Rethinking AI

Daron Acemoglu

MIT

Promise of AI

- Tremendous advances in AI over the last decade, and especially in the last few years, with transformers, large language models and more broadly generative AI.

- Recent issue of The Economist magazine:

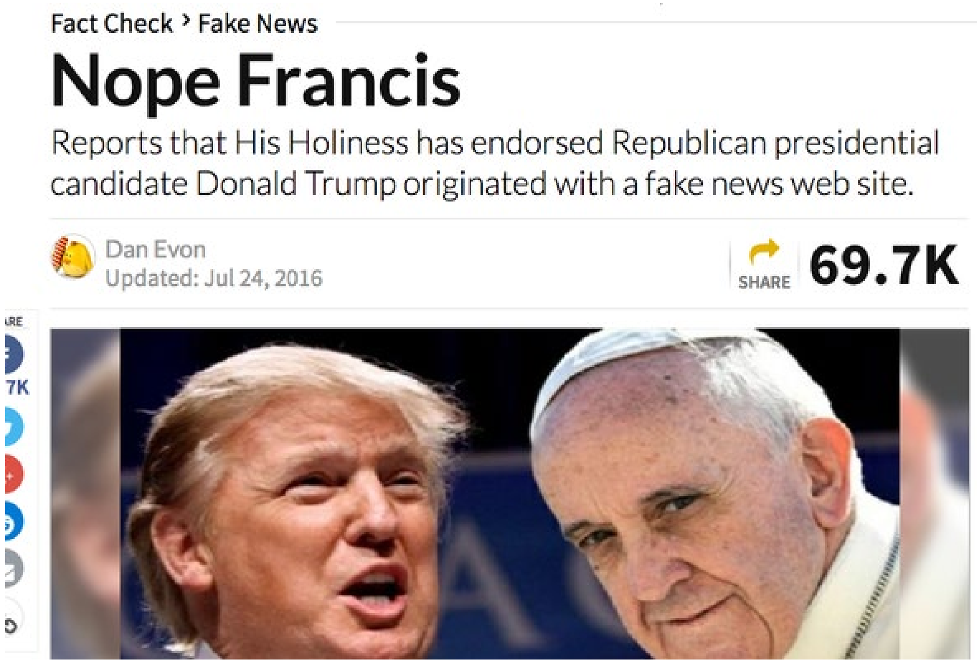
> The field's progress is precipitate and its promise immense.… The fear that machines will steal jobs is centuries old. But so far new technology has created new jobs to replace the ones it has destroyed. … Imposing heavy regulation, or indeed a pause, today seems an over-reaction.

- AI can in principle: Help fight pandemics, diagnose cancers, develop new drugs, expand and improve communication, transform entertainment, and even design new materials to help fight climate change.

What We Have Seen So Far

Yet, lots of causes for concern, from prior experience.

- Misinformation on the Internet and social media
- Emotional problems from excessive social media dependence
- New and more complex biases, e.g., pertaining to race
- Surveillance and massive data collection.
- An emerging monopoly/oligopoly over information and news.
- And jobs and inequality…



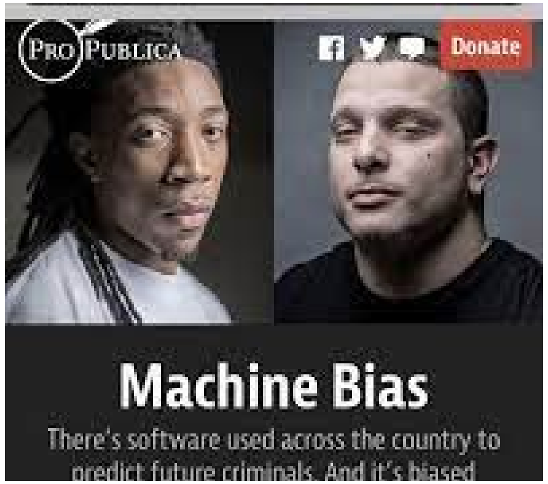
Fact Check › Fake News

Nope Francis

Reports that His Holiness has endorsed Republican presidential candidate Donald Trump originated with a fake news web site.

Dan Evon
Updated: Jul 24, 2016

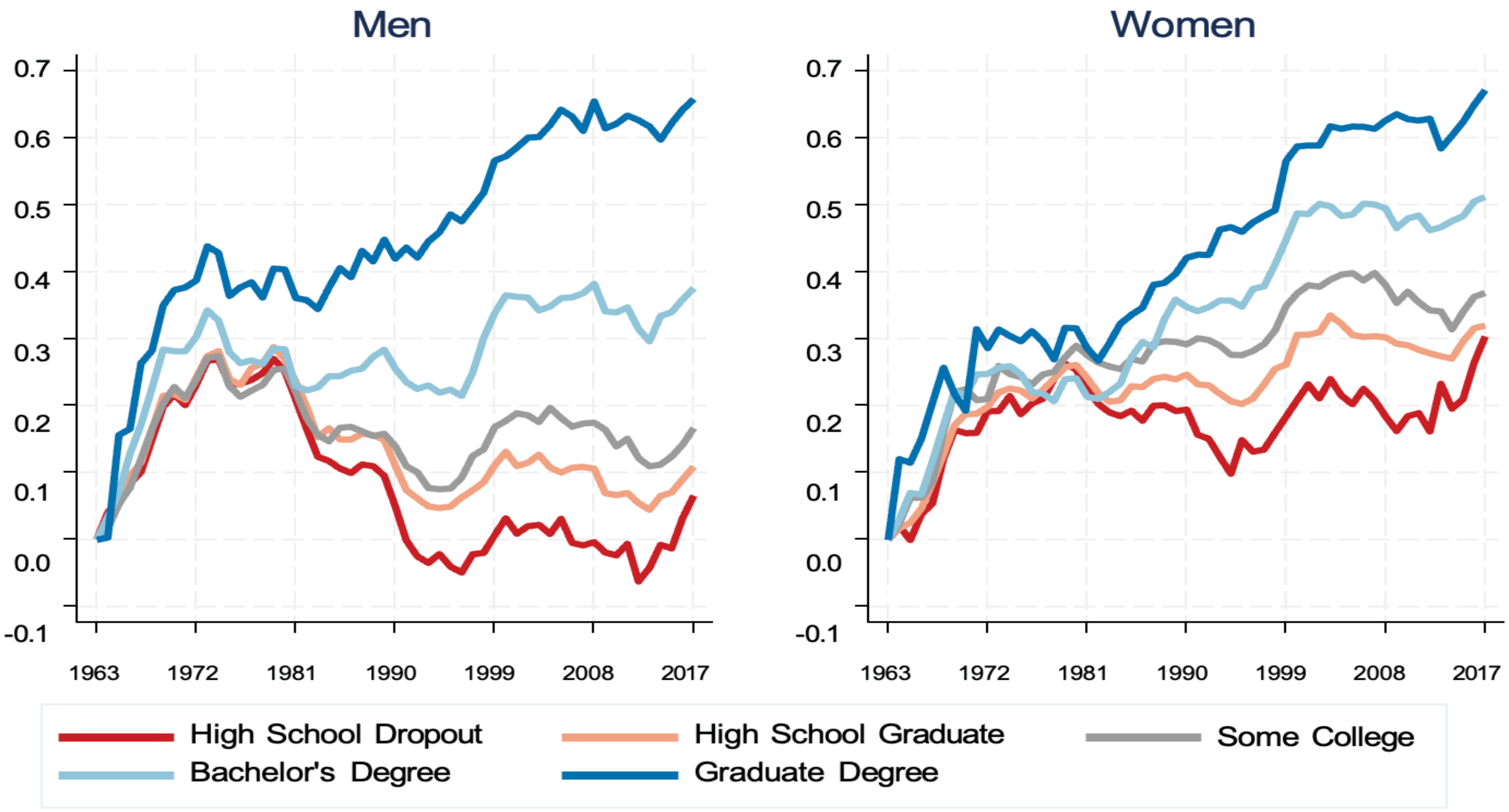
SHARE 69.7K



Machine Bias

There's software used across the country to predict future criminals. And it's biased

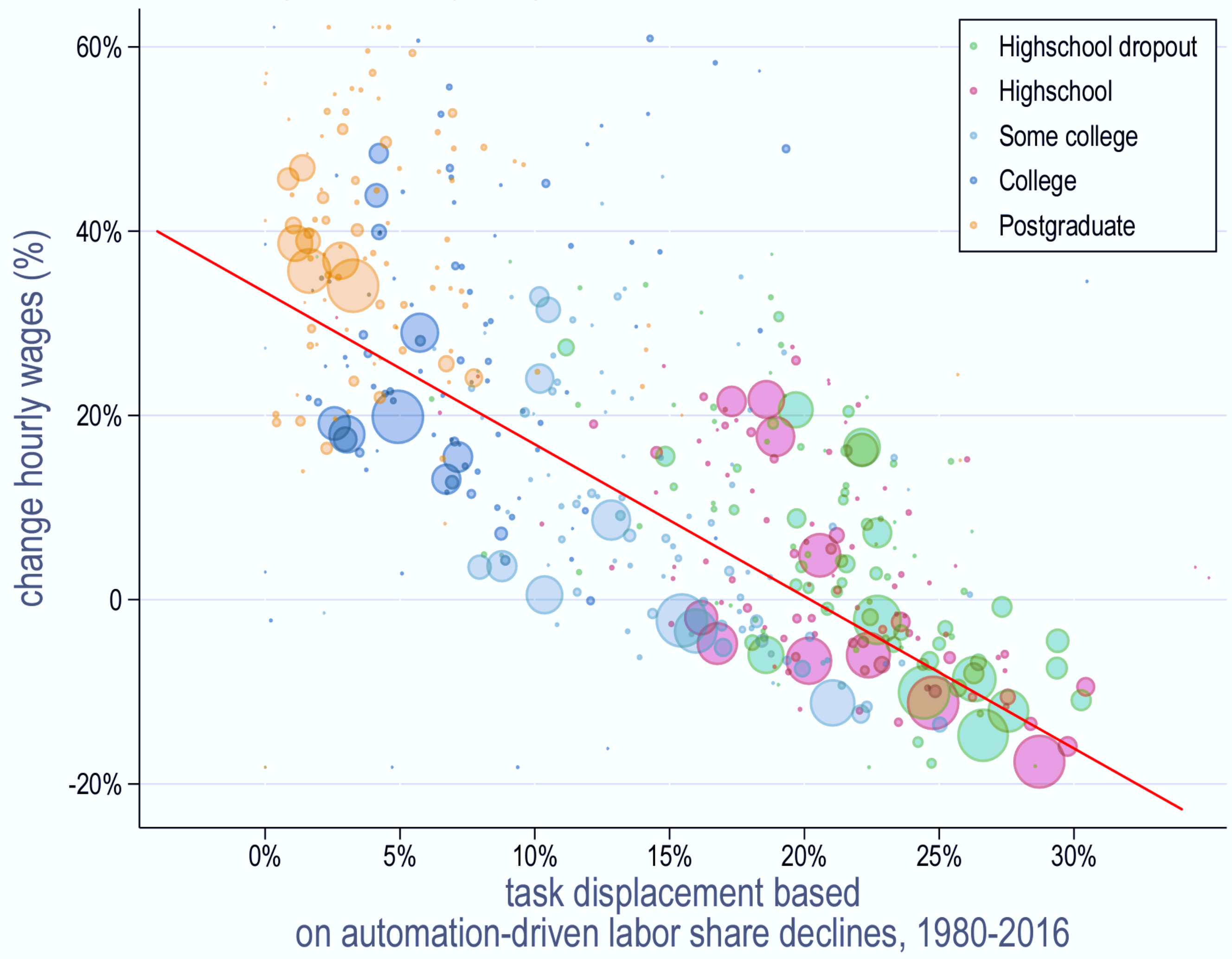
What We Have Seen So Far



Breakdown of shared prosperity during the era of digital technologies

What We Have Seen So Far

- Surge in inequality related to how we use digital technologies, in particular automation.
- As more and more tasks have been automated using numerically controlled machinery, office software and robotics, demographic groups that used to specialize in these jobs have experienced job loss and real income declines.
- The problem is not that digital technologies have automated work.
- The problem is that they have not created new tasks for the workers that have been displaced from their jobs.
- Early evidence that AI is going in the same automation direction.



Do we have to put up with these hugely costly disruptions? Or is there another way?
Can we make AI work better for society and for businesses?

Better AI (and LLM)

Two fundamentally different visions of AI:

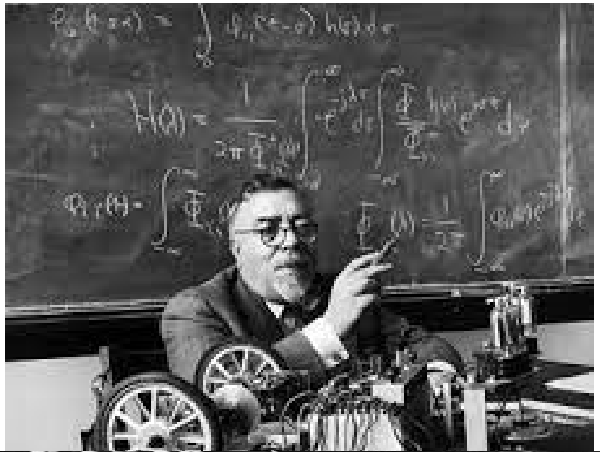
1. Machines designed to be *smarter and more powerful* than (most) humans.
2. Machines to *complement* human abilities.

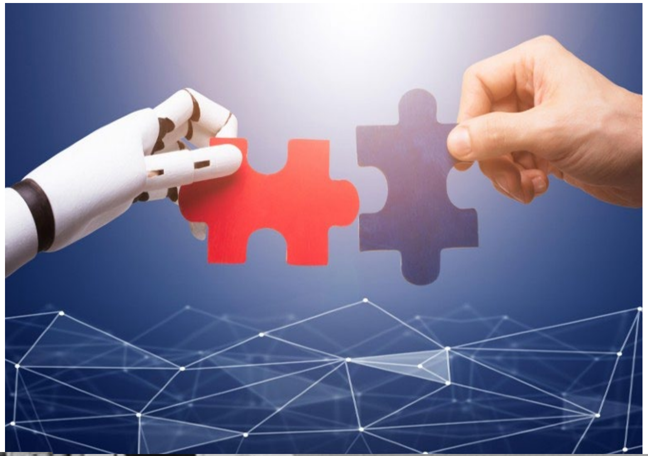
- The second vision – let's call it machine usefulness or "pro-human AI"– starts with Norbert Wiener (but also Edgar Allan Poe).

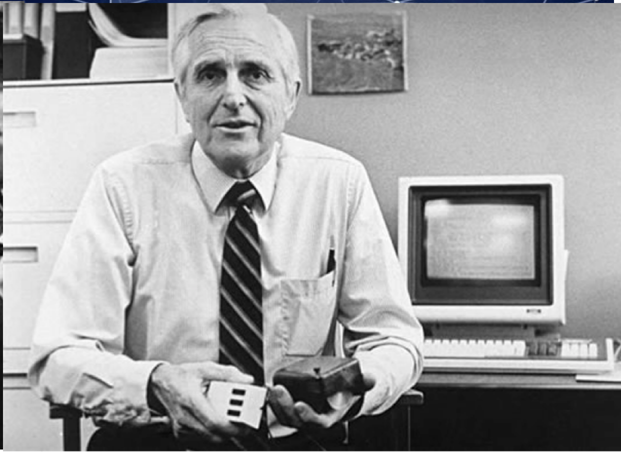
- Translated into practice by computer scientists such as JCR Licklider and Douglas Engelbart.

- Licklider: "human-machine symbiosis" as a way of complementing human capabilities; Engelbart suggested "augmenting human intellect".

- This requires humans to understand and appropriately use new technologies.

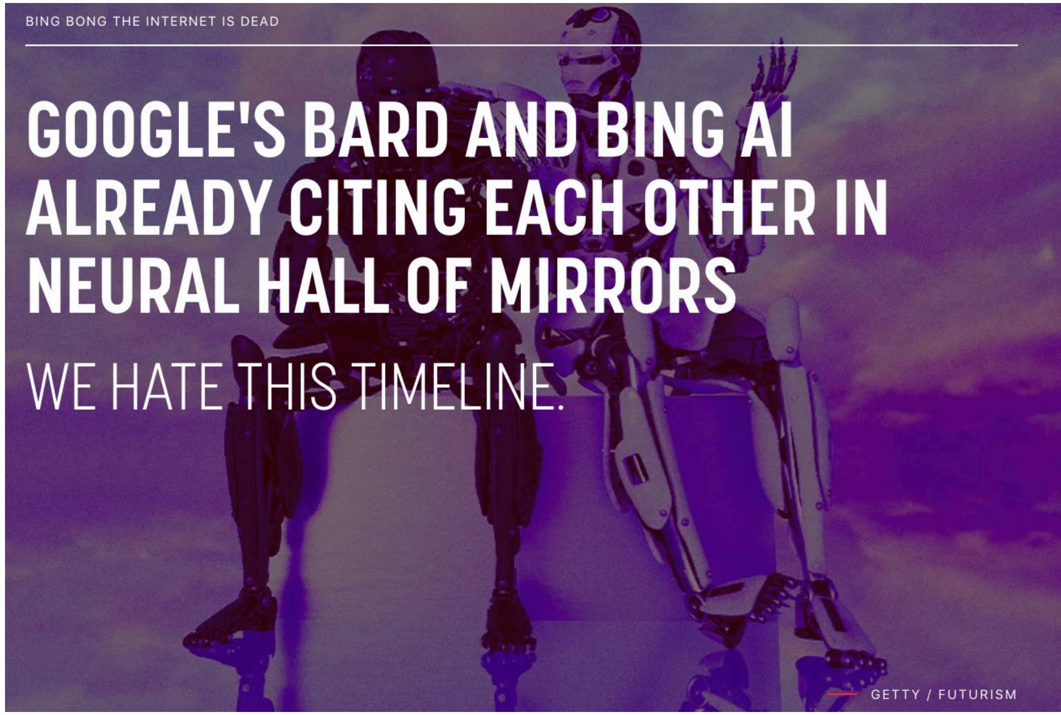






Roadblocks on Better LLM

- LLMs remain **illegible** and give “excessively authoritative” advice, rather than communicate with humans. Danger of more costly feedback loops.
- These will be magnified as LLMs start populating the Internet.
- Significant fraction of the Internet filled with LLM-produced content, LLMs will cite each other, and much of social media content will be influenced by LLMs.



BING BONG THE INTERNET IS DEAD

GOOGLE'S BARD AND BING AI ALREADY CITING EACH OTHER IN NEURAL HALL OF MIRRORS

WE HATE THIS TIMELINE.

GETTY / FUTURISM

Example:

Human Query: “Is policy X effective?”;

LLM: “No”.

Future Human Communication on social media: Policy X is not working.

LLM new training data: Policy X is not working, since much evidence on the web that it is doing so.

How to Do AI Better with LLMs?

- Better business models.
- Better architecture -- especially **"legible AI"** – to keep humans in the driving seat.

Some elements of such an architecture may be:

- incorporate reliability scores together with accurate source information
- allow reasoning exchanges, sensitivity analysis and broader interrogation by humans
- structure that facilitates human-complementary applications
- more selective use of (higher-quality) data
- internal guardrails, perhaps with two LLMs consistently checking each other (to prevent venturing into excessive authoritativeness) and internal structure to facilitate regulation

- There is the possibility of a **positive loop**: better architecture of LLMs enable businesses to use them more productively and encourage more useful technologies to develop.

Good news: It is possible.

Bad news: There is not where we are heading, for a variety of reasons (industry organization, the power of big tech companies and dominant visions in the field).