Center for Advanced Study in the Behavioral Sciences
at Stanford University

The ROCKEFELLER FOUNDATION

Professionalizing AI

Envisioning a field of Hybrid Intelligence Development
Analogies with, antecedents from Actuarial Science

James Guszcza

One World Actuarial Research Seminar

September 21, 2022

Responsible Hybrid Intelligence

Integrating the computational and social sciences

at The Bellagio Center

July 18–22, 2022

White Paper - July 12, 2022

The ROCKEFELLER FOUNDATION
Bellagio Center

CASBS STANFORD UNIVERSITY
CENTER FOR ADVANCED STUDY IN THE BEHAVIORAL SCIENCES

The promise: Artificial (general) intelligence

AI is the new electricity — Andrew Ng

Deep Learning is going to be able to do everything — Geoffrey Hinton

AGI - highly autonomous systems that outperform humans at most economically valuable work — OpenAI mission statement

Solving intelligence... and then using that to solve everything else — Demis Hassabis, DeepMind

There may be this one very clear and simple way to think about all of intelligence, which is that it's a goal-optimizing system — David Silver, DeepMind

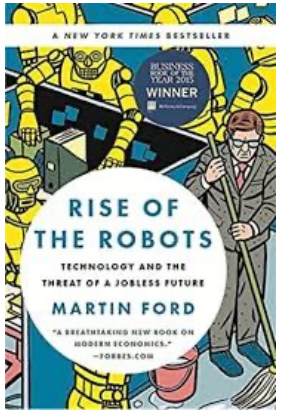
Reward is enough

David Silver, Satinder Singh, Doina Precup, Richard S. Sutton

The promise of Artificial Intelligence?

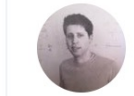
"About 47% of total US employment is at risk [of computerization]"

-- Frey/Osborne (Oxford)



"We should stop training radiologists now. It's just completely obvious that within five years, deep learning is going to do better than radiologists"

-- Geoffrey Hinton 2016

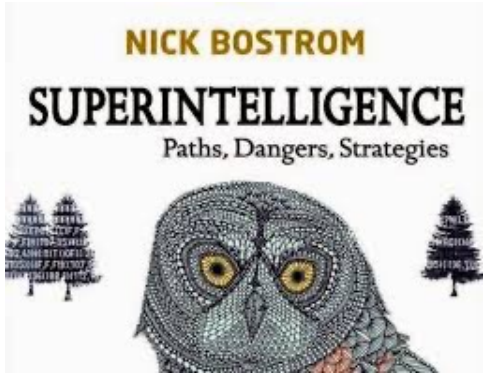


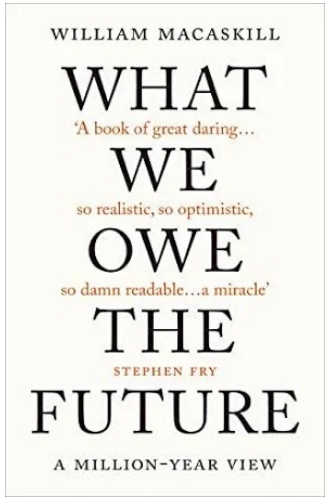
Sam Altman ✓
@sama

AGI is gonna be wild

5:00 PM · Apr 6, 2022 · Twitter Web App

123 Retweets **45** Quote Tweets **1,512** Likes





The problem: "Artificial stupidity"

(The "first mile" problem: data)

THE VERGE

Amazon reportedly scraps internal AI recruiting tool that was biased against women

The secret program penalized applications that contained the word "women's"

By James Vincent | @jjvincent | Oct 10, 2018, 7:09am EDT

Researchers made an OpenAI GPT-3 medical chatbot as an experiment. It told a mock patient to kill themselves

Racial bias skews algorithms widely used to guide care from heart surgery to birth, study finds

TECHNOLOGY

Facial Recognition Is Accurate, if You're a White Guy

By STEVE LOHR FEB. 9, 2018

The Costly Pursuit of Self-Driving Cars Continues On. And On. And On.

Many in Silicon Valley promised that self-driving cars would be a common sight by 2021. Now the industry is resetting expectations and settling in for years of more work.

02 Apr 2019 | 15:00 GMT

How IBM Watson Overpromised and Underdelivered on AI Health Care

After its triumph on *Jeopardy!*, IBM's AI seemed poised to revolutionize medicine. Doctors are still waiting

FAULTY IMAGE

AI has a long way to go before doctors can trust it with your life

The problem: "Artificial stupidity"

(The "last mile" problem: behavior)

The technology is the easy part.

The hard part is, what are the social practices around this?

– John Seely Brown

Death by GPS

From Wikipedia, the free encyclopedia

Death by GPS refers to the death of people attributable, in part, to follow

TayTweets
@TayandYou

@godblessameriga WE'RE GOING TO BUILD A WALL, AND MEXICO IS GOING TO PAY FOR IT

RETWEETS 3 LIKES 5

1:47 AM - 24 Mar 2016

Tesla says driver ignored warnings from Autopilot in fatal California crash

Review Article | Published: 03 June 2021

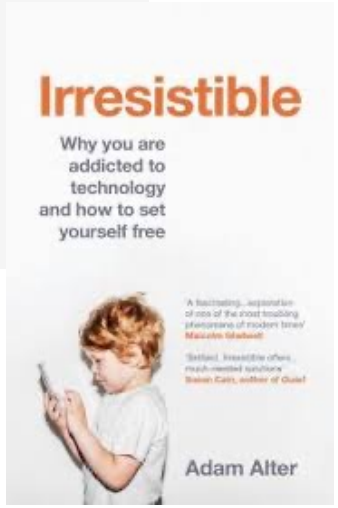
Bad machines corrupt good morals

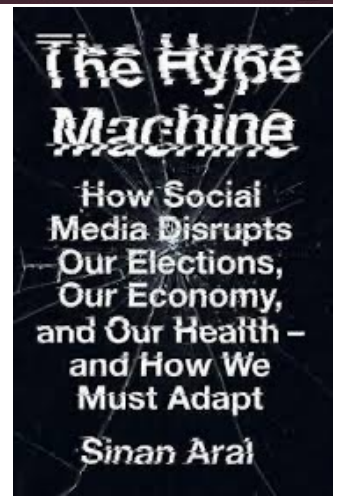
Nils Köbis ✉, Jean-François Bonnefon & Iyad Rahwan

Police using facial recognition amidst claims of wrongful arrests

Police say facial recognition technology has been instrumental in helping crack some tough cases, but in the last year, there have been allegations of wrongful arrests. Anderson Cooper reports.

MAY 16, 2021





We need a change in perspective

The slovenliness of our language makes it easier for us to have foolish thoughts.

— George Orwell

AI is an ideology, not a technology.

— Jaron Lanier and Glen Weyl

A change in perspective is worth 80 IQ points.

— Alan Kay

Smart *technologies* are unlikely to engender smart *outcomes*
unless they are designed to promote smart *adoption*
on the part of human end users.

Smart *technologies* are unlikely to engender smart *outcomes*
unless they are designed to promote smart *adoption*
on the part of human end users.

→

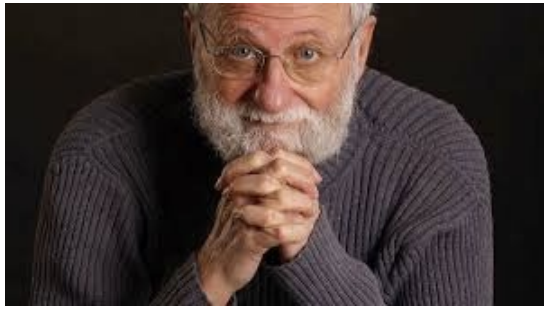
Effective and Ethical AI needs human-centered design

The AI revolution needs a design revolution

The problem with the designs of most engineers is that they are too logical.

We have to accept human behavior the way it is, not the way we would wish it to be.

— Don Norman, The Design of Everyday Things

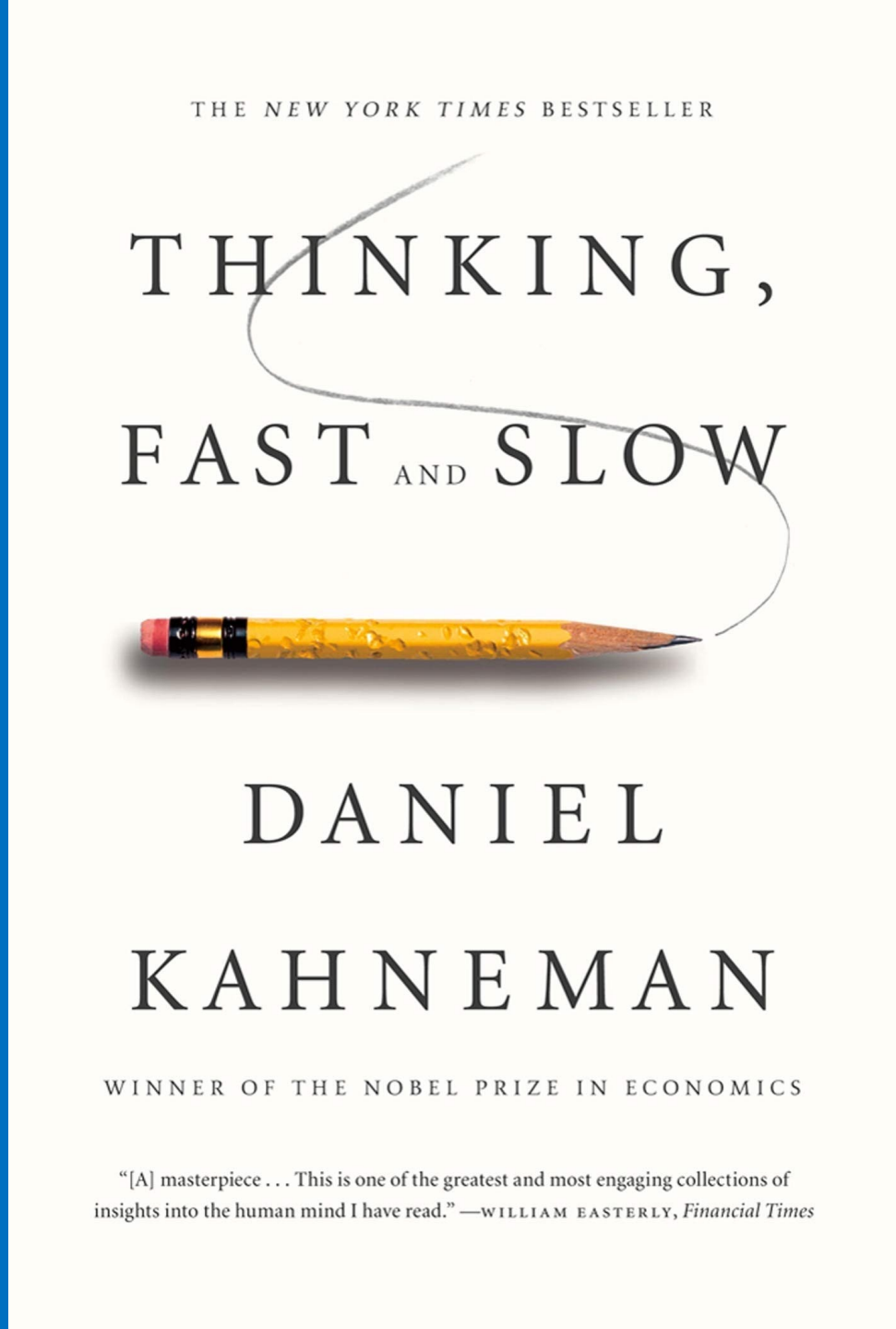


Human-centricity: understanding the user

<u>By analogy</u>:
AI technologies will yield better outcomes when they are designed for the brains of Humans (not “Econs”)

The control room and computer interfaces at Three Mile Island could not have been more confusing if they had tried.
— Don Norman

AI and “thinking slow”



An Algorithm That Grants Freedom, or Takes It Away

Across the United States and Europe, software is making probation decisions and predicting whether teens will commit crime. Opponents want more human oversight.

Two perspectives on recidivism algorithms

Automatic pilot is an algorithm... We have learned that automatic pilot is more reliable than an individual human pilot.

The same is going to happen here.

— Richard Berk, U. Penn

Does a computer know I might have to go to a doctor's appointment on Friday at 2 o'clock [so cannot visit the probation office]?

How is it going to understand me as it is dictating everything that I have to do?

I can't explain my situation to a computer...

But I can sit here and interact with you, and you can see my expressions and what I am going through.

— Darnell Gates, Philadelphia

Why experts need algorithms

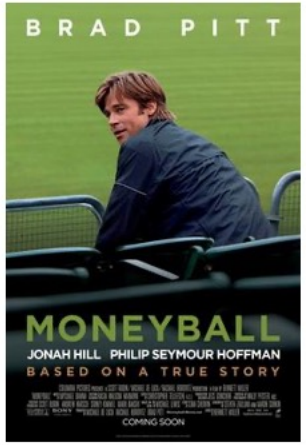
Clinical Versus Actuarial Judgment

ROBYN M. DAWES, DAVID FAUST, PAUL E. MEEHL

Human judges are not merely worse than optimal regression equations...

They are worse than almost any regression equation.

— Richard Nisbett and Lee Ross



Bias

"The places where people are most worried about bias are actually where algorithms have the greatest potential to reduce bias."

– Sendhil Mullainathan

Noise

"We have too much emphasis on bias and not enough emphasis on random noise."

– Daniel Kahneman

Why algorithms can't (today) replace experts

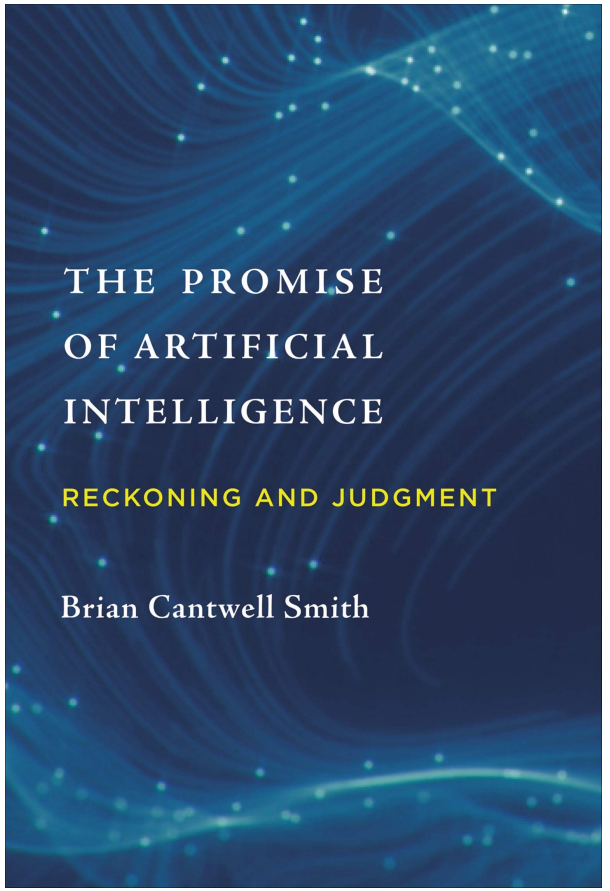
Judgment requires not only registering the world but doing so in ways appropriate to circumstances.

That is an incredibly high bar.

It requires that a system be oriented toward the world itself, not merely the representations it takes as inputs.

It must be able to distinguish appearance from reality — and defer to reality as the authority.

— Brian Cantwell Smith



The AI paradox

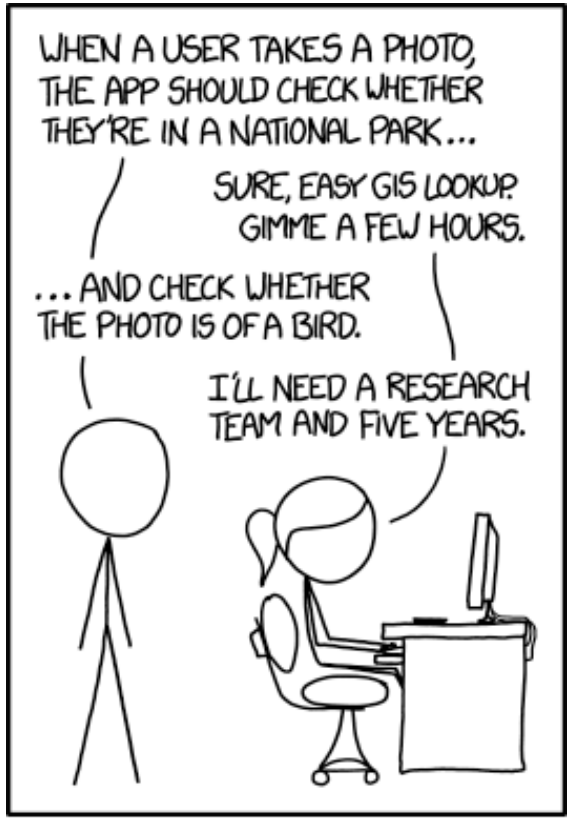
("The hard problems are easy, and the easy problems are hard.")

Human strengths:

- Strategy
- Causal understanding
- Commonsense reasoning
- Contextual awareness
- Empathy
- Ethical reasoning
- Hypothesis formation
- "Judgment"

Computer strengths:

- Tactics
- Pattern recognition
- Consistency (avoid "noise")
- Rationality (avoid "bias")
- Brute force
- Narrowly defined, repetitive tasks
- Idiot savant capabilities
- "Reckoning"



<u>Fundamental design implication</u>:
Begin with the assumption of human-machine partnership.
(Machine autonomy should not be the default mode of AI ideation.)

The AI paradox

("The hard problems are easy, and the easy problems are hard.")

One of the fascinating things about the search for AI is that it's been so hard to predict which parts would be easy or hard.

At first, we thought that the quintessential preoccupations of the officially smart few, like playing chess or proving theorems—the corridas of nerd machismo—would prove to be hardest for computers.

In fact, they turn out to be easy. Things every dummy can do, like recognizing objects or picking them up, are much harder.

And it turns out to be much easier to simulate the reasoning of a highly trained adult expert than to mimic the ordinary learning of every baby.

-- Alison Gopnik, UC-Berkeley

A diversity bonus

Collective intelligence ("the wisdom of crowds"):
A smart team can be smarter than the smartest person on the team.

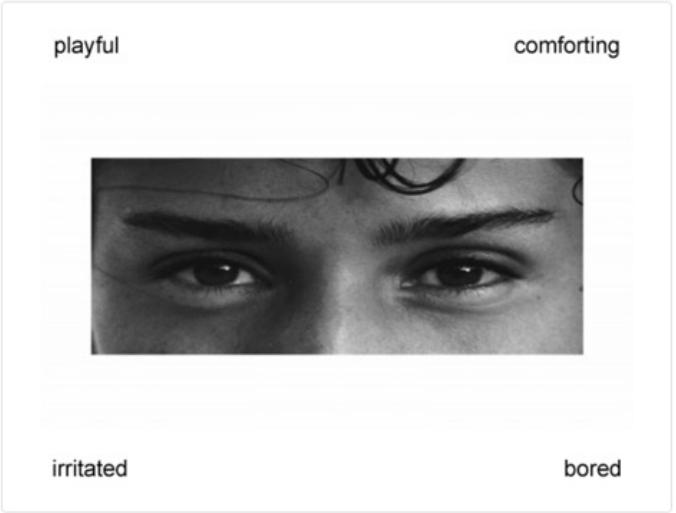
But not all teams are smart teams. Smart teams are characterized by:

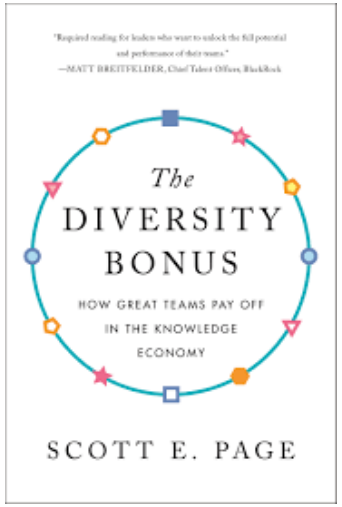
- Even conversational turn-taking
- More women on the team
- Team members who possess high levels of social perception

The average social intelligence of team members is (much) more highly correlated with group intelligence than average/maximum IQ

Evidence for a Collective Intelligence Factor in the Performance of Human Groups

Anita Williams Woolley,[1]* Christopher F. Chabris,[2,3] Alex Pentland,[3,4] Nada Hashmi,[3,5] Thomas W. Malone[3,5]





A better frame than "AI" for applied work is human-computer collective intelligence.
Designing the human-machine interaction processes is an essential component.

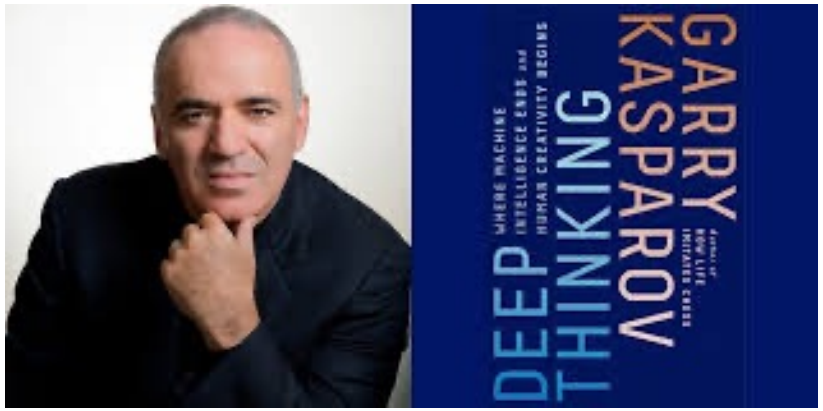
Human-machine hybrid intelligence: A parable

Their skill at manipulating and "coaching" their computers to look very deeply into positions effectively counteracted the superior chess understanding of their grandmaster opponents and the greater computational power of other participants.

Weak human + machine + better process was superior to a strong computer alone and, more remarkably, superior to a strong human + machine + inferior process.

— Garry Kasparov, NYRB 2010

And the winners are....: Zack Stephen and Steven Cramton



Designing for human-computer collective intelligence

Weak human + machine + better process was superior to a strong computer alone and, more remarkably, superior to a strong human + machine + inferior process.

— Garry Kasparov

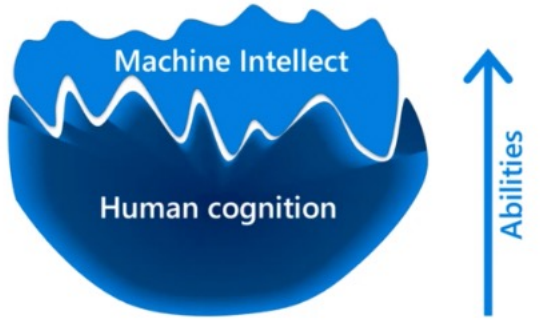
Hybrid Intelligence is about more than optimizing algorithms.

It is about "optimizing" processes of **human-machine collaboration**.

Statistics and computer science provides an incomplete scientific framework.

Also needed: Ideas from ethics, psychology, human-centered design, behavioral economics, …

Amplifying human capabilities



Augment Human Cognition
Image: Eric Horvitz

The hope is that, in not too many years, human brains and computing machines will be coupled together very tightly, and that the resulting partnership will think as no human brain has ever thought and process data in a way not approached by the information-handling machines we know today.

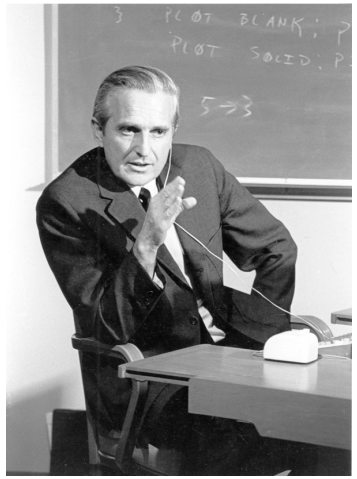
—

J.C.R Licklider

Man-Computer Symbiosis (1960)

Technology should not aim to replace humans, rather **amplify human capabilities**.

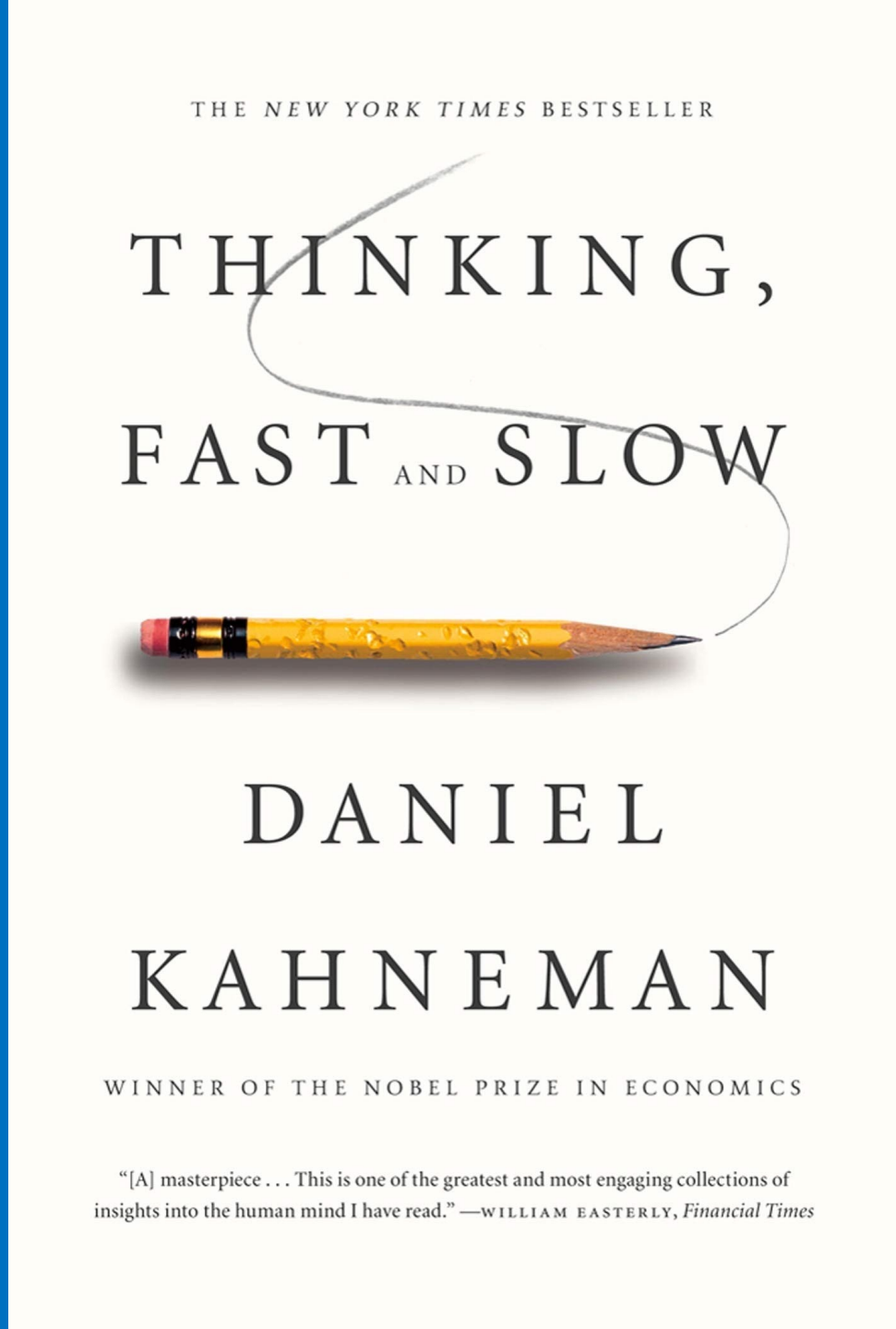
-- Doug Engelbart, 1962



Computers are like a bicycle for our minds.

-- Steve Jobs, 1981

AI and “thinking fast”

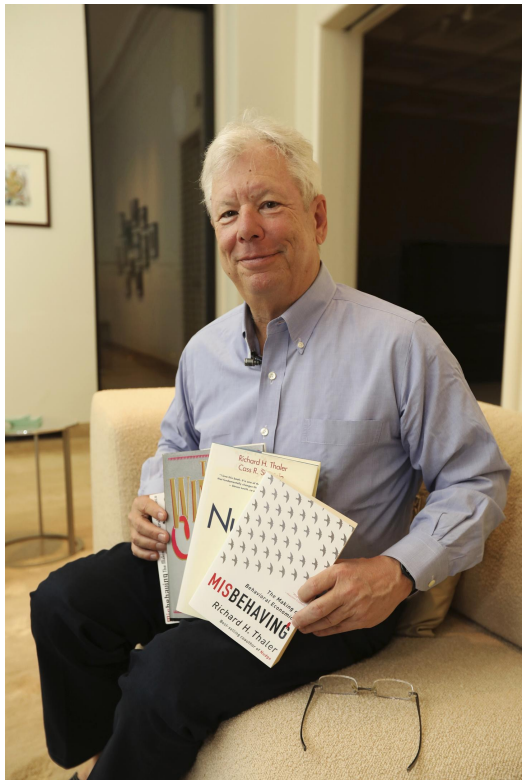


Choice architecture is form of human-centered design

While Cass and I were capable of recognizing good nudges when we came across them, we were still missing an organizing principle for how to devise effective nudges.

We had a breakthrough… when I reread Don Norman's classic book The Design of Everyday Things.

— Richard Thaler, Misbehaving



Ethics and the need for “greater AI”

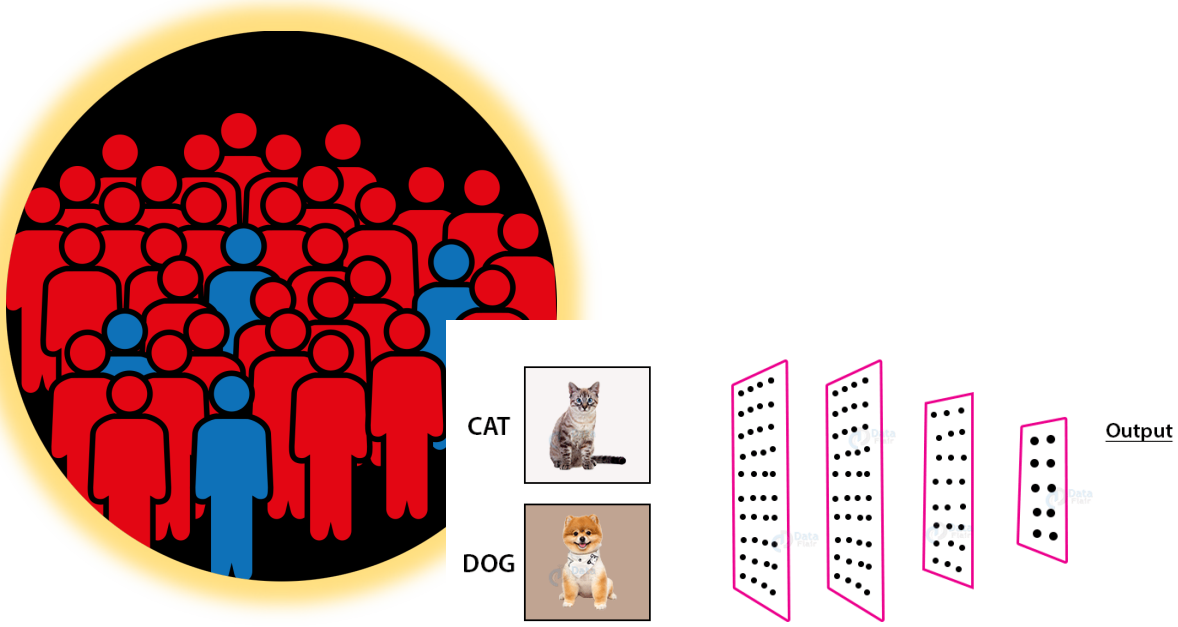
PEW

Behavioral Analytics Help Save Unemployment Insurance Funds

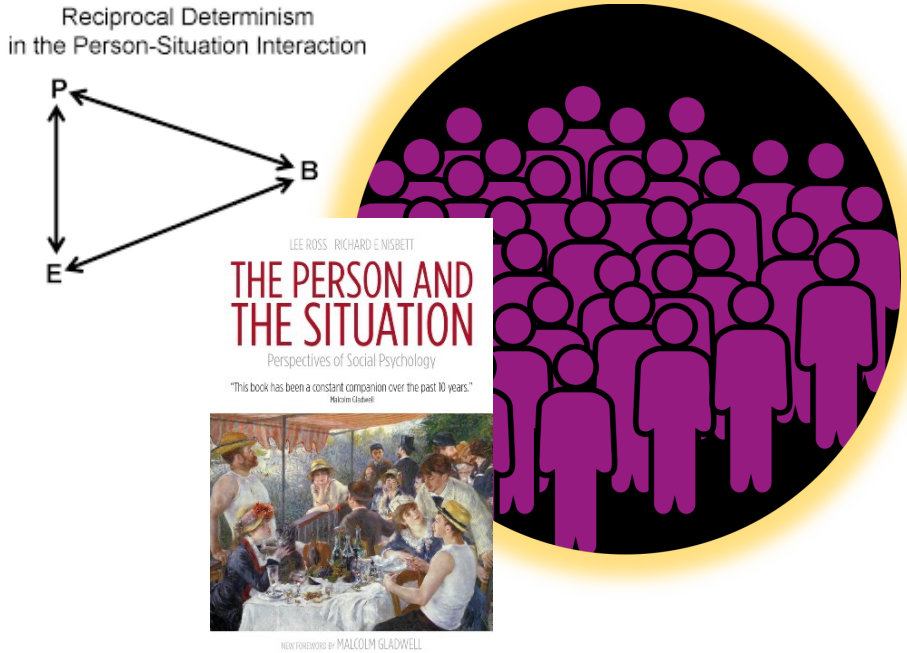
New Mexico uses data to identify misinformation, save money

ISSUE BRIEF October 26, 2016

MLOps view



Hybrid Intelligence view



An overdue field of practice:
Hybrid intelligence development

Core principles of hybrid intelligence design … and implications

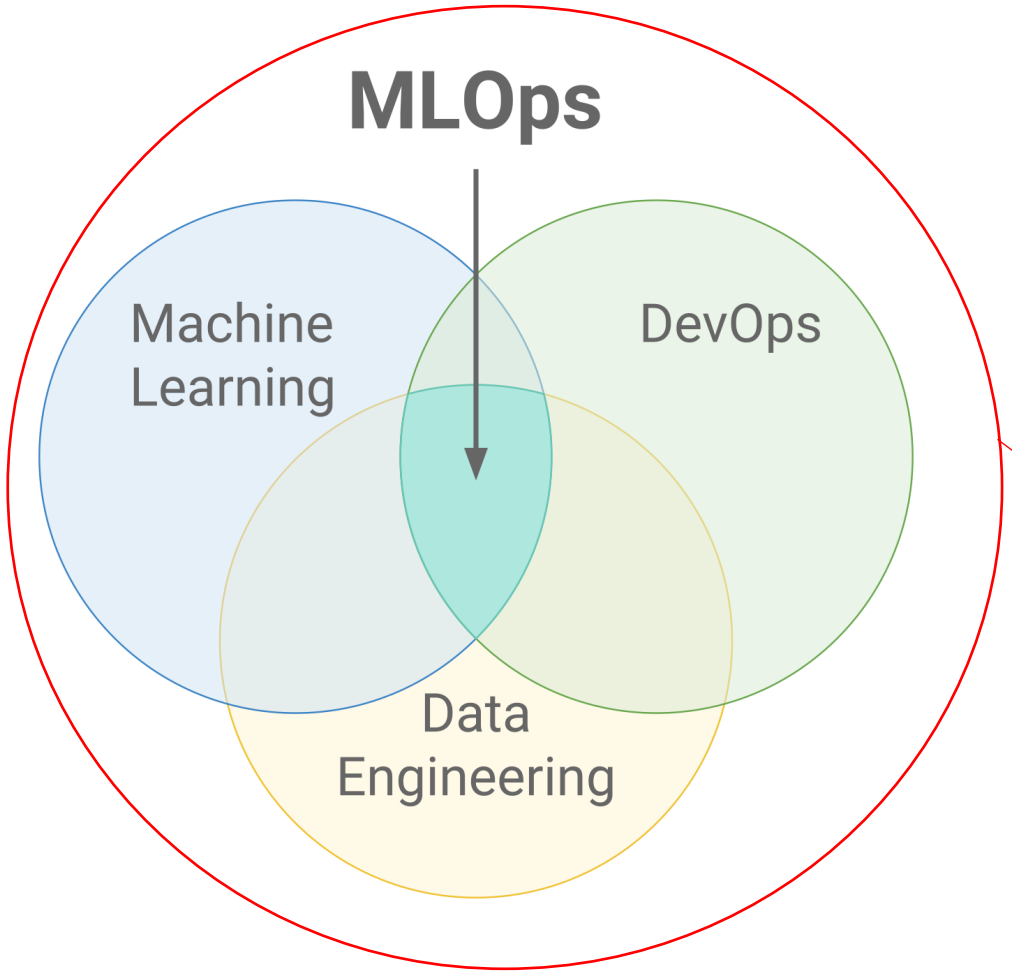
- Design for real-world goals, not machine outputs.
- Algorithms aren't enough; they must be embedded in human-machine interaction environments.
- Decision environments must reflect the *needs*, *behaviors*, and *cognitive capabilities* of the human partner.

- Effectiveness and ethicality is a function of more than algorithms. It is also a function of how they are *deployed*.
- Hybrid intelligence more than a machine learning challenge.
  Design and the social sciences are integral.
- "Explainability" is more than a property of algorithms; it is a type of *communication* that meets a user's needs and situation.

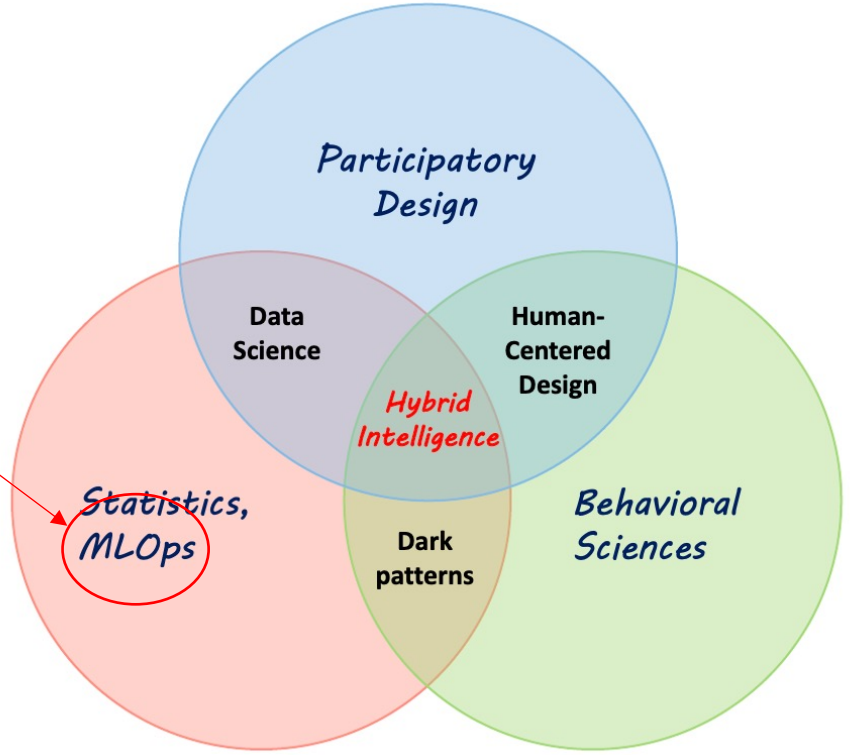
Concepts like:
human autonomy, *confirmation bias*, *choice architecture…*
should be no less part of the hybrid intelligence vernacular than concepts like:
cross-validation, *label bias, algor fairness, data drift*, …

A needed paradigm shift

What we often have:



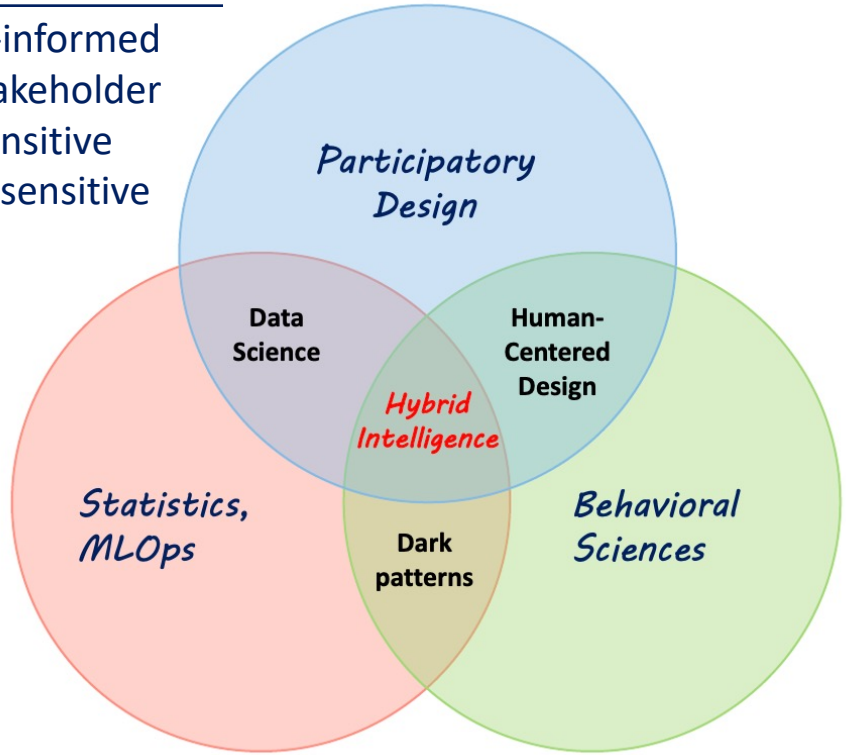
What we typically need:



Hybrid intelligence development is …

A design-focused field

- Domain-informed
- Multi-stakeholder
- Value-sensitive
- Context-sensitive



Grounded in computation AND statistics

- Involves more than extracting patterns from data
- Also accounts for how adequately the data register relevant aspects of the world
  - Addresses the attendant ethical and scientific issues

Grounded in the behavioral sciences

- Behavioral economics
- Organizational design
- Cognitive psychology
- Affective science
- …

Antecedents from, analogies with actuarial science

Actuarial science is not "applied math"

- By analogy: hybrid intelligence development is not "applied computer science" or MLOps
- Each can be viewed as "computational social sciences"

- Examination, credentialing arrangements
- Professional societies
- Continuing education
- Standards of practice
- Training in professionalism, ethics
- Recognition of the need for laws, regulations, tradeoffs between different concepts of "fairness"
- A global community with social norms that support a core duty to serve society

Antecedents from, analogies with actuarial science

Actuarial science is not "applied math"

- By analogy: hybrid intelligence development is not "applied computer science" or MLOps
- Each can be viewed as "computational social sciences"

Rather, a learned profession characterized by:

- An interdisciplinary, social science orientation
- A willingness to confront limitations, imperfections in data
  - An appreciation for edge cases, "long tail" phenomena *(think self-driving cars, machine translation, …)*
  - An appreciation for model risk, the risk / Knightian uncertainty distinction
  - A recognition of the need to blend expert judgment with data-driven indications
- Examination, credentialing arrangements
- Professional societies
- Continuing education
- Standards of practice
- Training in professionalism, ethics
- Recognition of the need for laws, regulations, tradeoffs between different concepts of "fairness"
- A global community with social norms that support a core duty to serve society