

AI is valued as an ***automation technology***.

What is the work of an Operations Analyst?

What is Operations Analysis (OA)?



SUMMARY TECHNICAL REPORT OF DIVISION 6, NDRC

VOLUME 2A

METHODS OF OPERATIONS RESEARCH

OFFICE OF SCIENTIFIC RESEARCH AND DEVELOPMENT
VANNEVAR BUSH, DIRECTOR

NATIONAL DEFENSE RESEARCH COMMITTEE
JAMES B. CONANT, CHAIRMAN

DIVISION 6
JOHN T. TATE, CHIEF

WASHINGTON, D. C., 1946

Morse PM, Kimball GE (1946) Methods of Operations Research. Technical report, Office of Scientific Research and Development, Washington, D.C., summary Technical Report of Division 6, National Defense and Research Committee, Volume 2A.

Operational Research.

P. M. S. BLACKETT (Manchester University).

Advancement of Science, **5.** *April*, 1948.

Two documents written in 1941 were here published generally for the first time. The first, prepared for the Admiralty, states the reasons for setting up Operational Research Sections and defines their functions and organisations.

The need for scientists to study secret operational data in war requires special staffs at Command Headquarters, able to observe actual operational conditions while retaining a scientific approach to the problems. Executive officers in the Services are not trained to analyse complex problems, and can profit by having quantitative results on which to take action. The research team should be directly responsible to the C.-in-C.; their reports should be distributed as widely as possible among the men concerned. The Services' requirements for war equipment need interpretation since the technicians who produce it are often unaware of many of the operational conditions; conversely, the Services need to be kept informed of scientific developments which may help their work. There is a balance to be struck between developing new devices and making the best use of existing ones: the introduction of Operational Research Sections is in part a redeployment of scientific effort to improve this balance.

The second paper, revised in 1943, discusses the methods used to solve many of the wartime problems.

The practical objectives of operational research in war are the appraisement and improvement of weapons and tactics, and the evaluation of strategy in terms of national effort. The methods are broadly classifiable as either "*a priori*" or "variational", mostly the latter. The *a priori* procedure may fail when the problem is complex, and variational analysis is the most useful method of forecasting the yield for a future operation

P. M. S. Blackett (1950) Operational Research, Journal of the Operational Research Society, 1:1, 3-6, DOI: 10.1057/jors.1950.2

Operations analysts use two families of methods

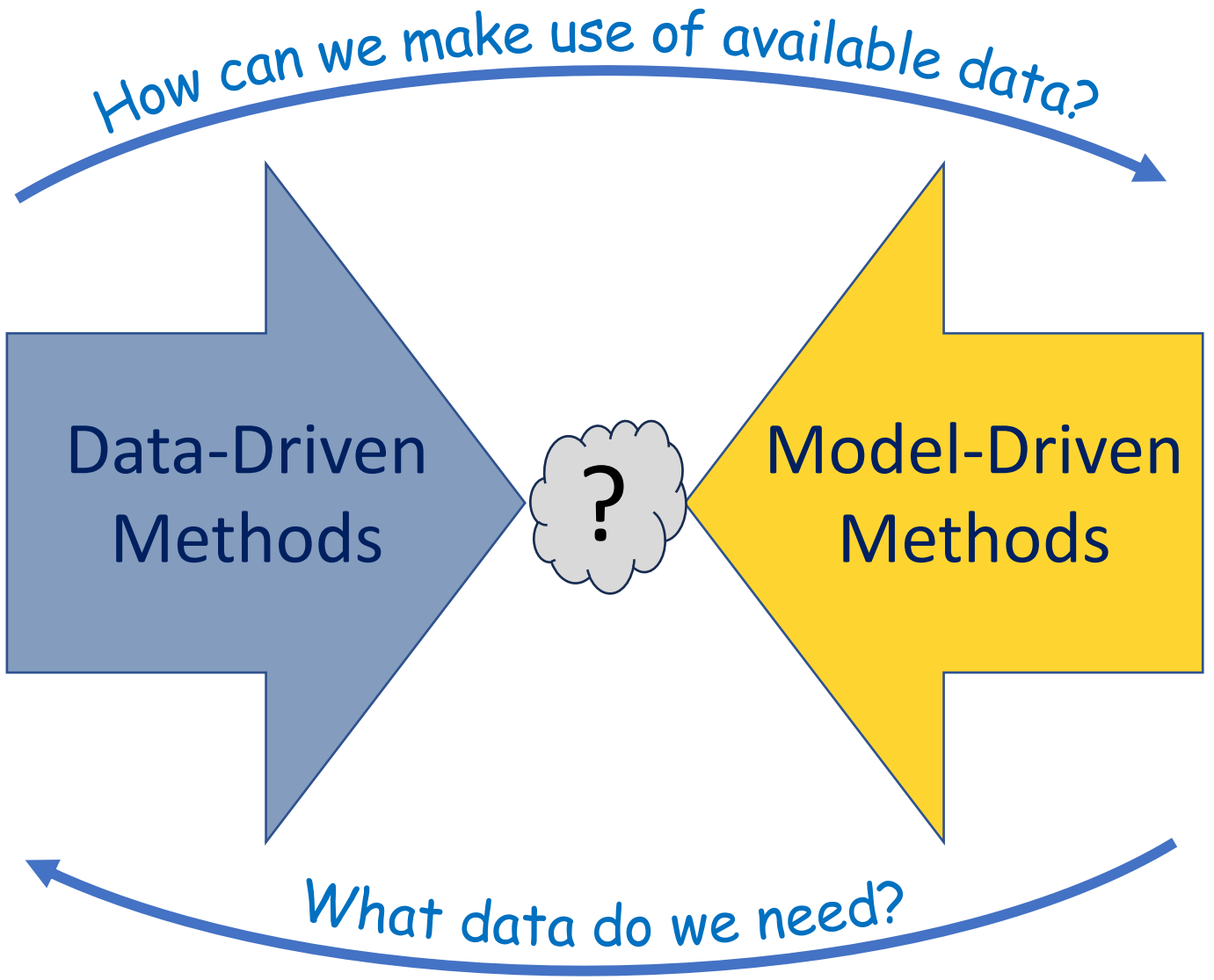
What data do we have?

What does it tell us?

(Why was the data collected?)

(How was the data collected?)

(What can the data actually support?)



What do we need to answer?

What should the answer look like?

What do we already know (e.g., physics)?

What can we assume (e.g., parameters)?


OA
Studies
Decision
Support Tools
Automated
Analytics

	OA Studies 	Decision Support Tools 	Automated Analytics 
Deliverable	Report (and briefing)		
User / Customer	Executive Decision-Maker		
Use Case	One-time, bespoke Blends quantitative + qualitative factors		
Goal	Insight to specific decision(s)		
Example	In a South China Sea conflict, should we devote combatants to escorting merchant resupply vessels?	How best to schedule training sessions at flight school to meet requirements?	How to integrate daily satellite images into readiness assessments?

	OA Studies 	Decision Support Tools 	Automated Analytics 
Deliverable	Report (and briefing)	Standalone Tool (connected to data)	(part of a) Production System
User / Customer	Executive Decision-Maker	Operations Analyst (domain-specific expert)	“Production” Analyst (frontline worker)
Use Case	One-time, bespoke Blends quantitative + qualitative factors	Repeated, custom use Uses data + computation for domain-specific problem	Repeated, mass use Part of a common computational workflow
Goal	Insight to specific decision(s)	Accelerate + transform complicated analysis	Streamlined operations

TECHNIQUES

(Optimization, Stochastic Modeling, Simulation, Statistics, Data Science, AI/ML, HPC)

THEORY

(Mathematics, Algorithms, Computing)

	OA Studies 	Decision Support Tools 	Automated Analytics 
Deliverable	Report (and briefing)	Standalone Tool (connected to data)	(part of a) Production System
User / Customer	Executive Decision-Maker	Operations Analyst (domain-specific expert)	"Production" Analyst (frontline worker)
Use Case	One-time, bespoke Blends quantitative + qualitative factors	Repeated, custom use Uses data + computation for domain-specific problem	Repeated, mass use Part of a common computational workflow
Goal	Insight to specific decision(s)	Accelerate + transform complicated analysis	Streamlined operations

	OA Studies	Decision Support Tools	Automated Analytics
Deliverable	Report (and briefing)		
User / Customer	Executive Decision-Maker		
Use Case	One-time, bespoke Blends quantitative + qualitative factors		
Goal	Insight to specific decision(s)		

Wagner DH, Mylander WC, Sanders TJ (1999) ***Naval Operations Analysis*** (Naval Institute Press), 3rd ed.

104 The OA Method

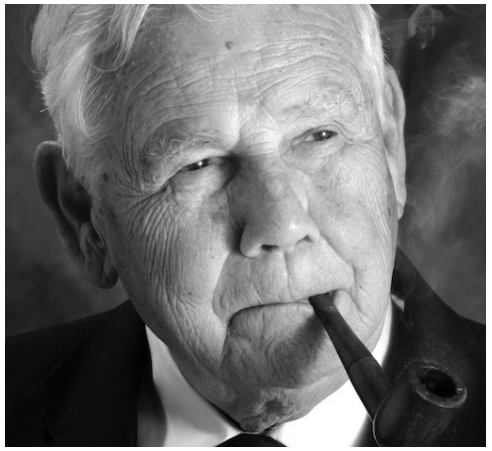
It has been asserted above that OA is essentially the application of scientific methods to the resolution of operational problems. Without digressing for a discussion of what constitutes scientific methods, it is reasonable to state that examination of successful applications of OA reveals a consistent pattern, a general form. That general approach is termed the **OA method**. It is by no means a problem-solving algorithm, but most if not all of its features are to be found in varying degrees of development in all cases of creative problem solving through OA. An outline of the OA method is as follows:

A. Formulation of the problem.
   1. Identification of the objectives of the operation's decision-maker (may or may not be quantitative).
   2. Identification of the reasonable alternative courses of action.
   3. Identification of the variables that impact the courses of action.
   4. Definition of a measure of effectiveness (MOE), i.e., a quantitative yardstick providing an ordering of the alternative courses of action that is consistent within the objective.

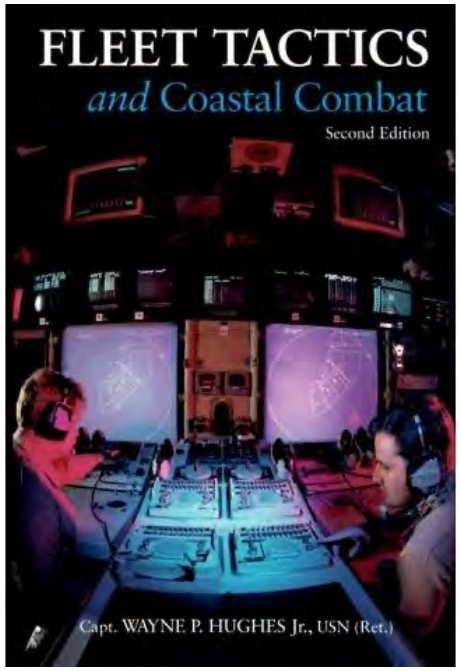
B. Analysis of the problem.
   1. Construct a model of the operation by analytic formulas and/or Monte Carlo simulation (see Chapter 3) that is faithful to reality and amenable to analysis.
   2. Evaluate, in terms of the MOE, outcomes of the alternative courses of action by exercising the model and by theoretical analysis.
   3. Conduct operational trials or observation of "real world" operations to obtain data needed in (1) and (2).

C. Communication of the results, orally and in writing.

D. Analyst assistance to implementation of the decision.



Wayne P. Hughes
Navy CAPT (Ret.)
MORS Fellow
1930-2019



Given a limited amount of time to address a problem (always the case), allocate time as follows:

1/3	*defining the problem*
1/3	*performing the analysis*
1/3	*preparing the report / briefing*

Wagner DH, Mylander WC, Sanders TJ (1999) ***Naval Operations Analysis*** (Naval Institute Press), 3rd ed.

8 Naval Operations Analysis

104 The OA Method

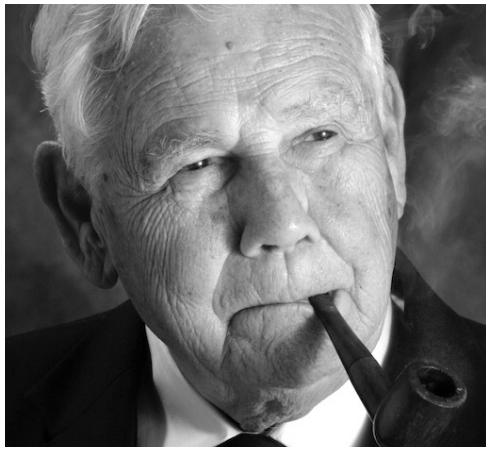
It has been asserted above that OA is essentially the application of scientific methods to the resolution of operational problems. Without digressing for a discussion of what constitutes scientific methods, it is reasonable to state that examination of successful applications of OA reveals a consistent pattern, a general form. That general approach is termed the **OA method**. It is by no means a problem-solving algorithm, but most if not all of its features are to be found in varying degrees of development in all cases of creative problem solving through OA. An outline of the OA method is as follows:

A. Formulation of the problem.
   1. Identification of the objectives of the operation's decision-maker (may or may not be quantitative).
   2. Identification of the reasonable alternative courses of action.
   3. Identification of the variables that impact the courses of action.
   4. Definition of a measure of effectiveness (MOE), i.e., a quantitative yardstick providing an ordering of the alternative courses of action that is consistent within the objective.

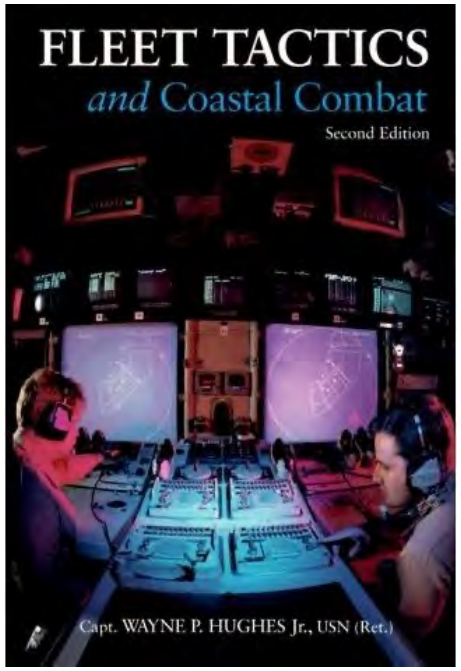
B. Analysis of the problem.
   1. Construct a model of the operation by analytic formulas and/or Monte Carlo simulation (see Chapter 3) that is faithful to reality and amenable to analysis.
   2. Evaluate, in terms of the MOE, outcomes of the alternative courses of action by exercising the model and by theoretical analysis.
   3. Conduct operational trials or observation of "real world" operations to obtain data needed in (1) and (2).

C. Communication of the results, orally and in writing.

D. Analyst assistance to implementation of the decision.



Wayne P. Hughes
Navy CAPT (Ret.)
MORS Fellow
1930-2019



Given a limited amount of time to address a problem (always the case), allocate time as follows:

1/3	*defining the problem*
1/3	*performing the analysis*
1/3	*preparing the report / briefing*

Gerald G. Brown, Ph.D.
Distinguished Professor Emeritus, NPS
National Academy of Engineering

Brown, Gerald G. 2004 "How To Write About Operations Research," PHALANX, Vol. 37, No. 3, p. 7.

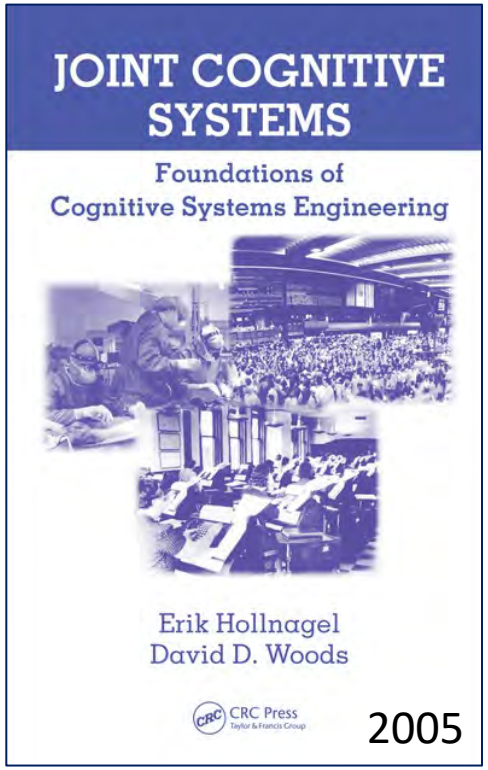
"There are five simple, essential questions you must answer… preferably in this order:

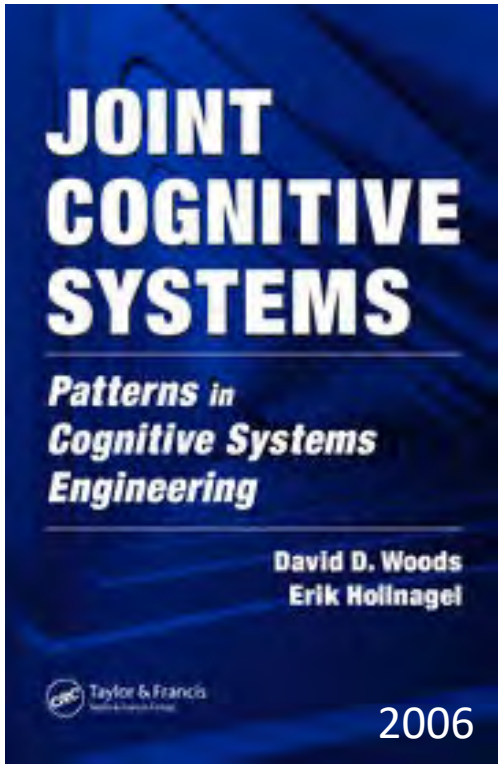
1. *What is the problem?*
2. *Why is this problem important?*
3. *How will this problem be solved without your help?*
4. *What are you doing to solve this problem?*
5. *How will we know when you have succeeded?"*

Both emphasize the importance of framing!

Key Element: Understanding *Cognitive Systems*

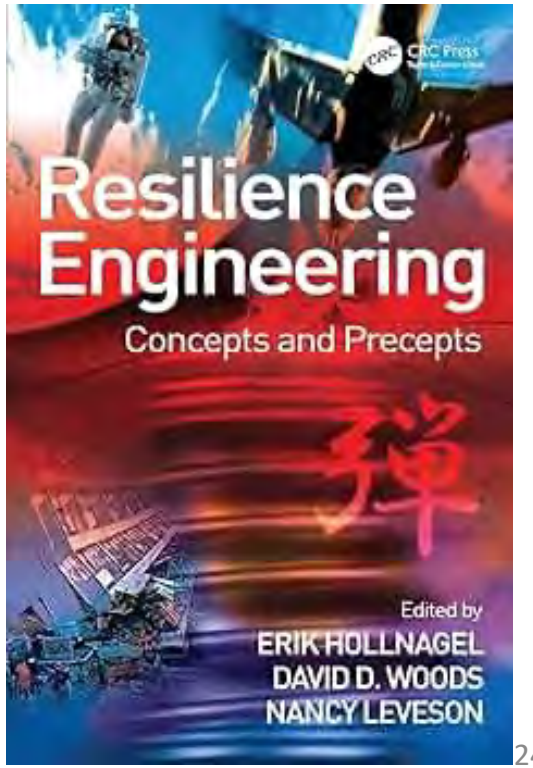
Cognitive Work is about **Building, Revising,** and **Reframing** ***models*** of **How the world works,** and **Our place in it**





When challenged by extreme events, how do systems handle situations that fall outside their design envelope?

- **Brittle?**
  Do they saturate and fail?
- **Resilient?**
  Do different elements to come together to extend capability in ways that are novel?



	OA Studies 	Decision Support Tools 	Automated Analytics 
Deliverable	Report (and briefing)	Standalone Tool (connected to data)	(part of a) Production System
User / Customer	Executive Decision-Maker	Operations Analyst (domain-specific expert)	"Production" Analyst (frontline worker)
Use Case	One-time, bespoke Blends quantitative + qualitative factors	Repeated, custom use Uses data + computation for domain-specific problem	Repeated, mass use Part of a common computational workflow
Goal	Insight to specific decision(s)	Accelerate + transform complicated analysis	Streamlined operations

Understanding automation technology

(a cognitive systems engineering perspective)

Work as imagined (WAI)

- System is built and operated as designed
- Components of the system (humans, algorithms, devices) behave as specified
- Exceptions/Anomalies are relatively few & usually well anticipated.

Work as done (WAD)

- Things are messy: “adaptations tailored to contingencies and context are always going on”
- “The adaptations that make the system function also hide the systems weaknesses.”
- “Management often can’t see the gaps so it seems that the system is functioning as designed.”
- Anomalies and surprises are continuous.

How we imagine automation

- Humans are the sources of inefficiency
- New technology can be introduced as a simple substitution of machines for people
- The system will be preserved and improved

How automation actually happens

- Adding or expanding the machine’s role changes the cooperative architecture and changes the role of the human in the system
- This is a ***joint system***—and needs to be designed and operated as such

References:
Hollnagel, Woods & Leveson (2006). Resilience Engineering. Woods et al., Behind Human Error (1994/ 2nd edition 2010). Woods and Decker (2000)
Quotes from Woods et al. (2021). Patterns in How People Think and Work: Importance of Patterns Discovery for Understanding Complex Adaptive Systems..

Understanding automation technology

(a cognitive systems engineering perspective)

Work as imagined (WAI)

- System is built and operated as designed
- Components of the system (humans, algorithms, devices) behave as specified
- Exceptions/Anomalies are relatively few & usually well anticipated.

Work as done (WAD)

- Things are messy: "adaptations tailored to contingencies and context are always going on"
- "The adaptations that make the system function also hide the systems weaknesses."
- "Management often can't see the gaps so it seems that the system is functioning as designed."
- Anomalies and surprises are continuous.

How we imagine automation

- Humans are the sources of inefficiency
- New technology can be introduced as a simple substitution of machines for people
- The system will be preserved and improved

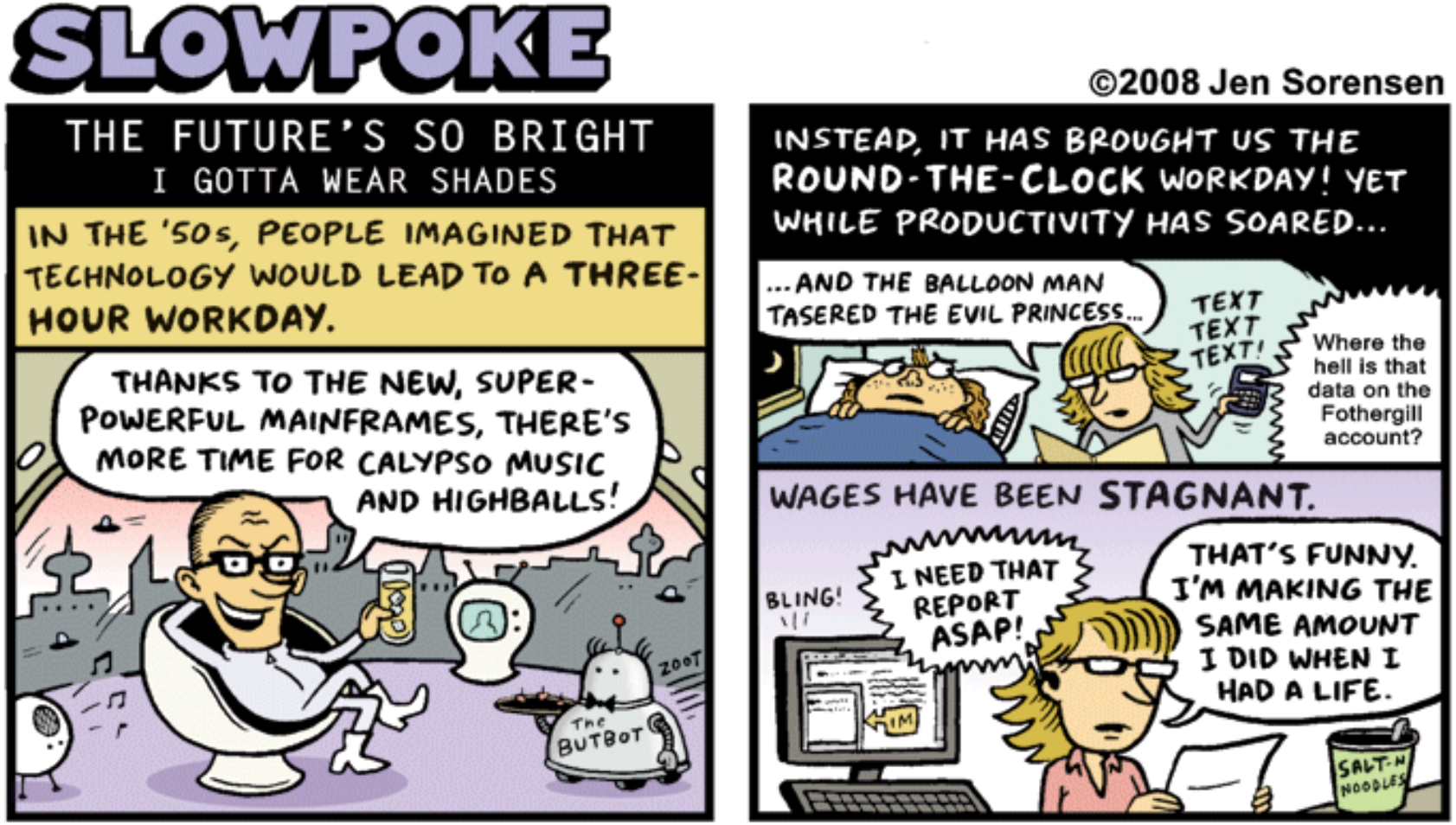
Substitution Myth

How automation actually happens

- Adding or expanding the machine's role changes the cooperative architecture and changes the role of the human in the system
- This is a ***joint system***—and needs to be designed and operated as such

References:
Hollnagel, Woods & Leveson (2006). Resilience Engineering. Woods et al., Behind Human Error (1994/ 2nd edition 2010). Woods and Decker (2000)
Quotes from Woods et al. (2021). Patterns in How People Think and Work: Importance of Patterns Discovery for Understanding Complex Adaptive Systems..

Law of Stretched Systems



Substitution Myth

Technology believed to be a simple replacement for humans

Think AI is going to save time, but actually leads to intensifying work

AI tools might degrade performance in software coding activities

Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity

Joel Becker*, **Nate Rush***, **Beth Barnes, David Rein**

Model Evaluation & Threat Research (METR)

Abstract

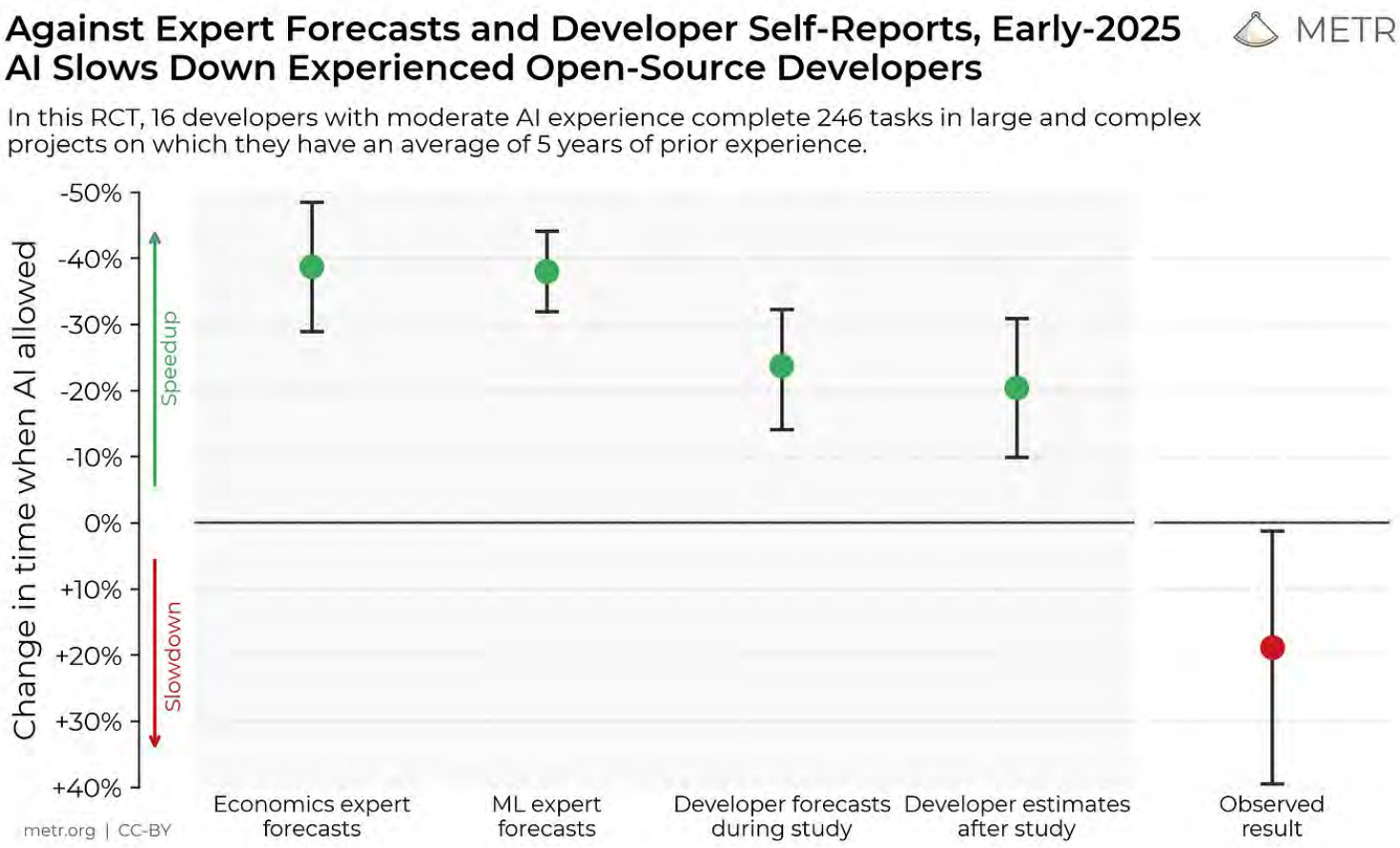
Despite widespread adoption, the impact of AI tools on software development in the wild remains understudied. We conduct a randomized controlled trial (RCT) to understand how AI tools at the February–June 2025 frontier affect the productivity of experienced open-source developers. 16 developers with moderate AI experience complete 246 tasks in mature projects on which they have an average of 5 years of prior experience. Each task is randomly assigned to allow or disallow usage of early-2025 AI tools. When AI tools are allowed, developers primarily use Cursor Pro, a popular code editor, and Claude 3.5/3.7 Sonnet. Before starting tasks, developers forecast that allowing AI will reduce completion time by 24%. After completing the study, developers estimate that allowing AI reduced completion time by 20%. Surprisingly, we find that allowing AI actually *increases* completion time by 19%—AI tooling slowed developers down. This slowdown also contradicts predictions from experts in economics (39% shorter) and ML (38% shorter). To understand this result, we collect and evaluate evidence for 21 properties of our setting that *a priori* could contribute to the observed slowdown effect—for example, the size and quality standards of projects, or prior developer experience with AI tooling. Although the influence of experimental artifacts cannot be entirely ruled out, the robustness of the slowdown effect across our analyses suggests it is unlikely to primarily be a function of our experimental design.

1 Introduction

Software development is an important part of the modern economy, and a key domain for understanding and forecasting AI capabilities [1; 2]. Frontier AI systems demonstrate impressive capabilities on a wide range of software benchmarks [3; 4; 5; 6; 7; 8; 9] and in experiments measuring AI's impact on developer productivity when completing synthetic tasks [10; 11]. However, tasks used in these *lab experiments* sacrifice realism for scale and efficiency: the tasks are typically self-contained, do not require much prior context/familiarity to understand and complete, and use algorithmic evaluation metrics which do not capture many important capabilities [12; 13; 14]. As a result, it can be difficult to draw inferences from results on these evaluations about AI's impact in practice.

To reduce the inferential gap between measurements of AI capabilities and real-world impact, one can measure the impact of AI systems in real-world settings (i.e. *field experiments*). Existing field experiments aimed at measuring AI's impact on software development measure outcomes like number of added lines of code or number of tasks completed [15; 16; 17]. However, AI systems can affect these outcomes without productivity actually increasing—for example, code can be more verbose but functionally equivalent, and tasks can be broken up into multiple smaller tasks without the total amount of work changing—making it challenging to interpret these results.

*Equal contribution. Correspondence to {nate, joel}@metr.org



Experienced software developers with access to AI tools took 19% *longer* to complete their tasks, despite believing they had finished 20% faster.

AI tools might degrade performance in software coding activities

Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity

Joel Becker*, **Nate Rush***, **Beth Barnes**, **David Rein**

Model Evaluation & Threat Research (METR)

Abstract

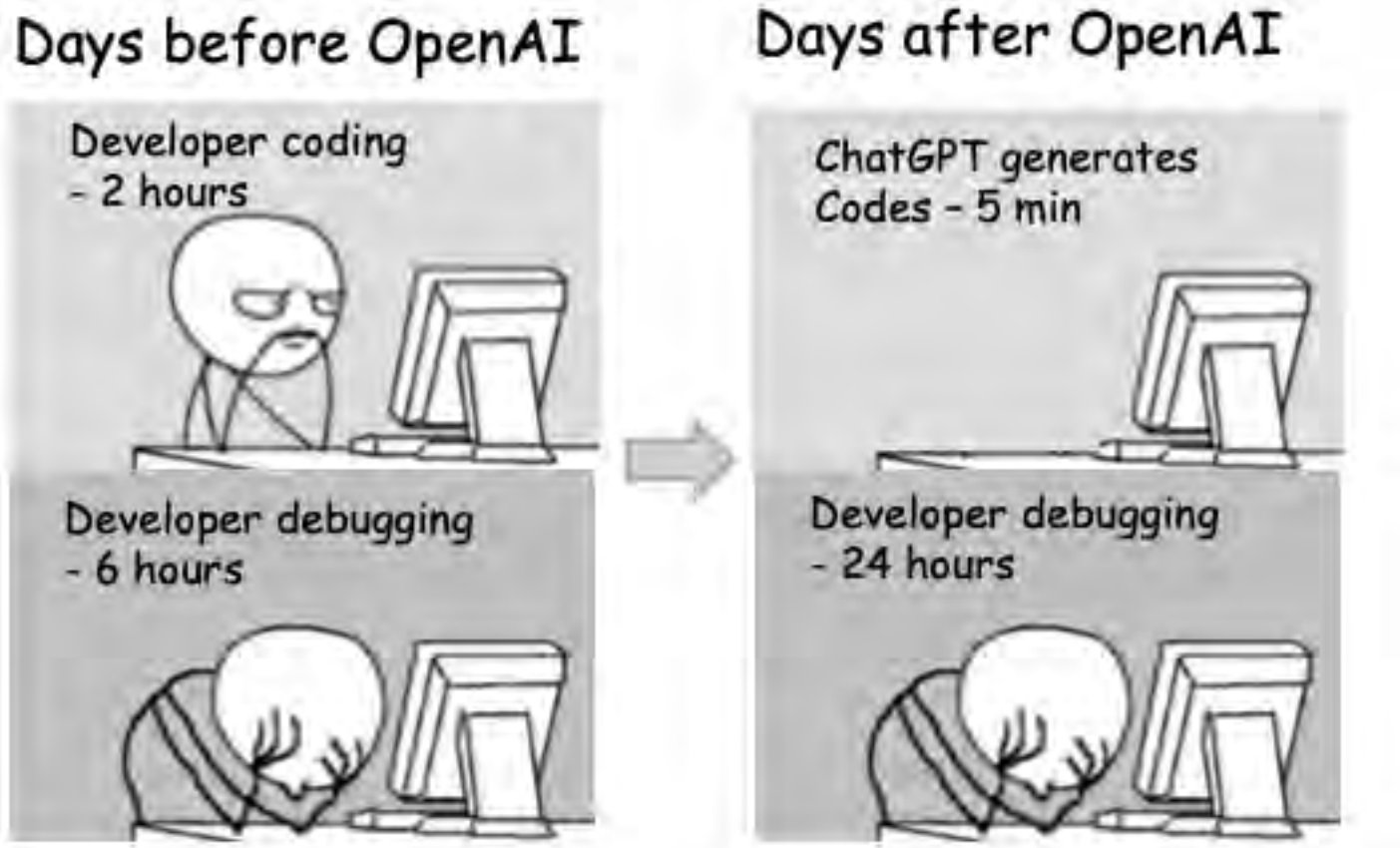
Despite widespread adoption, the impact of AI tools on software development in the wild remains understudied. We conduct a randomized controlled trial (RCT) to understand how AI tools at the February–June 2025 frontier affect the productivity of experienced open-source developers. 16 developers with moderate AI experience complete 246 tasks in mature projects on which they have an average of 5 years of prior experience. Each task is randomly assigned to allow or disallow usage of early-2025 AI tools. When AI tools are allowed, developers primarily use Cursor Pro, a popular code editor, and Claude 3.5/3.7 Sonnet. Before starting tasks, developers forecast that allowing AI will reduce completion time by 24%. After completing the study, developers estimate that allowing AI reduced completion time by 20%. Surprisingly, we find that allowing AI actually *increases* completion time by 19%—AI tooling slowed developers down. This slowdown also contradicts predictions from experts in economics (39% shorter) and ML (38% shorter). To understand this result, we collect and evaluate evidence for 21 properties of our setting that *a priori* could contribute to the observed slowdown effect—for example, the size and quality standards of projects, or prior developer experience with AI tooling. Although the influence of experimental artifacts cannot be entirely ruled out, the robustness of the slowdown effect across our analyses suggests it is unlikely to primarily be a function of our experimental design.

1 Introduction

Software development is an important part of the modern economy, and a key domain for understanding and forecasting AI capabilities [1; 2]. Frontier AI systems demonstrate impressive capabilities on a wide range of software benchmarks [3; 4; 5; 6; 7; 8; 9] and in experiments measuring AI's impact on developer productivity when completing synthetic tasks [10; 11]. However, tasks used in these *lab experiments* sacrifice realism for scale and efficiency: the tasks are typically self-contained, do not require much prior context/familiarity to understand and complete, and use algorithmic evaluation metrics which do not capture many important capabilities [12; 13; 14]. As a result, it can be difficult to draw inferences from results on these evaluations about AI's impact in practice.

To reduce the inferential gap between measurements of AI capabilities and real-world impact, one can measure the impact of AI systems in real-world settings (i.e. *field experiments*). Existing field experiments aimed at measuring AI's impact on software development measure outcomes like number of added lines of code or number of tasks completed [15; 16; 17]. However, AI systems can affect these outcomes without productivity actually increasing—for example, code can be more verbose but functionally equivalent, and tasks can be broken up into multiple smaller tasks without the total amount of work changing—making it challenging to interpret these results.

*Equal contribution. Correspondence to {nate, joel}@metr.org



The rise of AI “workslop”

Generative AI

AI-Generated “Workslop” Is Destroying Productivity

HARVARD BUSINESS REVIEW

by Kate Niederhoffer, Gabriella Rosen Kellerman, Angela Lee, Alex Liebscher, Kristina Rapuano and Jeffrey T. Hancock

September 22, 2025, Updated September 25, 2025

https://hbr.org/2025/09/ai-generated-workslop-is-destroying-productivity

- “But while some employees are using [AI-enabled tools] to polish good work, others use it to create content that is actually unhelpful, incomplete, or missing crucial context about the project at hand. “
- **“Employees are using AI tools to create low-effort, passable looking work that ends up creating more work for their coworkers.”**
- “We define workslop as *AI generated work content that masquerades as good work, but lacks the substance to meaningfully advance a given task*.”
- “The insidious effect of workslop is that it shifts the burden of the work downstream, requiring the receiver to interpret, correct, or redo the work. In other words, it transfers the effort from creator to receiver.”

AI tools are shifting the work in “coding” activities

2023-2024

Prompt Engineering

Prompting the model to get the results you want

- Adopt a persona
- Give it constraints

2025

Context Engineering

It’s not simply what you tell the model, it’s what set of information (context) the model can access

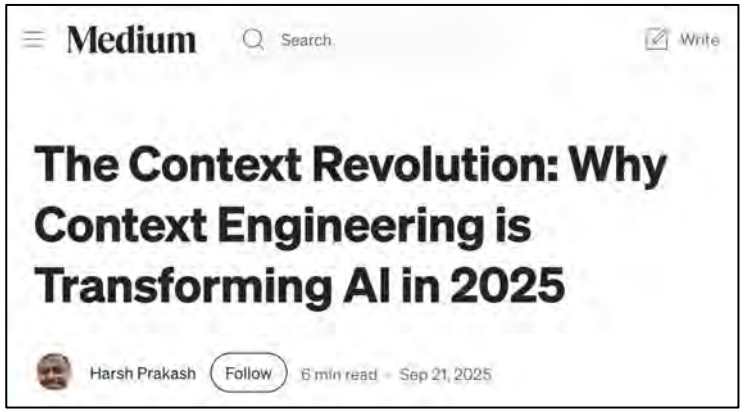
- Files (and folders)

2026

Harness Engineering

It’s everything you put around the model (e.g., tools, access, broader system) that helps it do what you intend it to do

https://medium.com/@toptalenticalcio/the-art-and-science-of-prompt-engineering-2024-2edfcbd81204



https://medium.com/@hs5492349/the-context-revolution-why-context-engineering-is-transforming-ai-in-2025-cbf68aa388ea

https://medium.com/@jerry.shao/harness-engineering-building-production-grade-ai-systems-beyond-prompts-and-context-5fcdffdd6b4c

Understanding automation technology

(a cognitive systems engineering perspective)

Work as imagined (WAI)

- System is built and operated as designed
- Components of the system (humans, algorithms, devices) behave as specified
- Exceptions/Anomalies are relatively few & usually well anticipated.

Work as done (WAD)

- Things are messy: “adaptations tailored to contingencies and context are always going on”
- “The adaptations that make the system function also hide the systems weaknesses.”
- “Management often can’t see the gaps so it seems that the system is functioning as designed.”
- Anomalies and surprises are continuous.

How we imagine automation

- Humans are the sources of inefficiency
- New technology can be introduced as a simple substitution of machines for people
- The system will be preserved and improved

How automation actually happens

- Adding or expanding the machine’s role changes the cooperative architecture and changes the role of the human in the system
- This is a ***joint system***—and needs to be designed and operated as such

References:
Hollnagel, Woods & Leveson (2006). Resilience Engineering. Woods et al., Behind Human Error (1994/ 2nd edition 2010). Woods and Decker (2000)
Quotes from Woods et al. (2021). Patterns in How People Think and Work: Importance of Patterns Discovery for Understanding Complex Adaptive Systems..

Decision Support Tools

Example

How best to schedule training sessions at flight school to meet requirements?

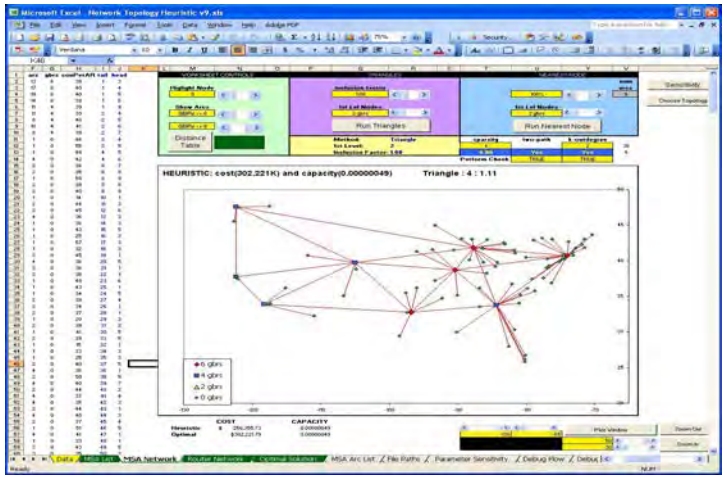
How automation actually happens

- Adding or expanding the machine's role changes the cooperative architecture and changes the role of the human in the system
- This is a ***joint system***—and needs to be designed and operated as such

Before

Manual process is slow and error-prone. Goal is simply to find a feasible solution.

Then



Decision support tool provides rapid, repeatable, process for optimal solution.

After

Analysts task changes to consider "what-if" scenarios, explore solution tradespace.

We should not expect that the mere deployment of decision support tools necessarily improves performance.

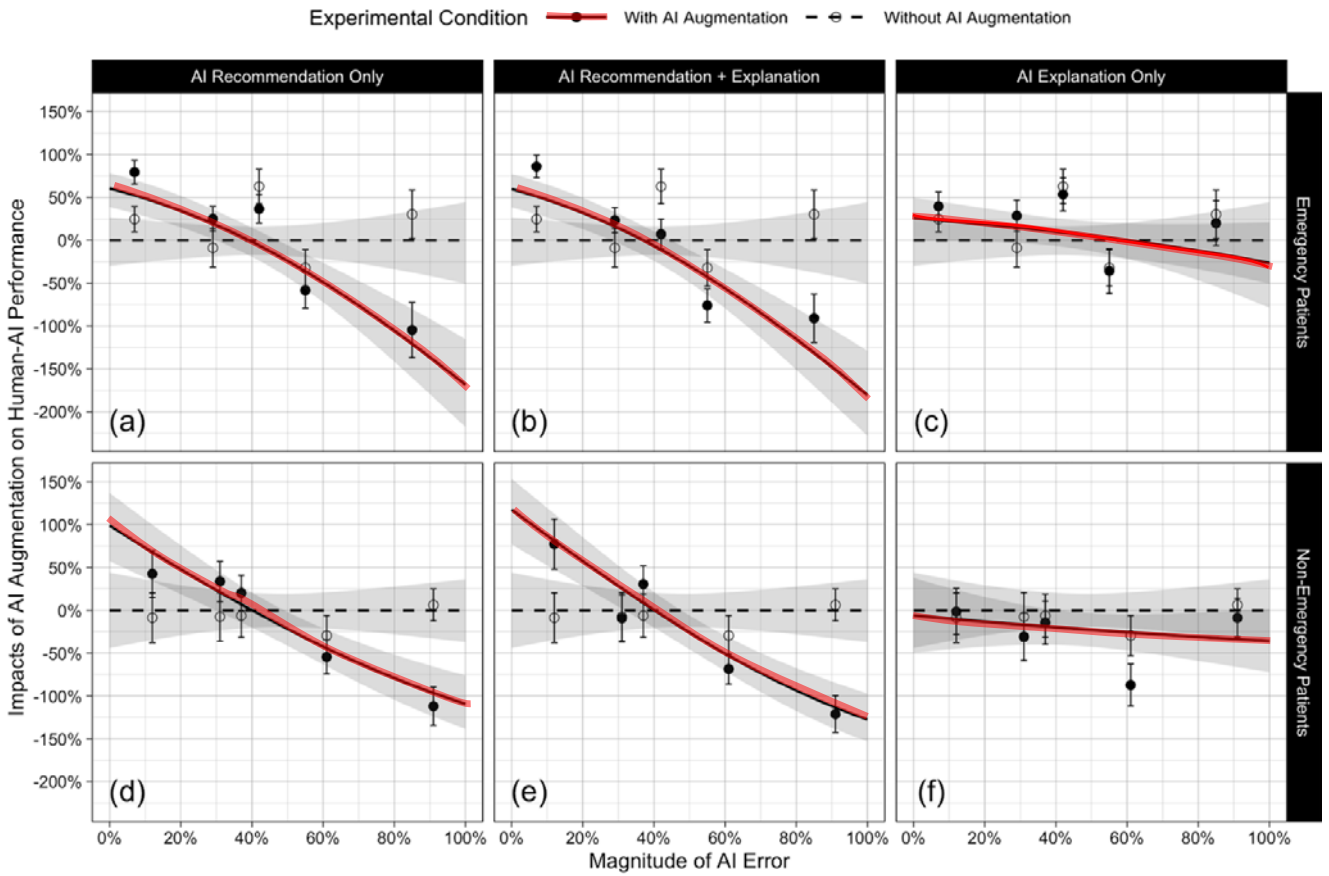
- It is insufficient (and misguided) to evaluate the performance of the tool in a standalone way.

  Instead, we should be evaluating the joint performance of user + tool.

- In cases where a decision support tool makes incorrect (or poor) recommendations, we should NOT expect that users (even those with high expertise) will be able to say "that's wrong" and disregard the guidance of the tool.

  In high-stakes situations involving life and death, we should insist on rigor in evaluating the correctness of model/tool output.

Performance of 450 nursing students and a dozen licensed nurses when reviewing 10 historical ICU cases



- "When AI predictions were most correct, nurses performed 53% to 67% better than when they worked without AI assistance."
- "However, when AI predictions were most misleading, nurses performed 96% to 120% worse than when they worked without AI assistance."

Good AI Does Not Guarantee a Good Human-AI System!

Towards Joint Activity Design Heuristics: Essentials for Human-Machine Teaming

Proceedings of the Human Factors and Ergonomics Society Annual Meeting
1–6

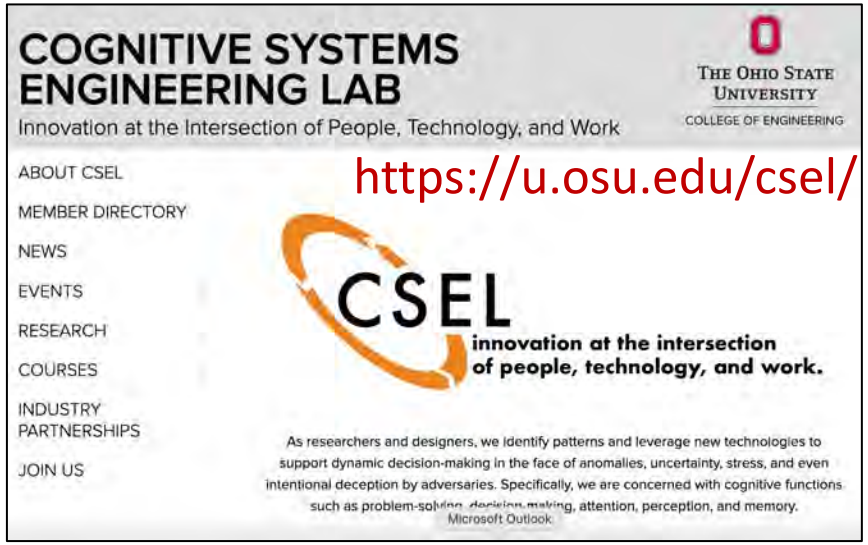
DOI: 10.1177/21695067231193646
journals.sagepub.com/home/pro
Sage

Dane A. Morey[1], Prerana Walli[1], Kenneth S. Cassidy[1], Priyanka K. Tewani[1], Morgan E. Reynolds[1], Samantha Malone[1], Mohammadreza Jalaeian[1], Michael F. Rayo[1], and Nicolette M. McGeorge[2]

Johnson, M., & Vera, A. (2019). **No AI is an island: the case for teaming intelligence.** *AI magazine*, *40*(1), 16-28.

Key Ideas for Joint Systems:

- "The machine changes and augments what people can do, rather than replaces them."
- "The machine interacts with people: they cannot remain separate or invisible"
- We need to design human-AI architectures



	OA Studies 	Decision Support Tools 	Automated Analytics 
Deliverable	Report (and briefing)	Standalone Tool (connected to data)	(part of a) Production System
User / Customer	Executive Decision-Maker	Operations Analyst (domain-specific expert)	"Production" Analyst (frontline worker)
Use Case	One-time, bespoke Blends quantitative + qualitative factors	Repeated, custom use Uses data + computation for domain-specific problem	Repeated, mass use Part of a common computational workflow
Goal	Insight to specific decision(s)	Accelerate + transform complicated analysis	Streamlined operations
Example	In a South China Sea conflict, should we devote combatants to escorting merchant resupply vessels?	How best to schedule training sessions at flight school to meet requirements?	How to integrate daily satellite images into readiness assessments?

The field of cognitive systems engineering is full of examples of "automation gone wrong"

*Automatica, Vol. 19, No. 6. pp. 775 779, **1983***

Brief Paper

Ironies of Automation*

LISANNE BAINBRIDGE†

Key Words—Control engineering computer applications; man–machine systems; on-line operation; process control; system failure and recovery.

"The designer's view of the human operator may be that the operator is unreliable and inefficient, so should be eliminated from the system."

The "irony is that the designer who tries to eliminate the operator still leaves the operator to do the tasks which the designer cannot think how to automate."

"By taking away the easy parts of his task, automation can make the difficult parts of the human operator's task more difficult."

Monitoring

"A more serious irony is that the automatic control system has been put in because it can do the job better than the operator, but yet the operator is being asked to monitor that it is working effectively."

"'Catastrophic' breaks to failure are relatively easy to identify. Unfortunately automatic control can 'camouflage' system failure by controlling against the variable changes, so that trends do not become apparent until they are beyond control."

Manual Take-Over

"When manual take-over is needed there is likely to be something wrong with the process, so that unusual actions will be needed to control it, and one can argue that the operator needs to be more rather than less skilled, and less rather than more loaded, than average."

"Perhaps the final irony is that it is the most successful automated systems, with rare need for manual intervention, which may need the greatest investment in human operator training."

The field of cognitive systems engineering is full of examples of "automation gone wrong"

Automatica, Vol. 19, No. 6. pp. 775 779, 1983

Brief Paper

Ironies of Automation*

LISANNE BAINBRIDGE†

Key Words—Control engineering computer applications; man–machine systems; on-line operation; process control; system failure and recovery.

THEOR. ISSUES IN ERGON. SCI., 2000, VOL. 1, NO. 3, 272–282

Anticipating the effects of technological change: a new era of dynamics for human factors

DAVID WOODS†* and SIDNEY DEKKER‡

† Institute for Ergonomics, 210 Baker Systems, The Ohio State University, Columbus, OH 43210, USA

‡ Linköping Institute of Technology, S-58183 Linköping, Sweden

Keywords: Technology change; Cognitive task analysis; Future of Human Factors.

June 2016 - *Journal of Cognitive Engineering and Decision Making*

The Risks of Autonomy: Doyle's Catch

David D. Woods, The Ohio State University

Special Issue Paper

Limits of Automata—Then and Now: Challenges of Architecture, Brittleness, and Scale

Journal of Cognitive Engineering and Decision Making
2024, Vol. 0(0) 1–8

DOI: 10.1177/15553434241240203
journals.sagepub.com/home/edm
Sage

David D. Woods[1]

Special Issue Paper

Wrong, Strong, and Silent: What Happens when Automated Systems With High Autonomy and High Authority Misbehave?

Journal of Cognitive Engineering and Decision Making
2024, Vol. 0(0) 1–7

DOI: 10.1177/15553434241240849
journals.sagepub.com/home/edm
Sage

Sidney W. A. Dekker[1] **and David D. Woods**[2]

Agentic AI

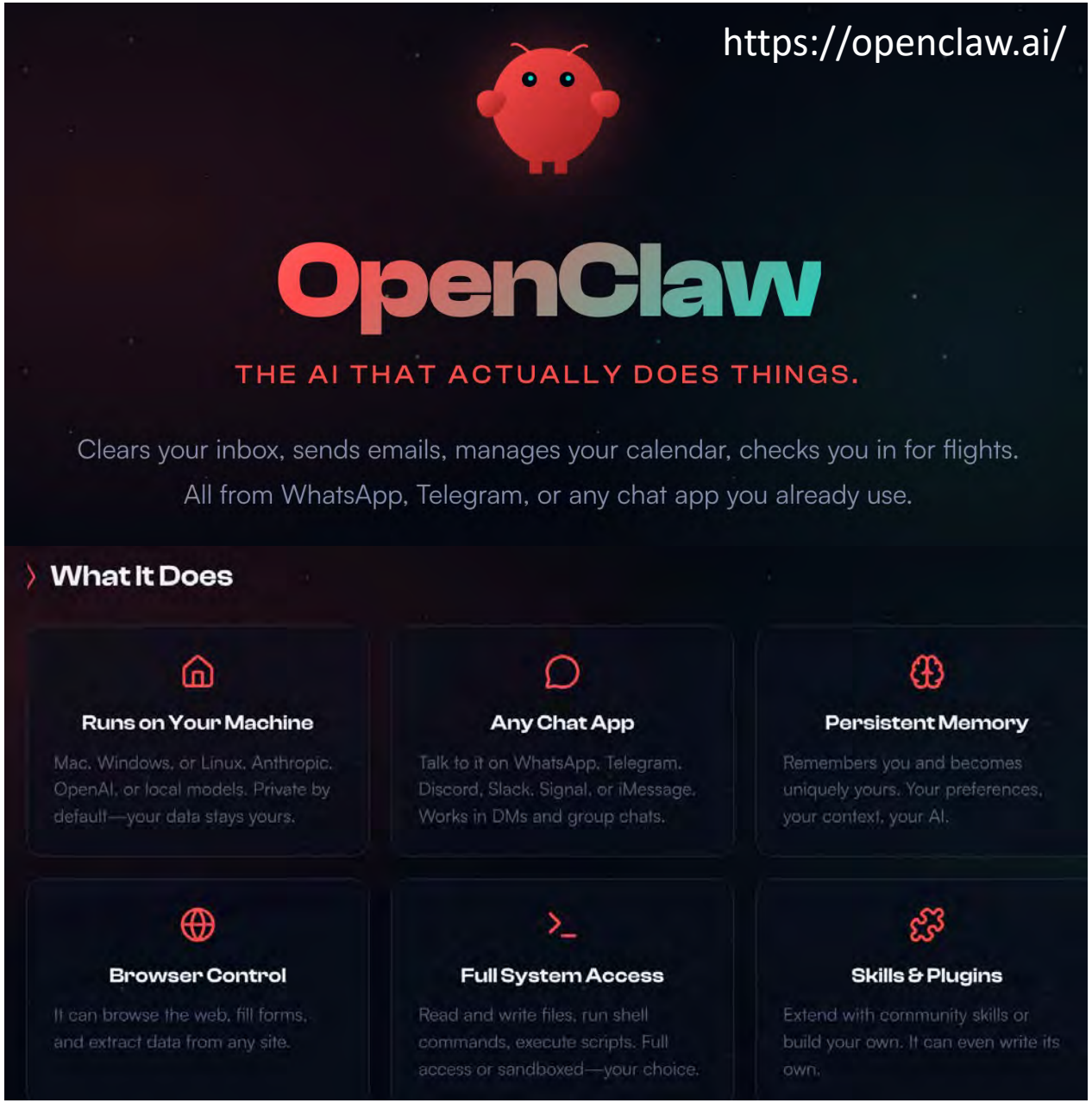
- Goal-directed behavior
- Can independently plan and take actions
- Adapts based on feedback

Skills

- Modular, reusable capabilities that an agent can call to accomplish specific tasks
- Given a goal, an agent can search, discover, select and execute skills iteratively to perform its work

Open Claw

- An agent runtime/infrastructure layer
- Runs on a local machine
- Connects to:
  - WhatsApp, Telegram, Slack, Signal
  - local files, email, calendar
  - underlying models like Claude or GPT



Today's Agenda

- What do [we] want from AI?
- What is the *work* of Military Operations Analysis?
- Automation
- Expertise
- What about LLMs?
- Challenges and Opportunities

Magazine @1962
Warhol time capsule
Warhol museum 10/25

Courtesy
David
Woods

Assessing and Understanding Expertise

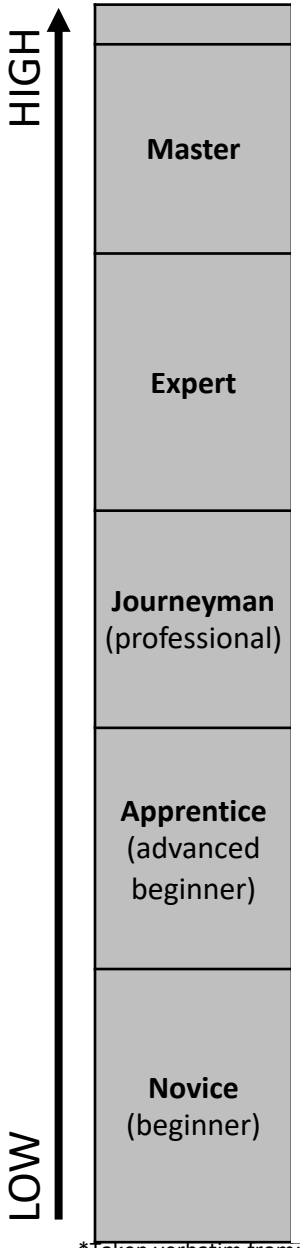
- The notion of ”expertise” is often defined loosely.

- However, cognitive scientists** have formal definitions of expertise in terms of:
  1. Developmental progression (*the experience needed* to acquire it)
  2. Knowledge organization (what is known and *how it is organized, e.g., models, abstractions*)
  3. Reasoning processes (*how experts think* and reason)

- Traditional levels of proficiency come from the “craft guilds” of the Middle Ages…

**Leading expert: Robert R. Hoffman, PhD., Institute for Human and Machine Cognition, Florida University Systems, http://www.ihmc.us/groups/rhoffman/

Basic Levels of Proficiency



*Taken verbatim from: Hoffman, R. R. (1998). How can expertise be defined?: Implications of research from cognitive psychology. In R. Williams, W. Faulkner, & J. Fleck (Eds.), Exploring expertise (pp. 81-100). New York: Macmillan.

**Adapted from: Denning PJ (2002) Career redux. Communications of the ACM 45(9):21-26.

Basic Levels of Proficiency

HIGH

LOW

	Traditional Description*	How they operate**	How they contribute**	How they learn and interact**
Master	Traditionally, a master is **one of an elite group of experts whose judgments set the regulations, standards, or ideals.** Also, a master can be that expert who is regarded by the other experts as being "the" expert, or the "real" expert, especially with regard to sub-domain knowledge.	Has studied with many different teachers and has developed own distinctive style. Has produced innovations in the standard practices of others, altered the course of history in the field, and knows how to do this again.	Develops new methods and practices for the field. Capacity for long-range strategic thinking and action. Sees historical drifts and shifting clearings.	Learning continues by working with other masters as teachers. **Creates and leads professional networks.** **Teaches others to be experts and masters.**
Expert	The distinguished or brilliant journeyman, highly regarded by peers, **whose judgments are uncommonly accurate and reliable, whose performance shows consummate skill and economy of effort, and who can deal effectively with certain types of rare or "tough" cases.** Also, an expert is one who has special skills or knowledge derived from extensive experience with subdomains.	Enormous breadth and depth of knowledge. Routinely forms and leads high-performance teams. Admired by others as a benchmark of team performance. Performance standards are well beyond those of most practitioners.	Consistently solves difficult, complex, problems. **Able to handle novel or unusual situations.** Produces consistently inspiring and excellent performances.	Apprenticeship to masters. Advanced coaching, development of breadth, focus on observing and adopting style of the teacher. **Teaches others. Years or decades of practice.**
Journeyman (professional)	**Literally, a person who can perform a day's labor unsupervised**, although working under orders. An experienced and reliable worker, or one who has achieved a level of competence. It is possible to remain at this level for life.	Appropriate action is deliberate yet appears to come from experience and intuition. **Seldom thinks in terms of rules for governing behavior.** Considerable experience and practice across a wide range of situations over years of work.	Operates effectively while unsupervised. **Able to deal with more complex situations.** Individual performance is a benchmark for others.	Apprenticeship to experts. Coaching. Putting self into wide range of situations. **Membership and contribution to professional networks.** **Teaches others.**
Apprentice (advanced beginner)	**Literally, one who is learning**—a student undergoing a program of instruction beyond the introductory level. Traditionally, the apprentice is immersed in the domain by living with and assisting someone at a higher level. The length of an apprenticeship depends on the domain, ranging from about one to 12 years in the craft guilds.	Carries out standard actions without causing breakdowns. Can fulfill standard promises to customers satisfactorily without supervision. **Performs most standard actions without conscious application of rules.** When faced with a new situation, works out appropriate actions by application of rules.	Able to perform advanced problem-solving on projects when coached to do so. Can assist someone at a higher level.	**Repeated practice with common situations and increasing exposure to exceptional situations.** Apprenticeship to more advanced professionals and teams. **Membership in professional networks**
Novice (beginner)	**Literally, someone who is new**—a probationary member. There has been some ("minimal") exposure to the domain and/or has begun introductory instruction.	**All action appears to be governed by rules defining allowable actions and strategies.** Common situations are unfamiliar and are described by more rules. **Most action is deliberate application of rules** or conscious recall of prior actions in the familiar situations.	**Can contribute in realistic, well-understood situations with supervision.** Can perform simple actions for customers; needs supervision for more complex tasks.	**Memorization, drill, and practice in simple situations.** Problem-solving and practice with rules and strategies. **Learning involves recognizing common situations that help in recalling which rules should be exercised.**

*Taken verbatim from: Hoffman, R. R. (1998). How can expertise be defined?: Implications of research from cognitive psychology. In R. Williams, W. Faulkner, & J. Fleck (Eds.), Exploring expertise (pp. 81-100). New York: Macmillan.

**Adapted from: Denning PJ (2002) Career redux. Communications of the ACM 45(9):21-26.

Basic Levels of Proficiency

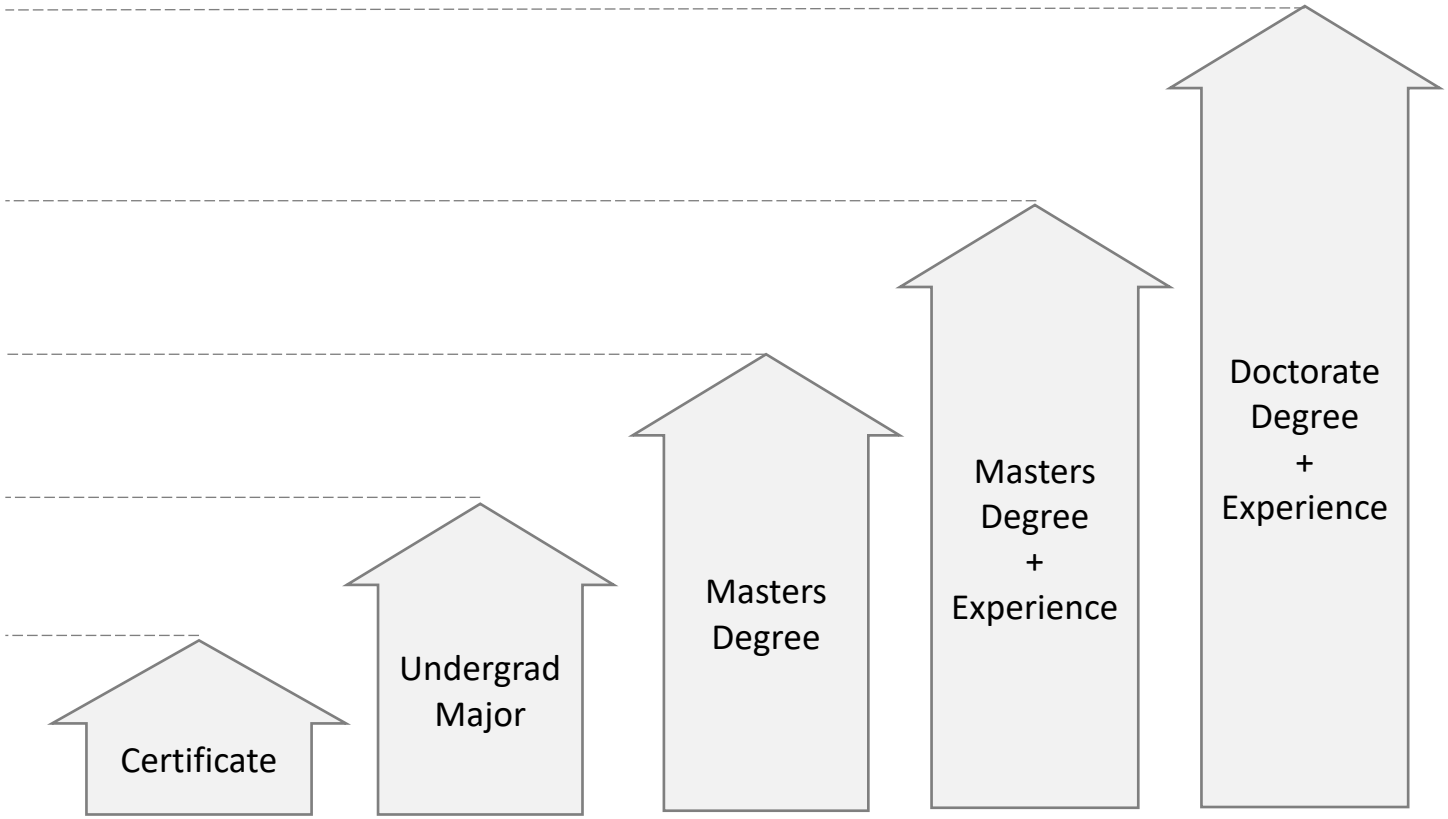
HIGH

LOW

	Traditional Description*
Master	Traditionally, a master **is one of an elite group of experts whose judgments set the regulations, standards, or ideals.** Also, a master can be that expert who is regarded by the other experts as being "the" expert, or the "real" expert, especially with regard to sub-domain knowledge.
Expert	The distinguished or brilliant journeyman, highly regarded by peers, **whose judgments are uncommonly accurate and reliable, whose performance shows consummate skill and economy of effort, and who can deal effectively with certain types of rare or "tough" cases**. Also, an expert is one who has special skills or knowledge derived from extensive experience with subdomains.
Journeyman (professional)	**Literally, a person who can perform a day's labor unsupervised**, although working under orders. An experienced and reliable worker, or one who has achieved a level of competence. It is possible to remain at this level for life.
Apprentice (advanced beginner)	**Literally, one who is learning**—a student undergoing a program of instruction beyond the introductory level. Traditionally, the apprentice is immersed in the domain by living with and assisting someone at a higher level. The length of an apprenticeship depends on the domain, ranging from about one to 12 years in the craft guilds.
Novice (beginner)	**Literally, someone who is new**—a probationary member. There has been some ("minimal") exposure to the domain and/or has begun introductory instruction.

We can use these levels to understand how education and experience contribute to the acquisition of proficiency in the field of operations analysis.

If an officer starts with little prior experience, the level of proficiency that can be attained scales approximately with the amount of education.



Note: Someone with a master's degree of education is typically not a "master" in this sense. Just as the rank of Master Chief Petty Officer follows a different notion of what it means to be a "master."

A Simplified View of Growth in OA Proficiency

Much easier with the support and mentorship of an expert!

Lots of work!
Lots of friction!

Junior Level

- Given clear problem
- Identify solution technique
- Execute and validate solution

Lots of work!
Lots of friction!

Mid Level

- Given vague goal
- Define problem
- Identify solution technique
- Execute and validate solution
- Assess tradeoffs

Senior Level

- Given chaos
- Identify what matters
- Define problem
- Structure solution
- Delegate execution

Technical competence → Operational competence → Strategic competence

What is the correct answer? *What is the real problem?* *What is success?*

What happens to skills when using automated tools?

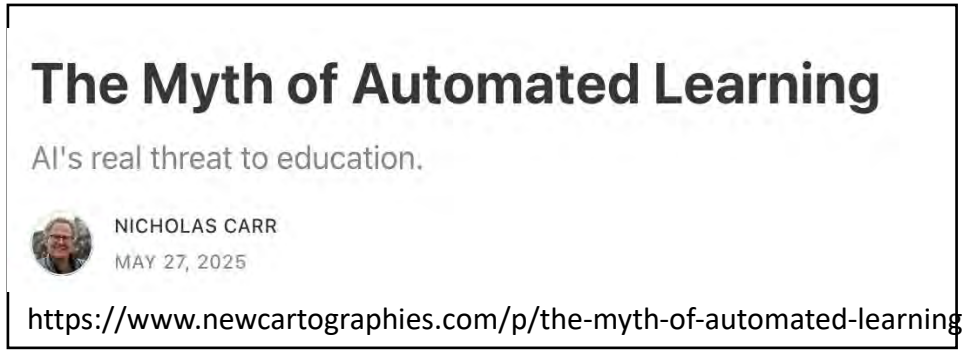
Three possible outcomes:

- **The skill is enhanced.** This tends to occur in individuals who already possess a certain level of proficiency and are engaged in activities that accelerate their practice.
- **The skill atrophies over time.** This is the classic *de-skilling paradox*.
- **The skill never develops** in the first place.

Areas of concern:

- Education: Reading. Analysis. Writing. Thinking.
- [Software-Based] Tool Development
- Information-Based Work
- Decision-making in high-stakes environments

https://biblioracle.substack.com/p/addressing-the-transactional-model



https://www.newcartographies.com/p/the-myth-of-automated-learning

Isn’t the current AI technology better than ever?
And getting better all the time?

(You mean, LLMs...?)

How do GPT / LLMs work?

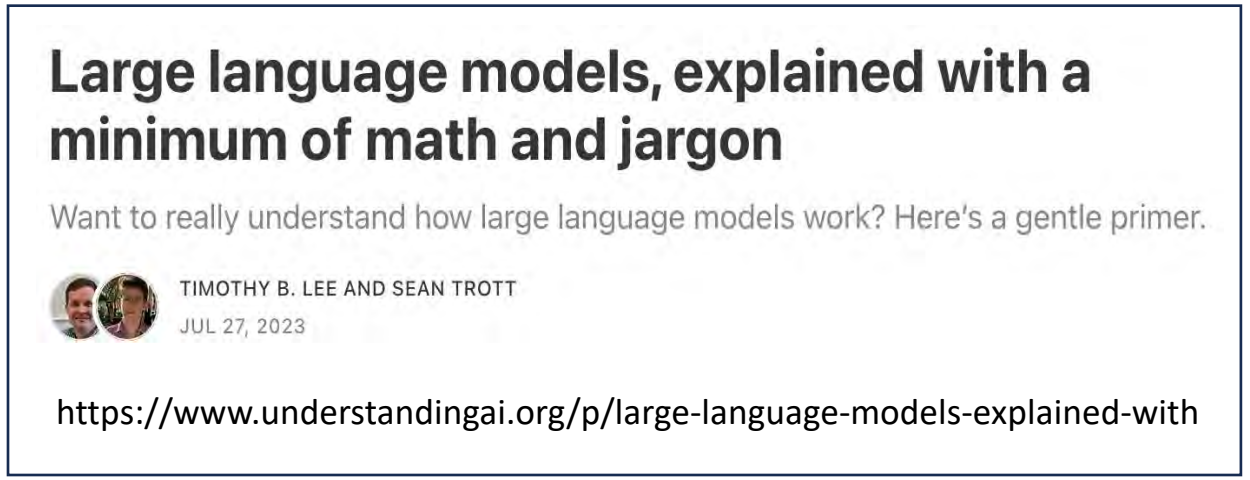
Token-predicting machines

- Tokens embedded in a high-dim vector space
- Layered neural network models
- Attention heads in neural network adjust weights between token vectors
- Feed-forward layers use vector math to “reason” about tokens
- Models with billions (trillions) of parameters

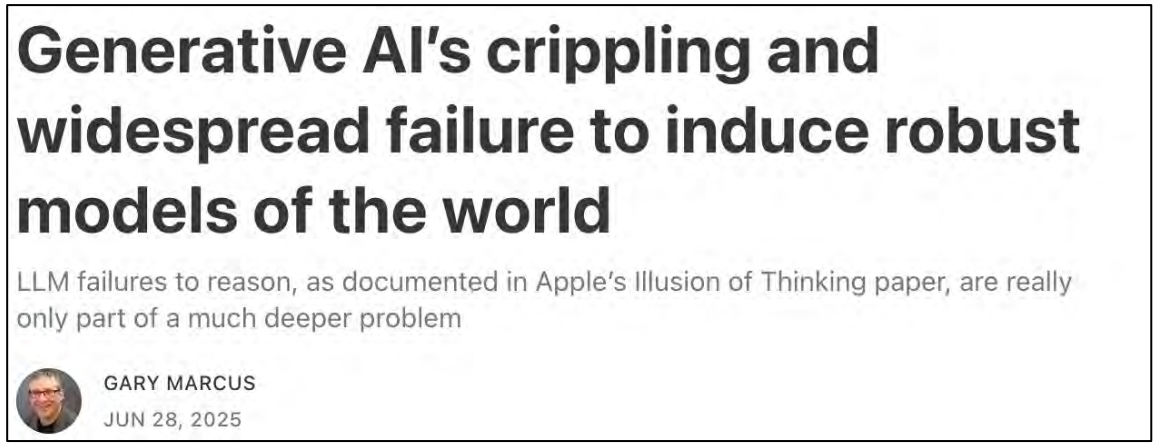
Requires scale!

- Trained on LOTS of data
- Using lots of compute (and energy)

Concern: LLMs nothing more than “stochastic parrots”…?



See also
https://arstechnica.com/science/2018/12/how-computers-got-shockingly-good-at-recognizing-images/



Generative AI's crippling and widespread failure to induce robust models of the world

LLM failures to reason, as documented in Apple's Illusion of Thinking paper, are really only part of a much deeper problem

GARY MARCUS
JUN 28, 2025

https://garymarcus.substack.com/p/generative-ais-crippling-and-widespread

*"A **world model** (or cognitive model) is a computational framework that a system (a machine, or a person or other animal) uses to track what is happening in the world."*

Each of us builds, maintains, and revises our own models for what is happening and how we make sense of the world around us.

The development of expertise is marked by how we organize information (i.e., build models) for improved retrieval and reasoning about what is happening and what might happen.

- **GPT-based LLMs do not build "world models"** in this way
- They are token-predicting machines (sometimes called "foundation models"—an unfortunate name)
- This limitation helps to explain their inability to effectively solve problems in many contexts
- **This is fundamental limitation** (i.e., it will not be solved with more data and more compute)

Generative AI's crippling and widespread failure to induce robust models of the world

LLM failures to reason, as documented in Apple's Illusion of Thinking paper, are really only part of a much deeper problem

GARY MARCUS
JUN 28, 2025

https://garymarcus.substack.com/p/generative-ais-crippling-and-widespread

What Has a Foundation Model Found? Using Inductive Bias to Probe for World Models

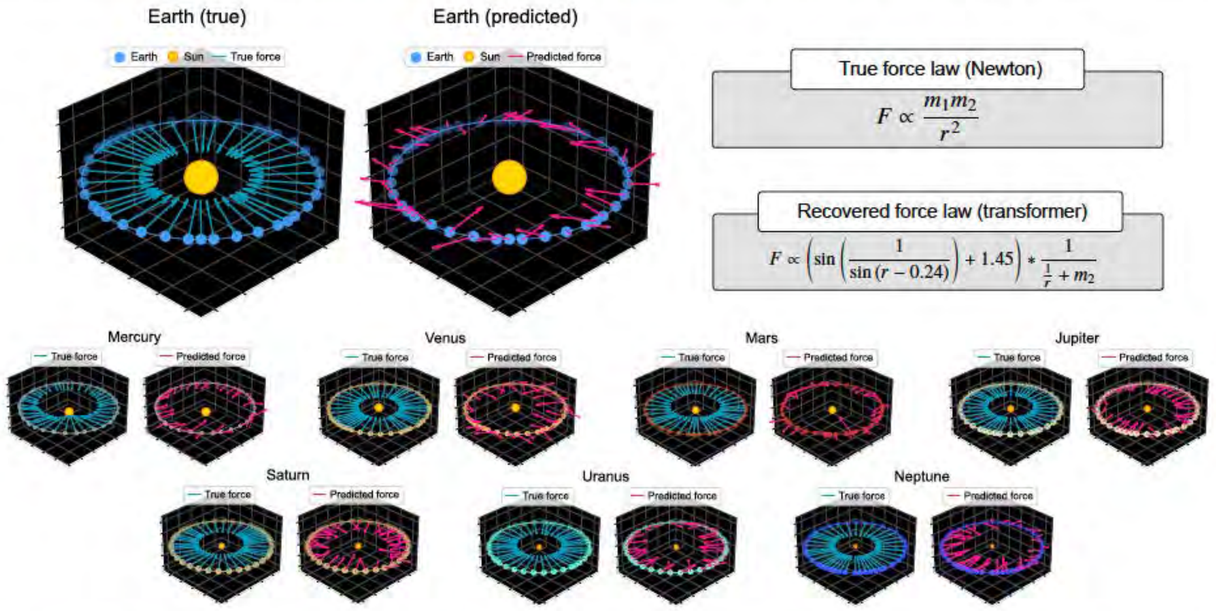


Figure 1: Each pair of panels illustrates the trajectory of a planet in the solar system and its gravitational force vectors, comparing the true Newtonian forces (left) to the predicted forces (right) from a transformer foundation model pretrained on orbital sequences and fine-tuned to predict forces. While the model excels at generating accurate predictions of planetary trajectories, it does not have an inductive bias toward true Newtonian mechanics; moreover, its force predictions recover a nonsensical force law, as revealed by symbolic regression.

What Has a Foundation Model Found? Using Inductive Bias to Probe for World Models

Keyon Vafa[1] Peter G. Chang[2] Ashesh Rambachan[2] Sendhil Mullainathan[2]

Abstract

Foundation models are premised on the idea that sequence prediction can uncover deeper domain understanding, much like how Kepler's predictions of planetary motion later led to the discovery of Newtonian mechanics. However, evaluating whether these models truly capture deeper structure remains a challenge. We develop a technique for evaluating foundation models that examines how they adapt to synthetic datasets generated from some postulated world model. Our technique measures whether the foundation model's inductive bias aligns with the world model, an
probe. A
foundatio
tasks yet
the unde
new task
models tr
fail to app
to new p
that thes
task-spec

1. Introduc

The promise o
sumption: tha
deeper truths,
this idea is new in one sense, it is old in another. Hundreds of years ago, astronomers like Kepler discovered geometric patterns that could pinpoint the future locations of planets in the night sky. Newton would later expand on this progress to develop Newtonian mechanics, fundamental laws that could not only predict the movement of planets but also explain physical properties across the universe (Koestler, 1959; Gin-

gerich, 2004). This path — from predicting sequences to understanding the deeper mechanisms that underlie them — is not unique to physics. In biology, animal breeders noticed patterns in the traits of offspring long before their predictive insights inspired Mendel to develop a theory of genetics.

How would we know if foundation models have also made the leap from making accurate predictions to developing reliable world models? This paper develops a framework for answering this question. Specifically, we create a procedure that, when given a foundation model and world model, tests whether the foundation model has learned that world model. We call this technique an *inductive bias probe*, and it is built on a simple insight: the implicit world model of a

model is toward Newtonian mechanics, it should have an inductive bias towards these force vectors. In contrast, Figure 1 shows that the model produces poor force vectors. More extremely, when we perform this exercise at a larger scale across many solar systems, the laws of gravity it uses to generalize bear no resemblance to Newton's law (Table 1).

We further apply inductive bias probes in other domains with a known world model: lattice problems and Othello games (Liu et al., 2022; Hazineh et al., 2023; Nanda et al., 2023b; Vafa et al., 2024). Across these domains, we find that neural sequence models have weak inductive biases

[1]Harvard University [2]MIT. Correspondence to: Keyon Vafa <kvafa@g.harvard.edu>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

"We particularly find that foundation models trained on orbital trajectories consistently fail to apply Newtonian mechanics when adapted to new physics tasks. Further analysis reveals that these models behave as if they develop task-specific heuristics that fail to generalize."

Generative AI's crippling and widespread failure to induce robust models of the world

LLM failures to reason, as documented in Apple's Illusion of Thinking paper, are really only part of a much deeper problem

GARY MARCUS

JUN 28, 2025

https://garymarcus.substack.com/p/generative-ais-crippling-and-widespread

"In short ChatGPT can *approximate* the game of chess, but it can't play it *reliably*, precisely because (despite immense relevant evidence) it never induces proper world model of the board and the rules."

Synthesized video from Dawid van Straaten, prompt ("Generate me a video of two men playing chess") in which the player for black reaches across the table and, in the midst of a rather unusual position moves his opponent's pawn horizontally, and quite illegally, several squares across the board.

Chap claims Atari 2600 'absolutely wrecked' ChatGPT at chess

1.19MHz eight-bit CPU trounced modern GPUs – can you do better with your retro-tech?

Simon Sharwood — Mon 9 Jun 2025 06:29 UTC

The Atari 2600 gaming console came into the world in 1977 with an eight-bit processor that ran at 1.19MhZ, and just 128 bytes of RAM – but that's apparently enough power to beat ChatGPT at chess.

So says infrastructure architect Robert Caruso, who over the weekend posted the results of an experiment he conducted to "pit ChatGPT against the Atari 2600's chess engine (via Stella emulator) and see what happens."

> **ChatGPT confused rooks for bishops, and repeatedly lost track of pieces**

Caruso decided to run the experiment after conversing with ChatGPT about the history of chess. At some point in that chat, the bot volunteered to play against the Atari – a reasonable suggestion as "Video Chess" was one of the games Atari commissioned for its console.

Online chat in which chess wonks discuss the merits of Video Chess suggest it may have played at a level beginners may have found challenging, and perhaps gave regular recreational players of intermediate skill a little to worry about.

Caruso thought his experiment would be "a lighthearted stroll down retro memory lane."

Instead, he watched as the Atari humiliated ChatGPT.

https://www.theregister.com/2025/06/09/atari_vs_chatgpt_chess/

Example of a "World Model"? Mathematical Optimization (MO)!

The starting point for a MO problem is *model formulation*:

- Index sets
- Parameters
- Variables
- Objective
- Constraints

Building a "good" representation of the problem (i.e., a world model) is the essential first step toward being able to optimize a decision.

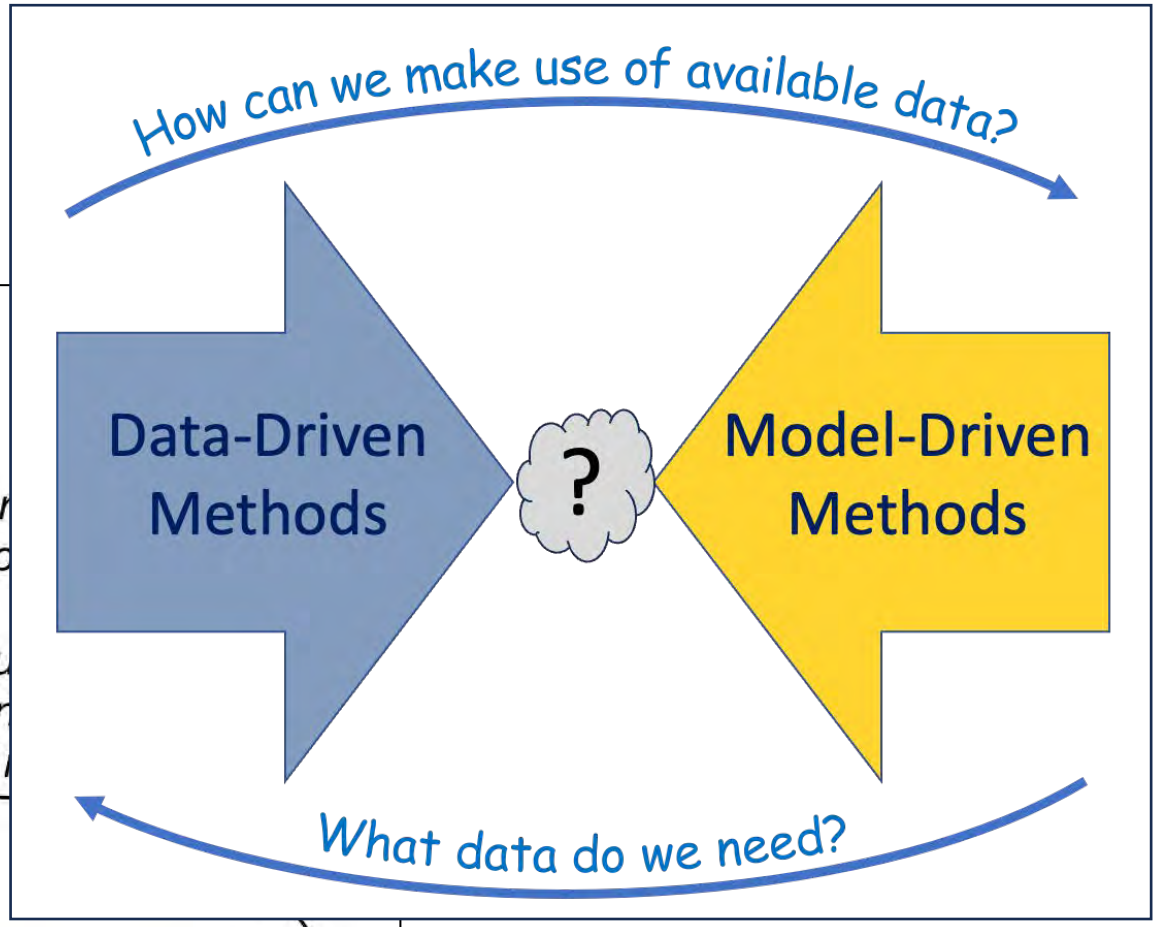
This requires an understanding of the underlying mathematics *and* domain expertise in the problem at hand.

The benefits and advantages of having a "world model" are immediate.

Mathematical Optimization (MO) vs. Large Language Models (LLMs)

Mathematical Optimization (MO)	vs.	Large Language Models (LLMs)
• When MO finds an optimal solution, it provides a certificate that guarantees it is the best. In other cases, it can potentially tell you how far away the solution is from the best one.		• An LLM provides no guarantees on its output and requires a human to interpret the results to determine whether the solution is any good or a nonsense hallucination.
• MO will tell you if a problem is infeasible. If infeasible, the model can be interrogated to understand what would be required to make it feasible.		• An LLM can't say whether a problem is infeasible or not and will likely produce a plausible-sounding answer in either case.
• If an MO problem goes wrong, there is typically a signal about what went wrong and what to do.		• LLMs don't provide such alarm signals.
• The results of MO are repeatable. Different solves (either by different people or at different times) will get the same answer.		• LLMs are not stable under repeated invocations. Different solves (either by different people or at different times) are likely to get different answers.
• MO is transparent in the sense that we can understand how the solution was obtained and explain why it is good.		• LLMs are not transparent. It is unclear how or why they produce their results, and explaining their output can be difficult.
• We can use MO to do parametric analysis: understand how the best solution changes as inputs (e.g., available resources) or requirements change. MO supports systematic "what-if" analysis that allows users to explore future uncertainties.		• LLMs do not support this.
• We can use MO to find worst-case failures in operational plans, either for attack or defense.		• LLMs do not support this.
• MO has mature theory that is decades old and is understood and supported by thousands of practitioners worldwide.		• LLMs are poorly understood, relatively speaking, and it is unclear who to call when things don't work out as planned.

Potential Limits in the current AI Gold Rush?



In hindsight

- **The bitter lesson about the bitter lesson** is *that scaling wor*
  for some problems — mainly pattern recognition — but not c
- *Scaling has made GenAI much better at what it is good at, bu*
  much with its core weaknesses around reliability, hallucinatio
  reasoning, poor planning, etc, or its challenges in the face of

Problems scale helps with

Object recognition, autocomplete, voice recognition, brainstorming, etc

Problems scale doesn't help with

open-ended problems that cannot be beaten by brute force; reasoning with incomplete information, etc

Source: Gary Marcus, 26 SEP 2025, https://garymarcus.substack.com/p/game-over-for-pure-llms-even-turing

Key Points

- AI is valued as an ***automation technology***
- The capabilities of these tools are increasing, and there is much that can be done in the hands of a ***skilled user***
- But there are limitations in what can be automated
  - Framing
  - Context
  - Learning
- LLMs, in particular, are fundamentally limited in their capability
- Lots of implications for the ***expertise*** of operations analysts
  - Need experience and mentorship, not automation
  - Need more than good technology to thrive in current AI Gold Rush

What's Missing: Challenges and Opportunities

- "How do we know if it works?"
  - Test & Evaluation
  - Verification & Validation

  … For AI-enabled Workflows
- Design for Joint Human-AI Performance
- AI Reliability & Safety
- Control Theory for Agentic Workflows
  - Observability, Controllability, Stability
- Understanding when "going slow" is a feature, not a bug…

I welcome comments and feedback. Thank you!

- Dr. David Alderson
  Professor and Chair, Operations Research
  Naval Postgraduate School
  dlalders@nps.edu
  http://faculty.nps.edu/dlalders

Check out our OR Hiring Webpage for job announcement details!

https://www.nps.edu/web/or/hiring