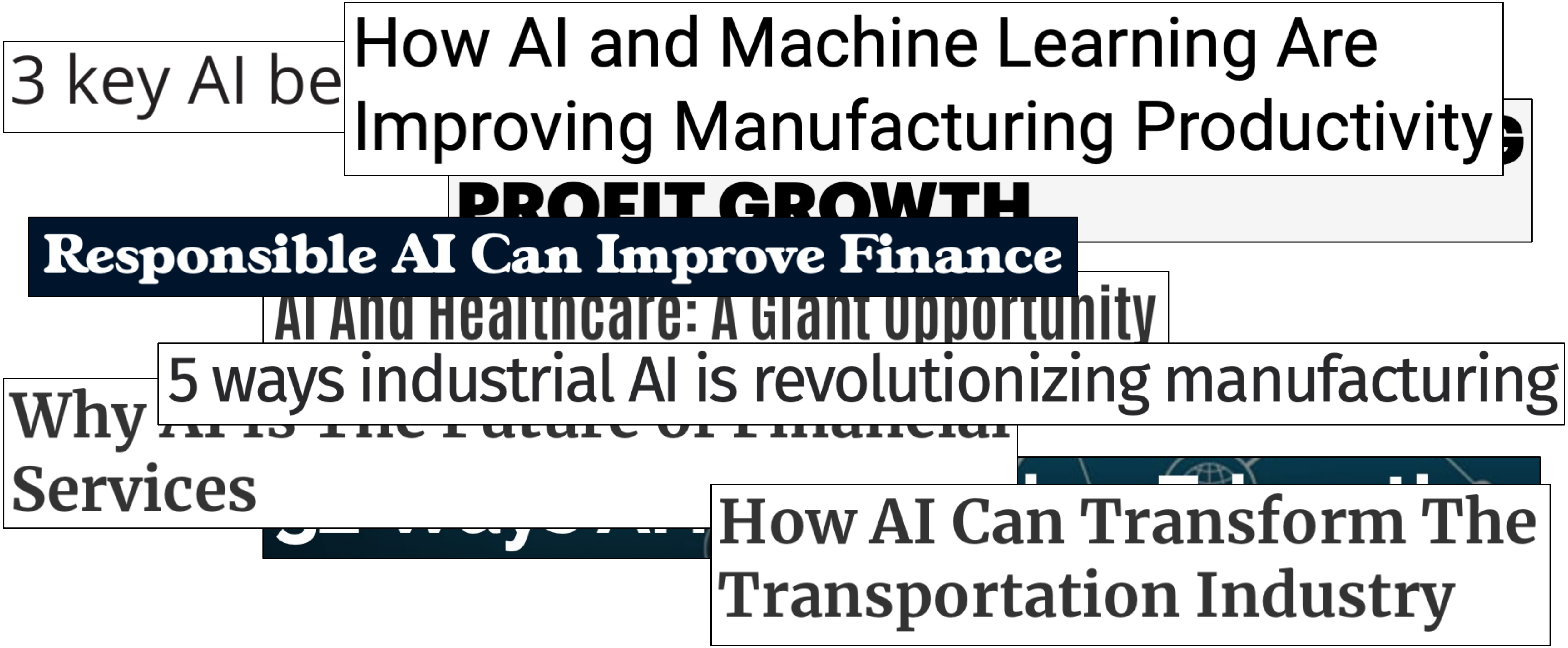
AI ETHICS AS TRANSLATIONAL ETHICS

DAVID DANKS

DATA SCIENCE // PHILOSOPHY
UNIVERSITY OF CALIFORNIA, SAN DIEGO

TRANSFORMATIVE AI THAT HELPS



3 key AI be

How AI and Machine Learning Are Improving Manufacturing Productivity

PROFIT GROWTH

Responsible AI Can Improve Finance

AI And Healthcare: A Giant Opportunity

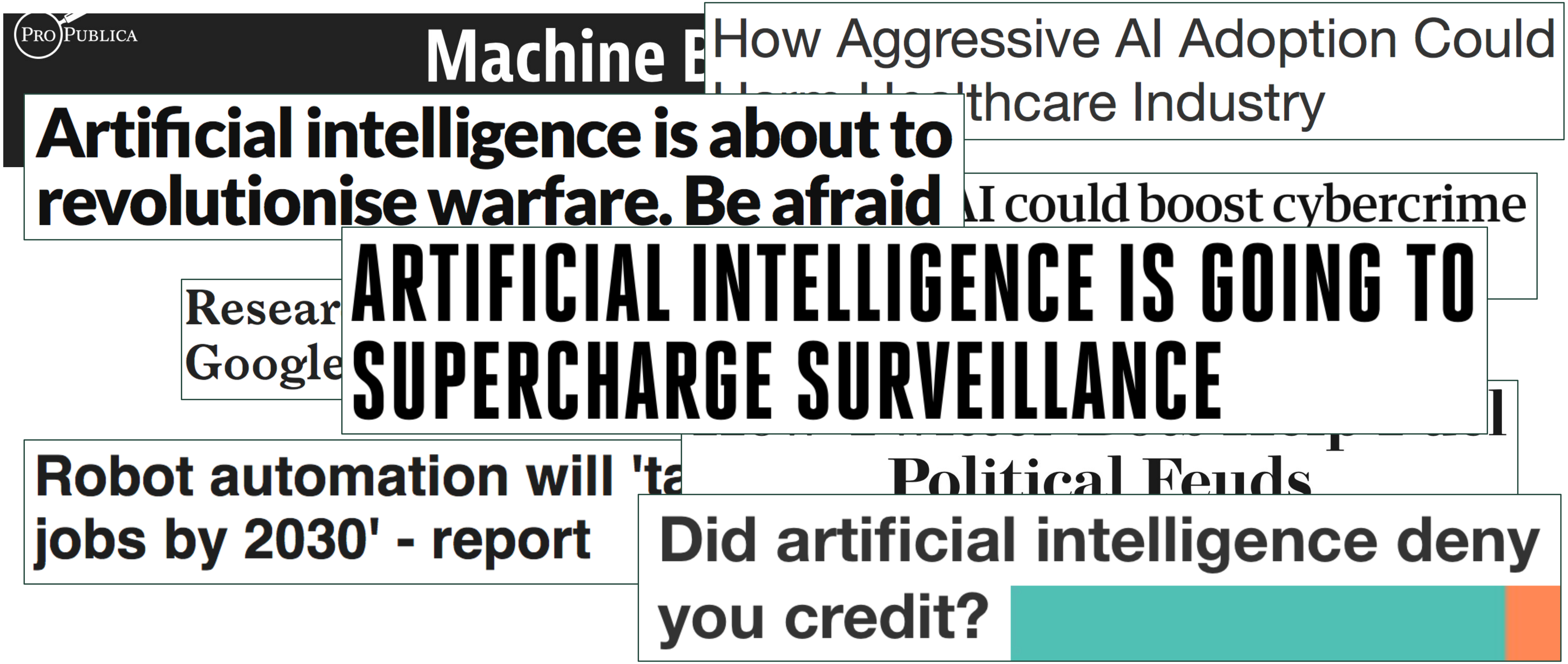
5 ways industrial AI is revolutionizing manufacturing

Why AI Is The Future of Financial Services

32 Ways AI

How AI Can Transform The Transportation Industry

TRANSFORMATIVE AI THAT HARMS



PRO PUBLICA

Machine B

How Aggressive AI Adoption Could Harm Healthcare Industry

Artificial intelligence is about to revolutionise warfare. Be afraid

AI could boost cybercrime

Resear
Google

ARTIFICIAL INTELLIGENCE IS GOING TO SUPERCHARGE SURVEILLANCE

Political Feuds

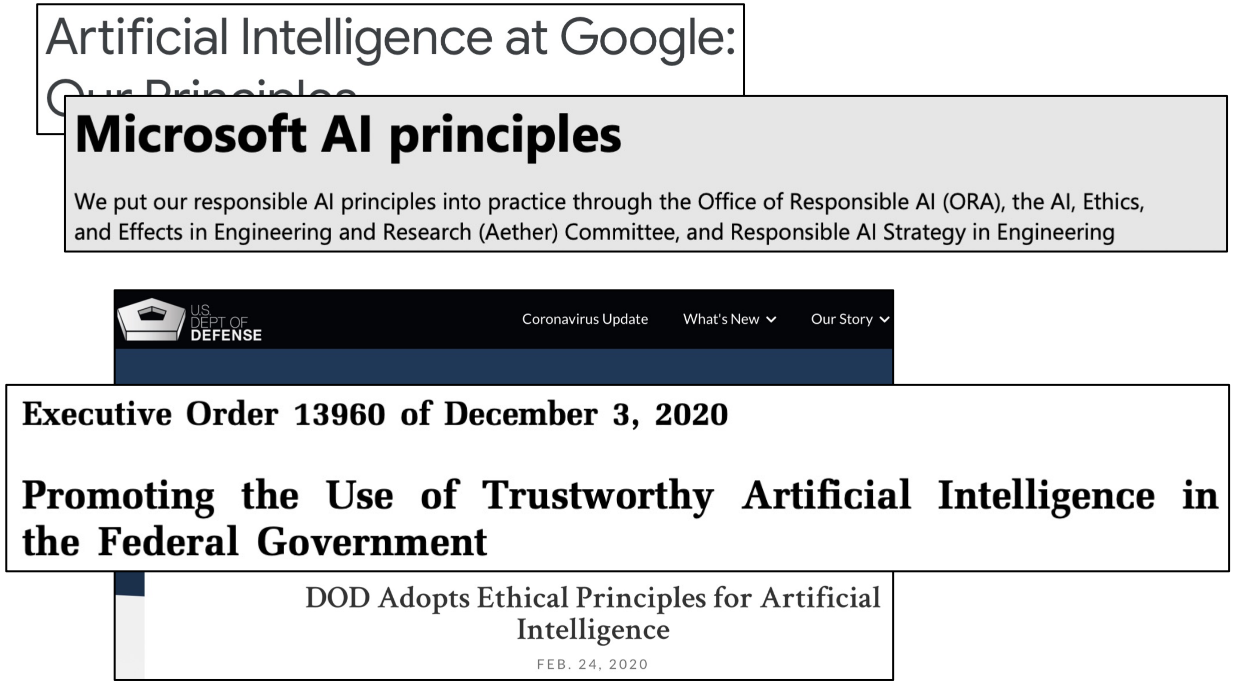
Robot automation will 'ta
jobs by 2030' - report

Did artificial intelligence deny you credit?

APPROACHES TO ETHICAL AI

1. Ethical AI principles & features
2. Ethical algorithms
3. Ethical system behaviors

ETHICAL PRINCIPLES & FEATURES



Artificial Intelligence at Google: Our Principles

Microsoft AI principles

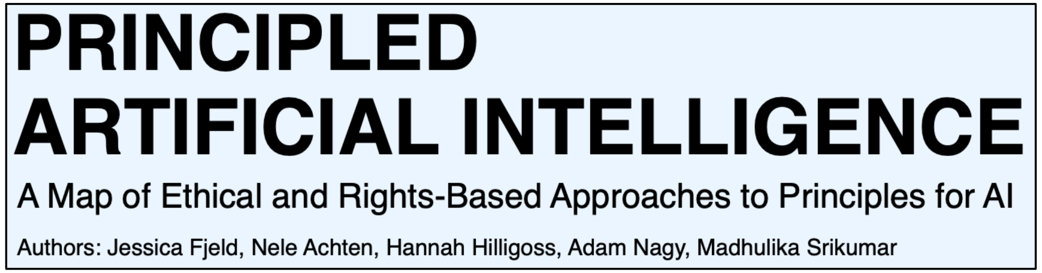
We put our responsible AI principles into practice through the Office of Responsible AI (ORA), the AI, Ethics, and Effects in Engineering and Research (Aether) Committee, and Responsible AI Strategy in Engineering

Executive Order 13960 of December 3, 2020

Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government

DOD Adopts Ethical Principles for Artificial Intelligence

FEB. 24, 2020



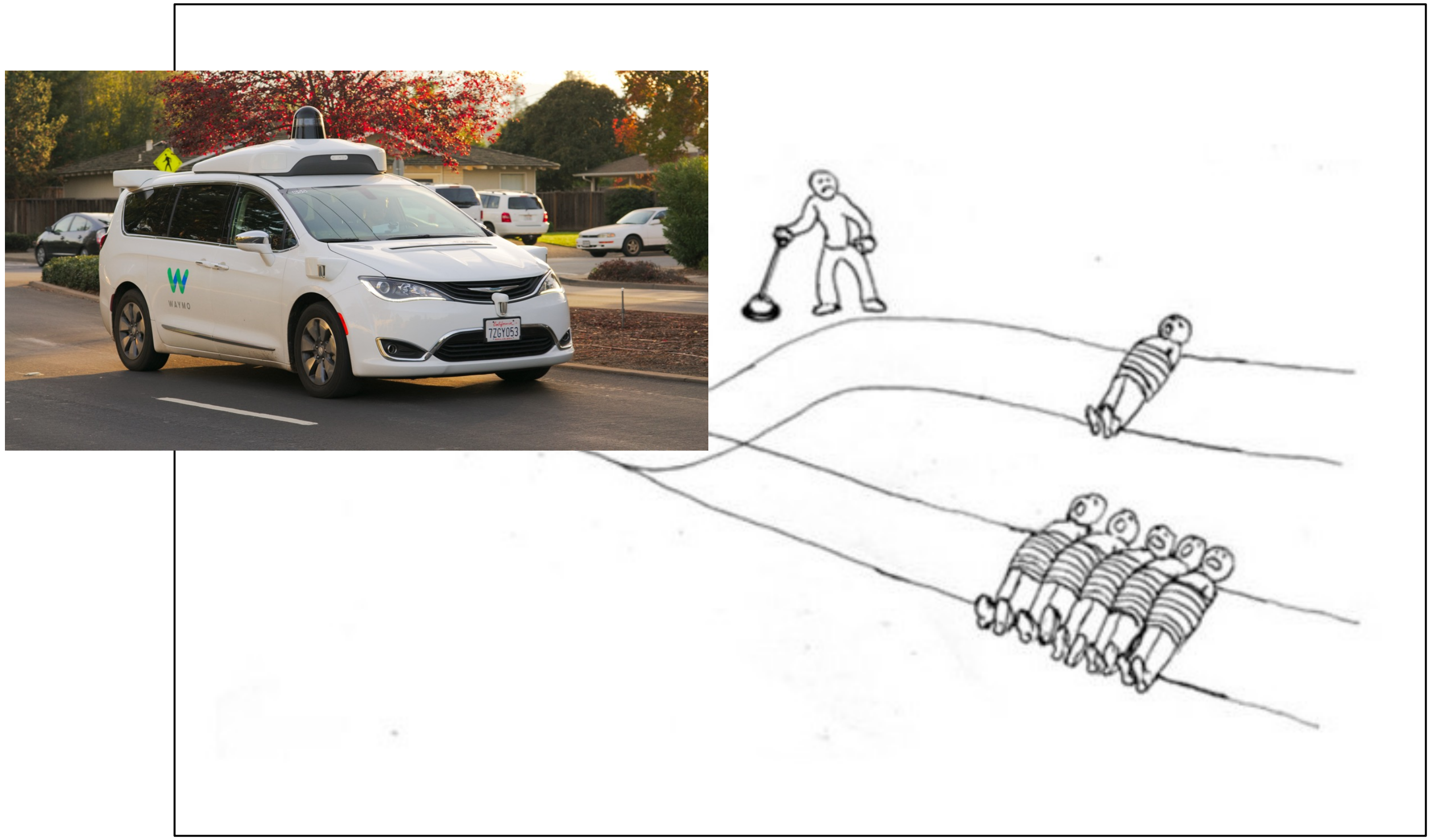
PRINCIPLED ARTIFICIAL INTELLIGENCE

A Map of Ethical and Rights-Based Approaches to Principles for AI

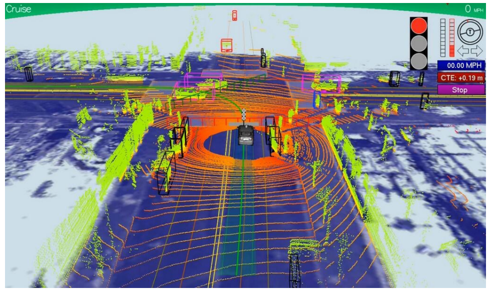
Authors: Jessica Fjeld, Nele Achten, Hannah Hilligoss, Adam Nagy, Madhulika Srikumar

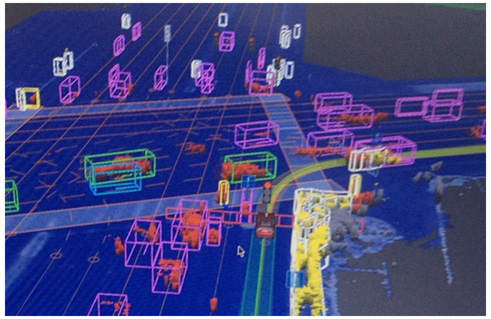
- Limited impact on practice
- Context-insensitive
- Team-sensitive
- Lack of interoperability

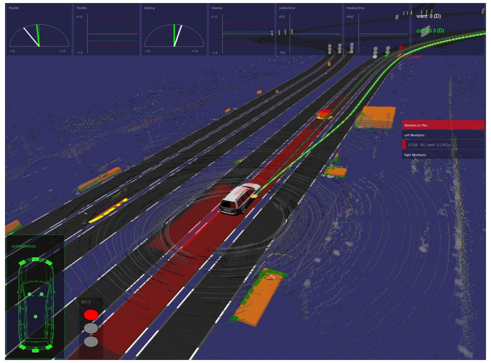
ETHICAL ALGORITHMS



ETHICAL ALGORITHMS







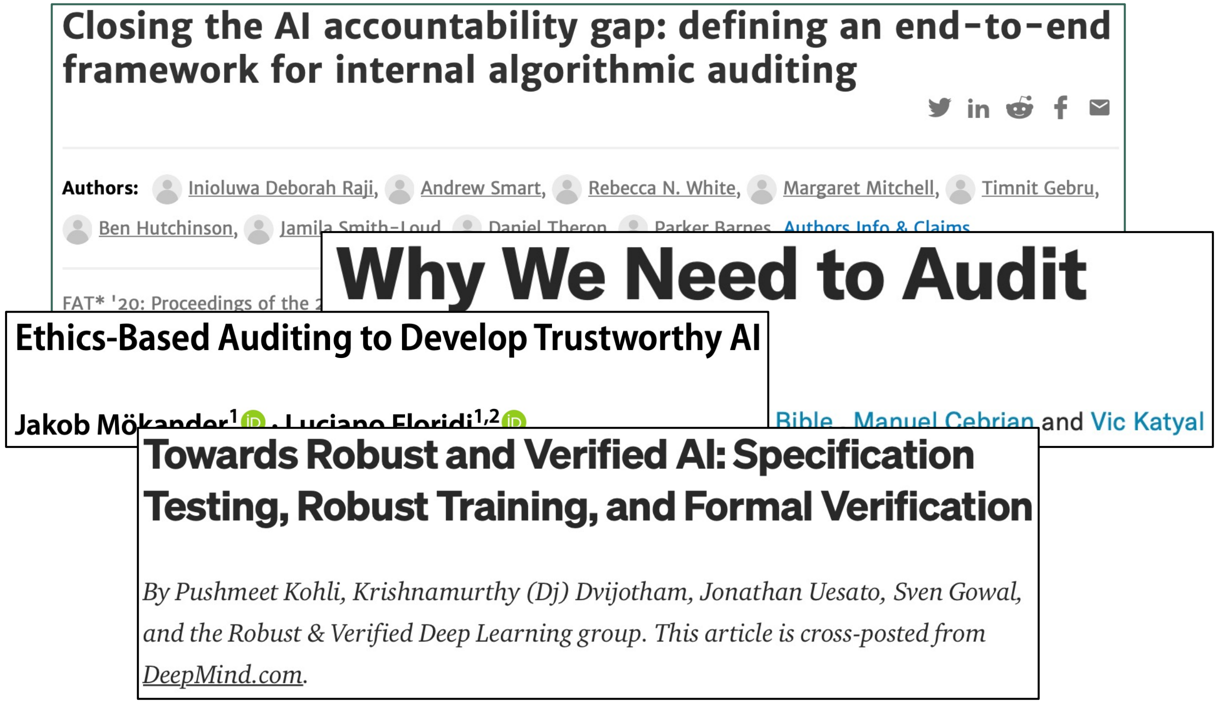
No explicit Trolley Problem calculus…

…“just” finding a low-cost path through the landscape

Carnegie Mellon University
Carnegie Mellon University
Carnegie Mellon University
Carnegie Mellon University

CoBots
CORAL research group
Manuela Veloso (CMU)

ETHICAL BEHAVIORS



Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing

Authors: Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, Parker Barnes, Authors Info & Claims

FAT* '20: Proceedings of the 2

Why We Need to Audit

Bible, Manuel Cebrian and Vic Katyal

Ethics-Based Auditing to Develop Trustworthy AI

Jakob Mökander[1], Luciano Floridi[1,2]

Towards Robust and Verified AI: Specification Testing, Robust Training, and Formal Verification

By Pushmeet Kohli, Krishnamurthy (Dj) Dvijotham, Jonathan Uesato, Sven Gowal, and the Robust & Verified Deep Learning group. This article is cross-posted from DeepMind.com.

- Require implausible specificity
- Only work for "closed worlds"
- Insensitive to values of different groups
- Risks of Goodhart's Law

A POTENTIAL DIAGNOSIS?

- Each focused on “basic research” on one component
- ⇒ Each ignores key complexities
  - Principles ignore practice
  - Algorithms ignore technology
  - Behaviors ignore ethics

- But if not basic research, then what?

TRANSLATIONAL ETHICS

- *Translational medicine*: Substantive research to apply basic biomedical advances into clinical practice

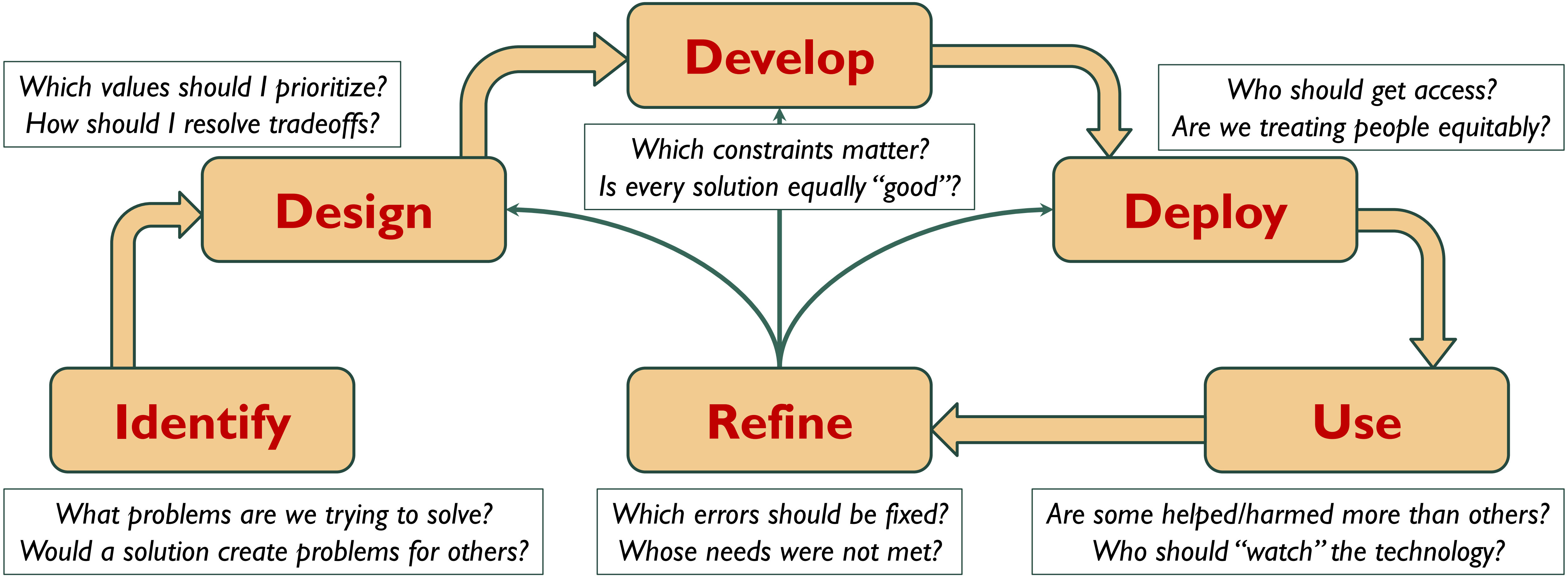
- Generally requires *both*: (cf. Baddeley, 1978)
  - Applied basic research
  - Basic applied research

TRANSLATIONAL ETHICS

- *Translational* ***AI ethics***: Substantive research to apply basic **ethical** advances into **technological** practice

- Generally requires *both*: (cf. Baddeley, 1978)
  - Applied basic research: **Translation of AI, HCI, ethics, sociology, …**
  - Basic applied research: **Novel practices, processes, methods, …**

INTERVENTION POINTS



What problems are we trying to solve?
Would a solution create problems for others?

Which values should I prioritize?
How should I resolve tradeoffs?

Which constraints matter?
Is every solution equally "good"?

Who should get access?
Are we treating people equitably?

Are some helped/harmed more than others?
Who should "watch" the technology?

Which errors should be fixed?
Whose needs were not met?

TOWARDS TRANSLATIONAL AI ETHICS

- **People**: Interdisciplinary education; Collaboration training; …
- **Processes**: Datasheets; Model cards; Ethical triage; Audits (both pre- and post-deployment); …
- **Policies**: Smarter regulation; Better industry standards; Improved incentives; …
- **Partnerships**: Value ↔ Code mappings; Ethical interoperability; …

All *start* w/ practice, technology, & ethics

All multi-disciplinary

THANKS!

www.daviddanks.org
ddanks@ucsd.edu // david@danks.org

Key conversationalists:

- Dwight Barry
- Sina Fazelpour
- Emily LaRosa
- Zack Lipton
- Alex John London
- Osonde Osoba
- Alka Patel
- Heather Roff
- Kerstin Vignard

(and many others)