Center for Advanced Study in the Behavioral Sciences
at Stanford University

AICT

Human-centered

why ethical and effective AI needs behavioral design

Jim Guszcza
AICT/CAS
Joint Seminar
September 8, 2021

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

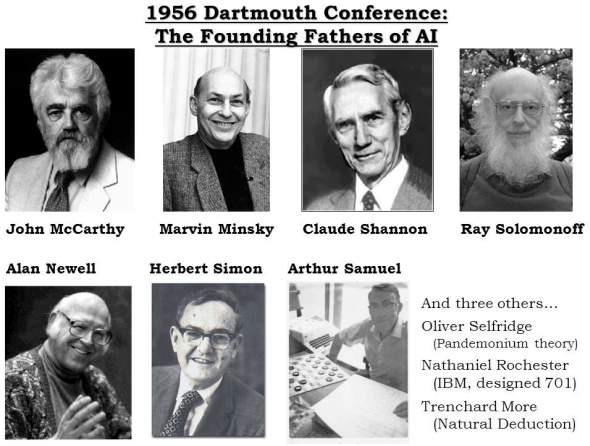
Solving intelligence... and then using that to solve everything else — Demis Hassabis, DeepMind

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

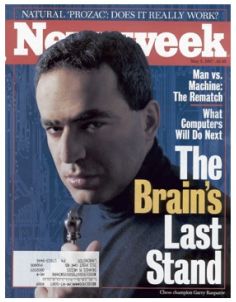
David Silver, Satinder Singh, Doina Precup, Richard S. Sutton

The AI “master narrative”



“Every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it”

-- 1956 Dartmouth Conference



“Google’s AlphaGo is demonstrating for the first time that machines can truly learn and think in a human way”

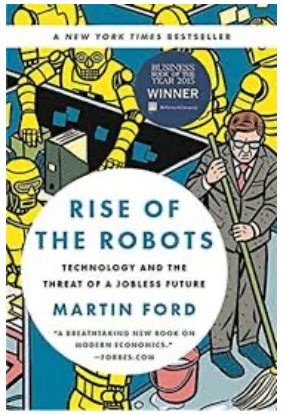
-- New York Times, March 2016

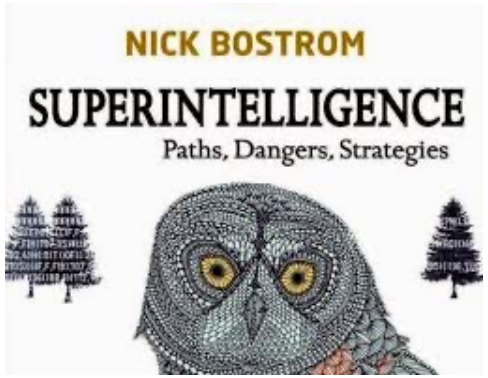
“Before the prospect of an intelligence explosion we humans are like small children playing with a bomb”

-- Nick Bostrom (Oxford)

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

-- Frey/Osborne (Oxford)



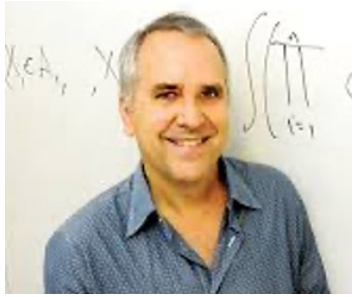


“AI” and its discontents

Public dialog… too often uses the term AI as an intellectual wildcard, one that makes it difficult to reason about the scope and consequences of emerging technology…

This is not the classical case of the public not understanding the scientists—here the scientists are often as befuddled as the public.

— Michael Jordan, UC Berkeley



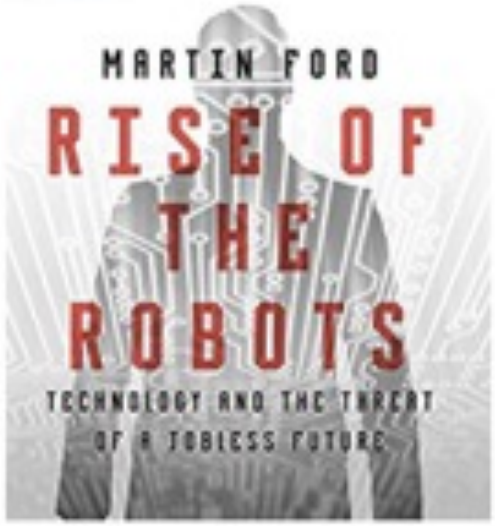
AI is an ideology, not a technology.

— Jaron Lanier and Glen Weyl

AI as “automation”

“… about 47% of total US employment is at risk.”

-- Frey/Osborne (Oxford U)



THE FUTURE OF EMPLOYMENT: HOW SUSCEPTIBLE ARE JOBS TO COMPUTERISATION?*

Carl Benedikt Frey† and Michael A. Osborne‡

September 17, 2013

Abstract

We examine how susceptible jobs are to computerisation. To assess this, we begin by implementing a novel methodology to estimate the probability of computerisation for 702 detailed occupations, using a Gaussian process classifier. Based on these estimates, we examine expected impacts of future computerisation on US labour market outcomes, with the primary objective of analysing the number of jobs at risk and the relationship between an occupation's probability of computerisation, wages and educational attainment. According to our estimates, about 47 percent of total US employment is at risk. We further provide evidence that wages and educational attainment exhibit a strong negative relationship with an occupation's probability of computerisation.

Can underwriting be “computerized”?

THE FUTURE OF EMPLOYMENT: HOW SUSCEPTIBLE ARE JOBS TO COMPUTERISATION?*

Carl Benedikt Frey† and Michael A. Osborne‡

September 17, 2013

	Computerisable			
Rank	Probability	Label	SOC code	Occupation
1.	0.0028		29-1125	Recreational Therapists
2.	0.003		49-1011	First-Line Supervisors of Mechanics, Installers, and Repairers
3.	0.003		11-9161	Emergency Management Directors
4.	0.0031		21-1023	Mental Health and Substance Abuse Social Workers
5.	0.0033		29-1181	Audiologists
6.	0.0035		29-1122	Occupational Therapists
7.	0.0035		29-2091	Orthotists and Prosthetists
8.	0.0035		21-1022	Healthcare Social Workers
9.	0.0036		29-1022	Oral and Maxillofacial Surgeons
10.	0.0036		33-1021	First-Line Supervisors of Fire Fighting and Prevention Workers
11.	0.0039		29-1031	Dietitians and Nutritionists
12.	0.0039		11-9081	Lodging Managers
13.	0.004		27-2032	Choreographers
695.	0.99		13-2082	Tax Preparers
696.	0.99		43-5011	Cargo and Freight Agents
697.	0.99		49-9064	Watch Repairers
698.	0.99		13-2053	Insurance Underwriters
699.	0.99		15-2091	Mathematical Technicians
700.	0.99		51-6051	Sewers, Hand
701.	0.99		23-2093	Title Examiners, Abstractors, and Searchers
702.	0.99		41-9041	Telemarketers

The most important aspect of a statistical analysis is not what you do with the data, it's what data you use

NOVEMBER 15, 2017

Stanford algorithm can diagnose pneumonia better than radiologists

Stanford researchers have developed a deep learning algorithm that evaluates chest X-rays for signs of disease. In just over a month of development, their algorithm outperformed expert radiologists at diagnosing pneumonia.

"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

03 May 2021 | 18:30 GMT

Andrew Ng X-Rays the AI Hype

AI pioneer says machine learning may work on test sets, but that's a long way from real world use

"It turns out," Ng said, "that when we collect data from Stanford Hospital, then we train and test on data from the same hospital, indeed, we can publish papers showing [the algorithms] are comparable to human radiologists in spotting certain conditions."

But, he said, "It turns out [that when] you take that same model, that same AI system, to an older hospital down the street, with an older machine, and the technician uses a slightly different imaging protocol, that data drifts to cause the performance of AI system to degrade significantly. In contrast, any human radiologist can walk down the street to the older hospital and do just fine.

The problem of “artificial stupidity”

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

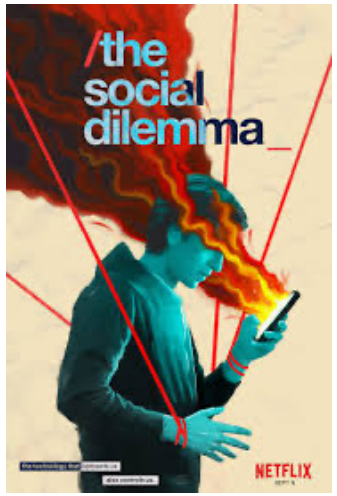
Racial bias skews algorithms widely used to guide care from heart surgery to birth, study finds

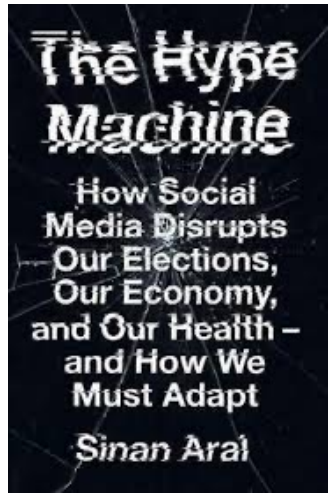
TECHNOLOGY

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

By STEVE LOHR FEB. 9, 2018

Tesla says driver ignored warnings from Autopilot in fatal California crash





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.

→

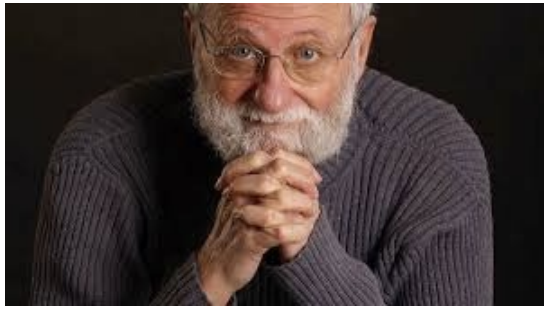
Effective and Ethical AI needs human-centered design

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

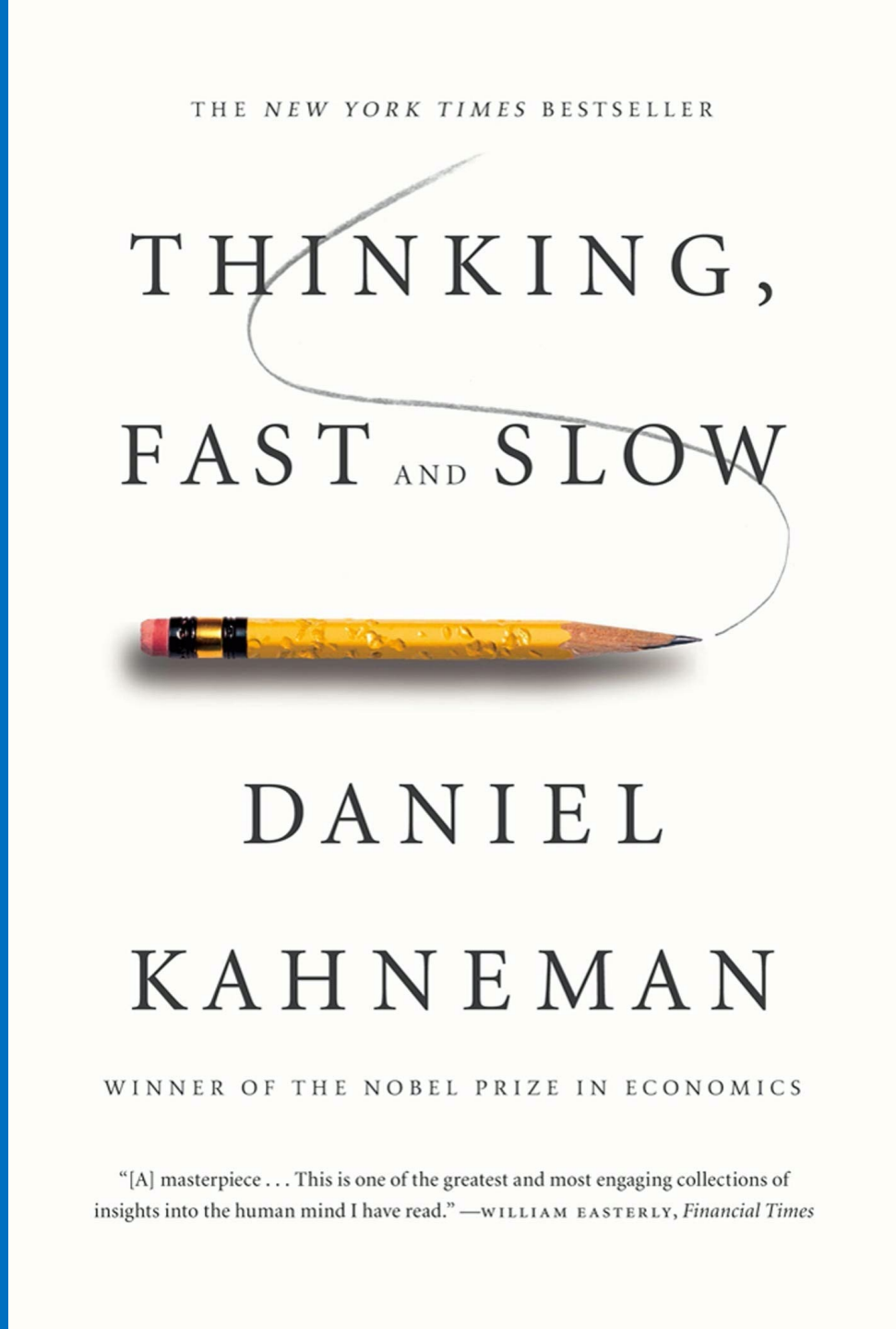


Human-centricity: understanding the user

<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.

Algorithms can “reckon”…

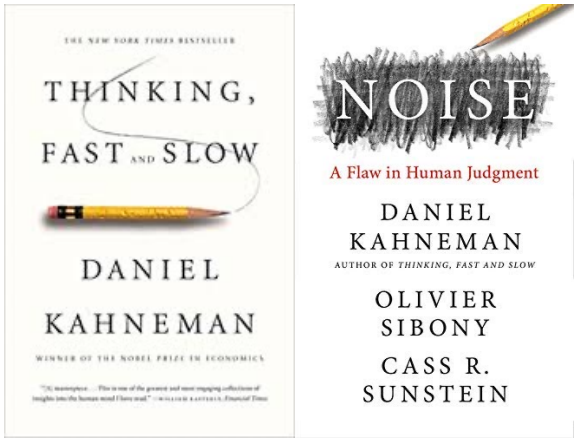
Automatic pilot is an algorithm… We have learned that automatic pilot is more reliable than an individual human pilot.

The same is going to happen here.

— Richard Berk, U. Penn

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

— Sendhil Mullainathan, U. Chicago



… but algorithms cannot "judge"

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…

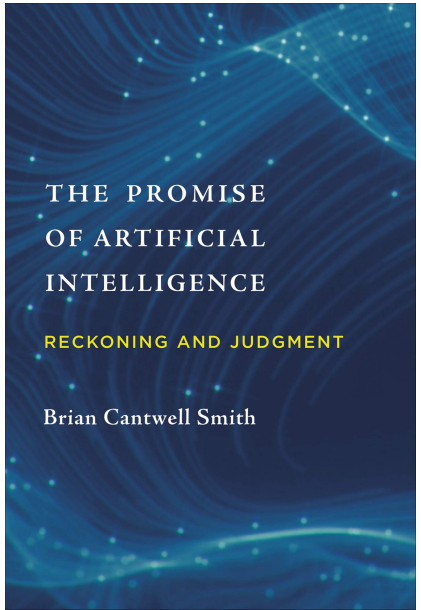
But I can sit here and interact with you, and you can see my expressions and what I am going through.

— Darnell Gates, Philadelphia

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

That is an incredibly high bar.

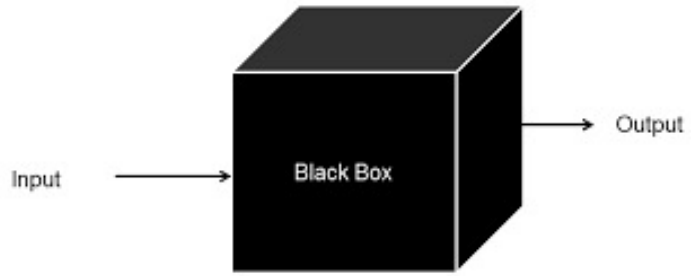
— Brian Cantwell Smith, U. Toronto



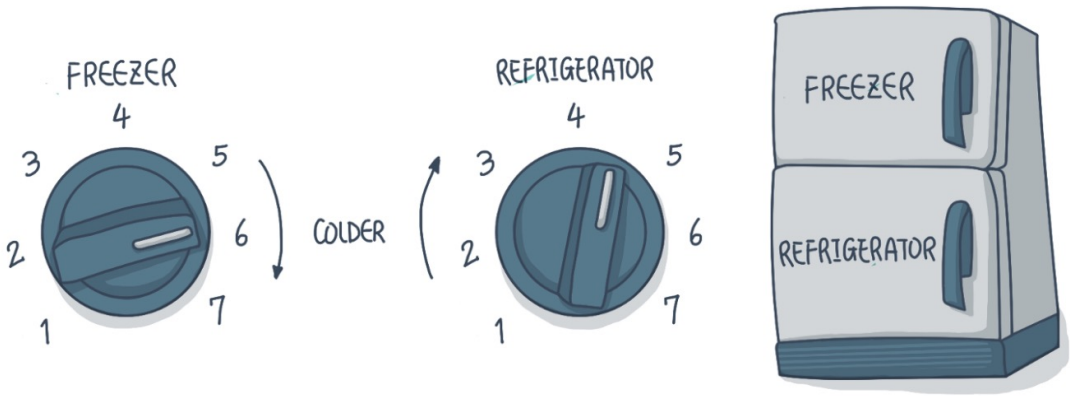
An Algorithm That Grants Freedom, or Takes It Away

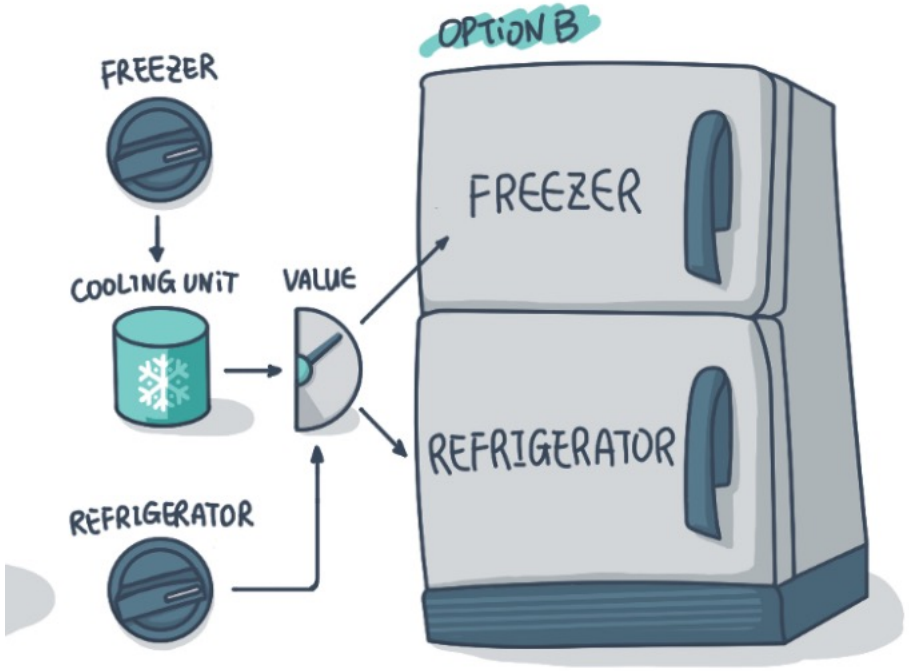
Across the United States and Europe, software is making probation decisions and predicting whether teens will commit crime. Opponents want more human oversight.

They are angered by a growing dependence on automated systems that are taking humans and transparency out of the process. It is often not clear how the systems are making their decisions. Is gender a factor? Age? ZIP code? It's hard to say, since many states and countries have few rules requiring that algorithm-makers disclose their formulas.



Internal behavior of the code is unknown





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

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"

Fundamental design principle:

Begin with the assumption of the need for **human-machine partnerships**.

(Automating tasks should not be the default mode of AI design.)

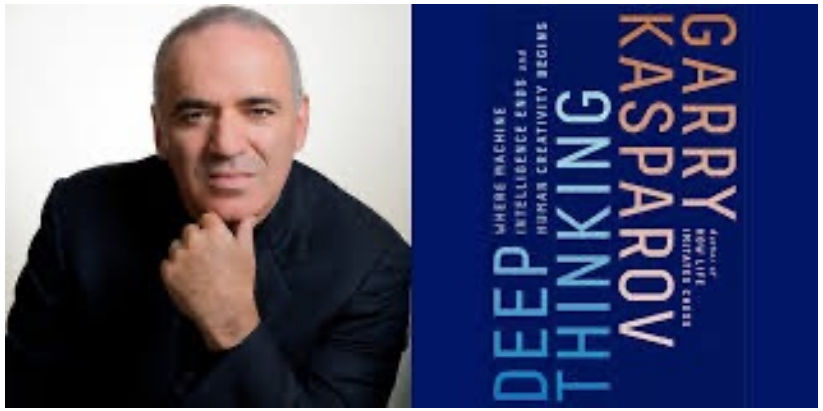
Human-computer symbiosis: a parable

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*

"Greater AI" is about more than optimizing algorithms.

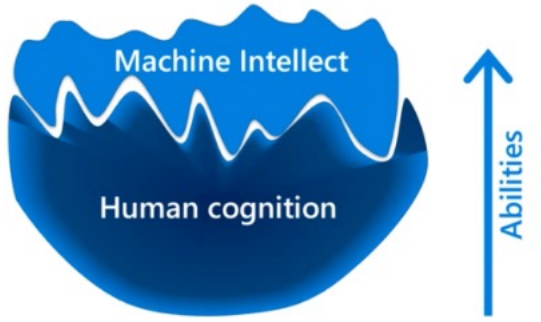
It is about "optimizing" processes of **human-machine collaboration**.

Statistics and computer science provides an incomplete conceptual framework.

Also needed: training/education

And also: 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.

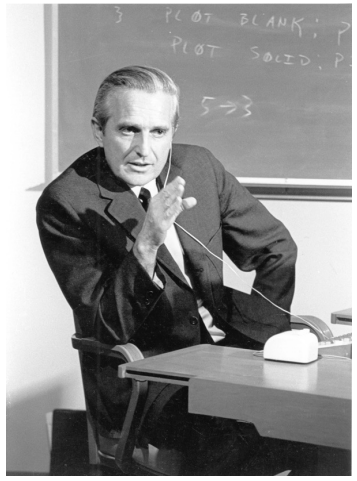
—

J.C.R Licklider

Man-Computer Symbiosis (1960)

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

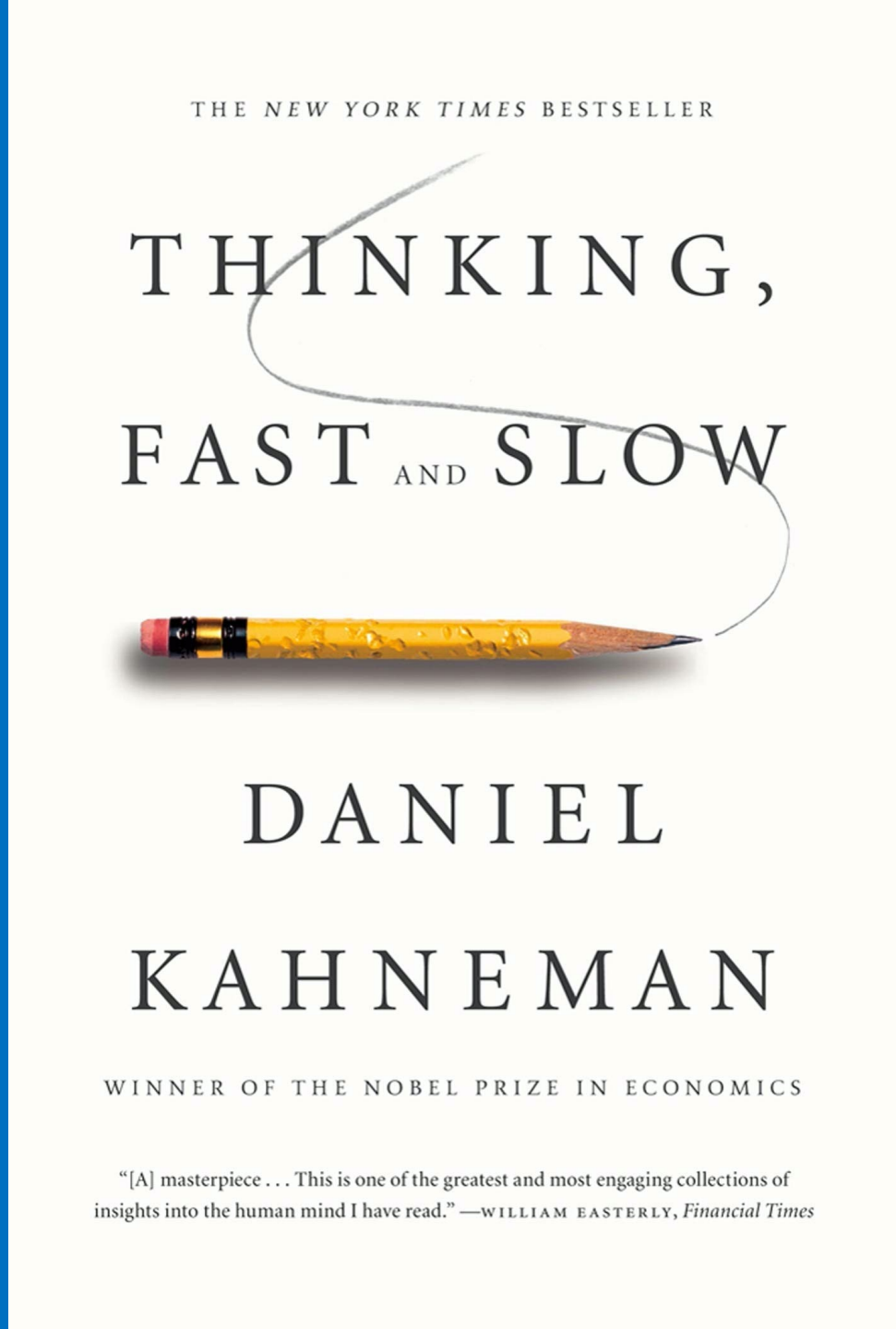
-- Doug Engelbart, 1962



Computers are like a bicycle for our minds.

-- Steve Jobs, 1981

AI and “thinking fast”

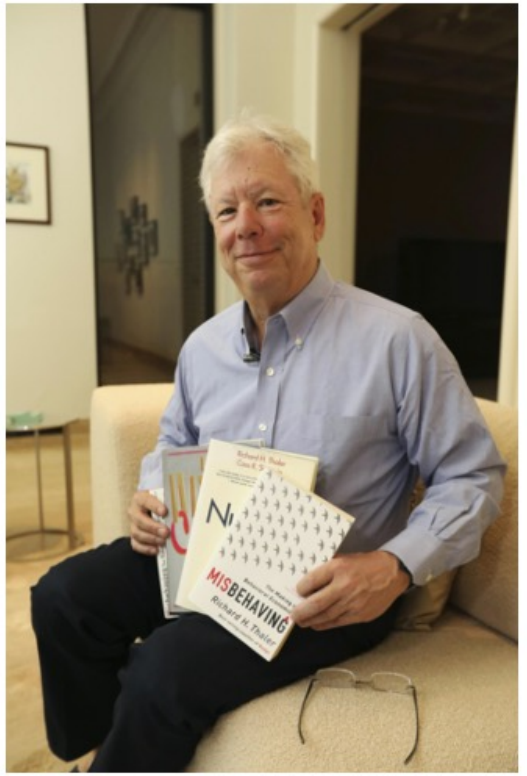


Choice architecture is form of human-centered design

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



Looking for welfare fraud in Rotterdam

Last year in Rotterdam, a rumor circulating in two predominantly low-income and immigrant neighborhoods claimed that the city government had begun using an experimental algorithm to catch citizens who were committing welfare and tax fraud.

The algorithm produces “risk reports” on individuals who should be questioned by investigators. In Rotterdam, where the system was most recently used, 1,263 risk reports were produced in two neighborhoods.

“You’re putting me in a system that I didn’t even know existed,” said Mr. Bouchkhachakhe, who works for a logistics company.

Ethics and the need for “greater AI”

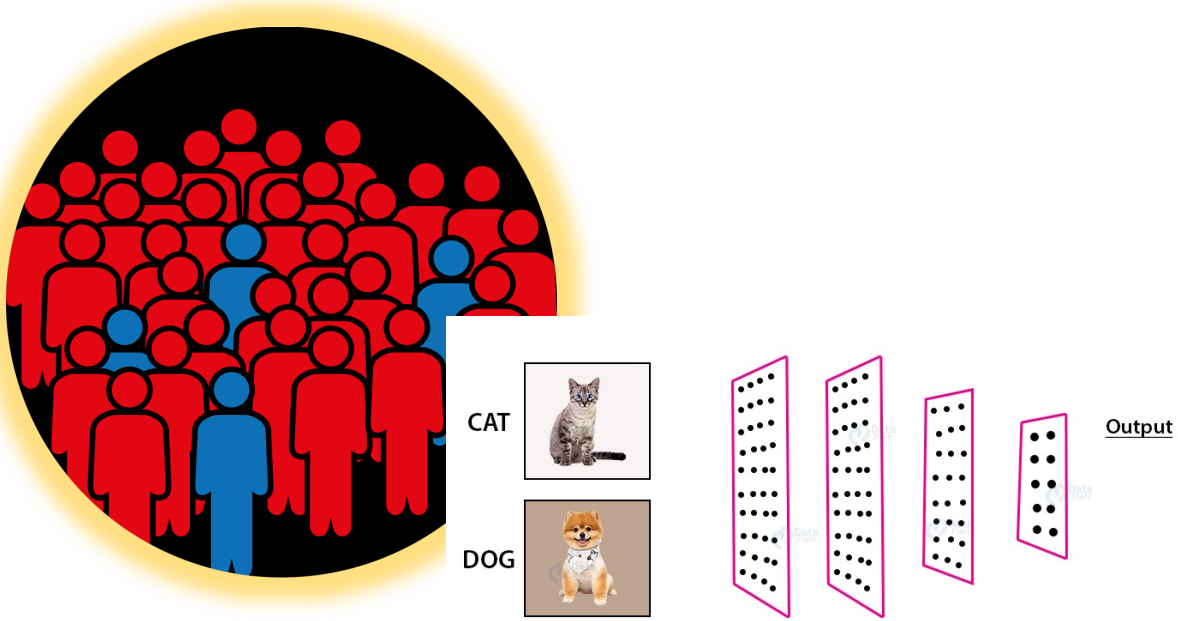
PEW

Behavioral Analytics Help Save Unemployment Insurance Funds

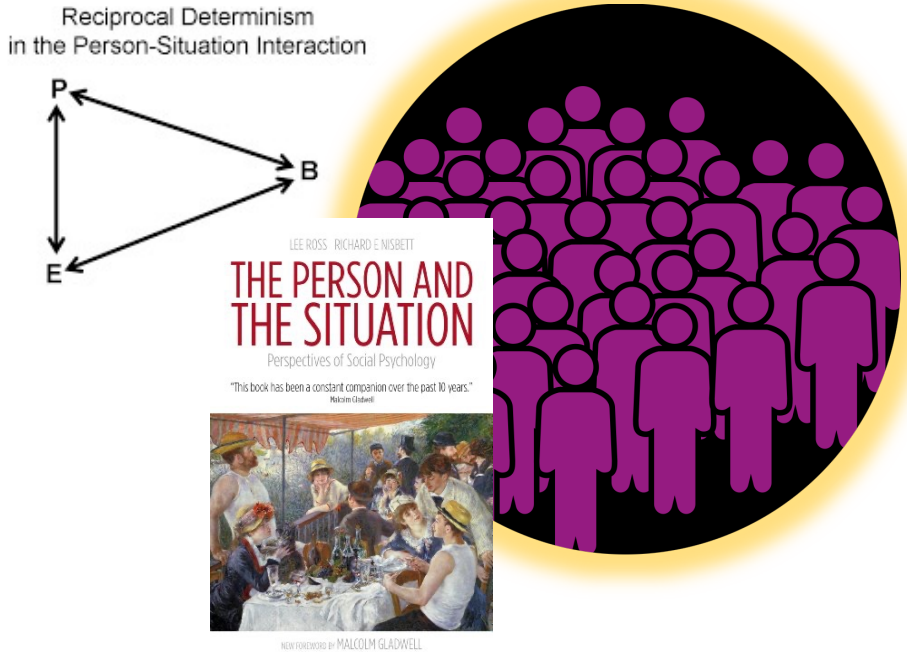
New Mexico uses data to identify misinformation, save money

ISSUE BRIEF October 26, 2016

Naïve view



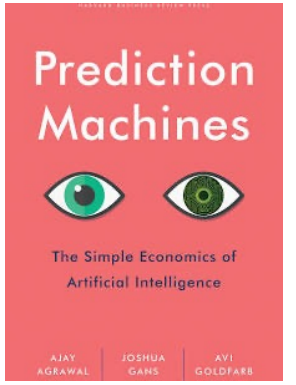
“Nudge” view



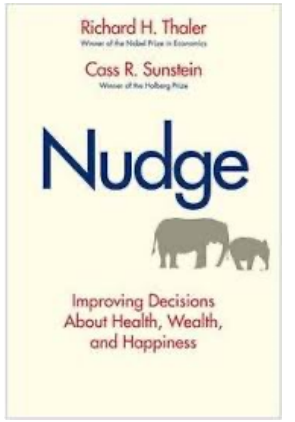
Interventions, not just predictions

Some implications:

- AI is often framed in terms of machines that make better predictions.
  We must also focus on interventions.



- Possible interventions are often construed in "classical economics" terms:
  Presenting rational actors with information, rewards, or punishments.

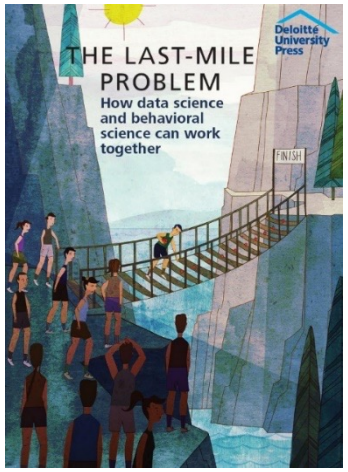


- Choice architecture – human centered design of choice environments
  – can be part of the applied AI toolkit.
  - Ethical deliberation is crucial here: "behavioral AI" must be human autonomy-enhancing.

- More generally: Principles of ethics and psychology belong in the design phase of applied AI projects
  No less than principles of data collection, statistical analysis, and machine learning

Further reading

“The Last Mile Problem: How data science and behavioral science can work together”
Deloitte Review, January 2015
http://dupress.com/articles/behavioral-economics-predictive-analytics/



“The Importance of Misbehaving: A conversation with Richard Thaler”
Deloitte Review, January 2016
https://dupress.deloitte.com/dup-us-en/deloitte-review/issue-18/behavioral-economics-richard-thaler-interview.html

“Cognitive collaboration: Why humans and computers think better together”
Deloitte Review, January 2017
https://dupress.deloitte.com/dup-us-en/deloitte-review/issue-20/augmented-intelligence-human-computer-collaboration.html

“Smarter together: Why artificial intelligence needs human-centered design”
Deloitte Review, January 2018
https://www2.deloitte.com/insights/us/en/deloitte-review/issue-22/artificial-intelligence-human-centric-design.html

“Superminds: How humans and machines can work together”
(Interview with Thomas Malone, MIT Sloan School of Management
Deloitte Review, January 2019
https://www2.deloitte.com/insights/us/en/focus/technology-and-the-future-of-work/human-and-machine-collaboration.html

“Human values in the loop: Design principles for ethical AI”
Deloitte Review, January 2020
https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/design-principles-ethical-artificial-intelligence.html

“Superminds, not Substitutes: Designing human-machine collaboration for a better future of work”
Deloitte Review, August 2020
https://www2.deloitte.com/us/en/insights/focus/technology-and-the-future-of-work/ai-in-the-workplace.html