Responsible Data Science

Applied Ethics

April 21, 2025

Lucas Rosenblatt

Center for Data Science &
Computer Science and Engineering
New York University

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai

AI ethics from our finest news source!

the ONION

4/22/19 7:02am • SEE MORE: SCIENCE

NEWS IN BRIEF

Computer Scientists Say AI's Underdeveloped Ethics Have Yet To Move Beyond Libertarian Phase

"While companies like Facebook and Google have allocated millions to making sure machine learning is guided by basic moral and ethical values, early prototypes, which achieved self-awareness, have yet to move beyond self-importance," said MIT robotics research engineer Dr. Alvin Dubicki.

Dubicki hypothesized that **even the most advanced labs are decades away from developing neural networks sophisticated enough to analyze large quantities of data and output much else besides paraphrased Ayn Rand quotes**.

costs, benefits & risks

Costs, benefits, and risks

- Barebones DS pipeline:

Aims → Task → Data

Crude cost-benefit analysis

$$\text{Benefit}_{\text{aims}} > \text{Cost}_{\text{task}} + \text{Cost}_{\text{data}} \rightsquigarrow \text{Do it}$$

Costs, benefits, and risks

- *Aim*: YouTube wants to optimize views
- *Task*: Experiment with recommendation engine
- *Data*: User profiles, page views, time spent watching, etc

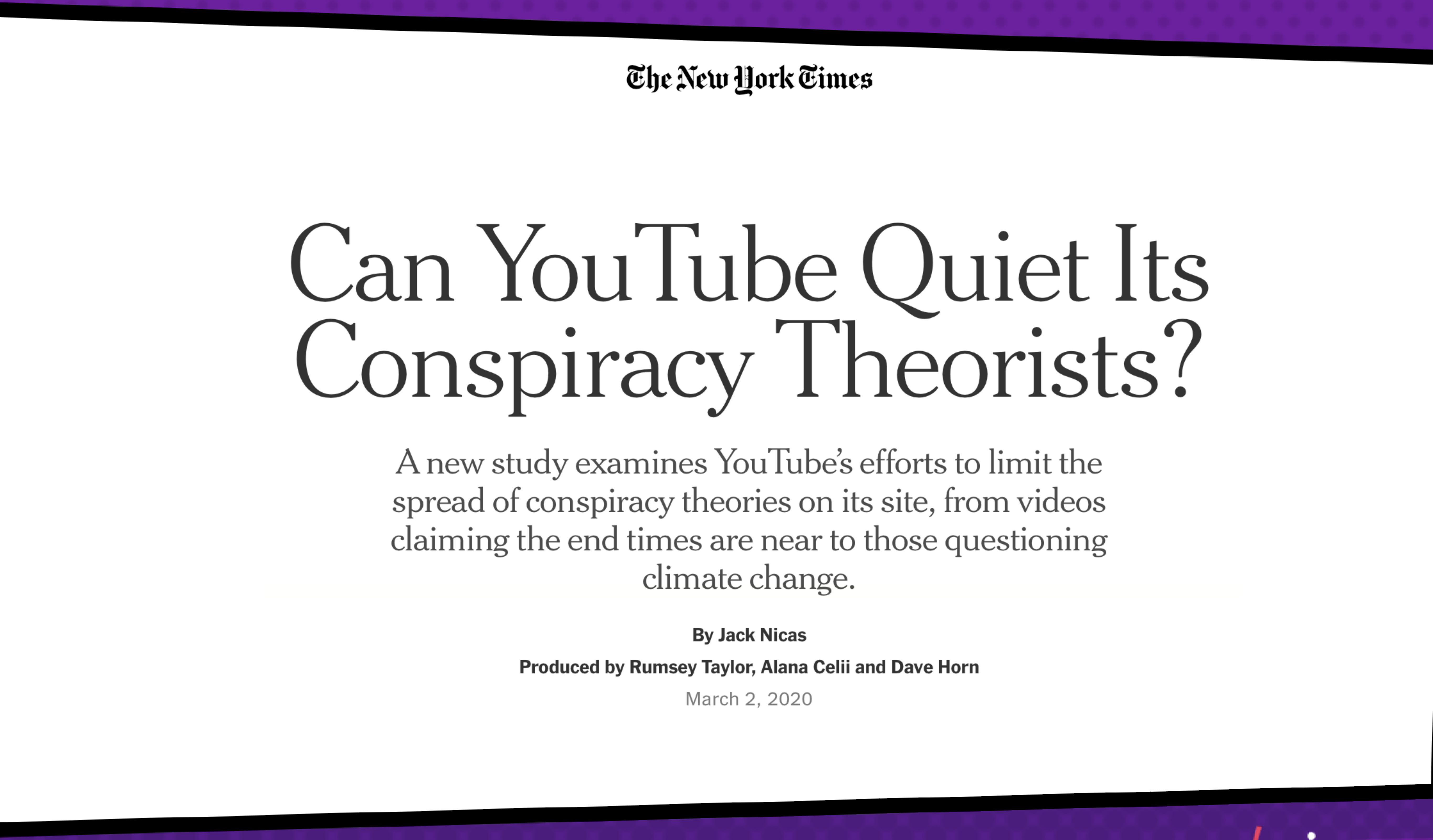
Costs, benefits, and risks

Stakeholders

- *Aim*: YouTube wants to optimize views
- *Task*: Experiment with recommendation engine
- *Data*: User profiles, page views, time spent watching, etc

Are there any other stakeholders?

Costs, benefits, and risks



The New York Times

Can YouTube Quiet Its Conspiracy Theorists?

A new study examines YouTube's efforts to limit the spread of conspiracy theories on its site, from videos claiming the end times are near to those questioning climate change.

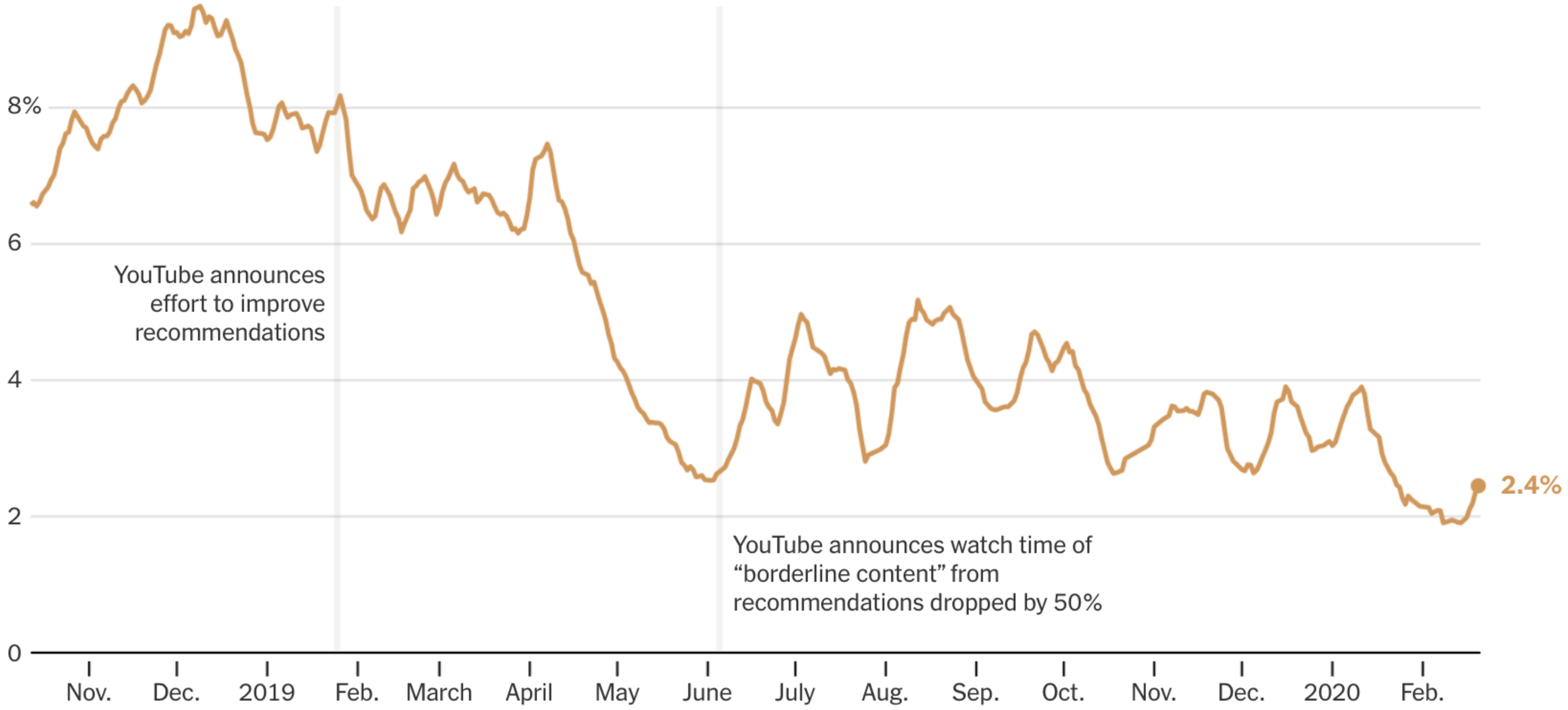
By Jack Nicas

Produced by Rumsey Taylor, Alana Celii and Dave Horn

March 2, 2020

Costs, benefits, and risks

This is the share of conspiracy videos recommended from top news-related clips



Note: Recommendations were collected daily from the “Up next” column alongside videos posted by more than 1,000 of the top news and information channels. The figures include only videos that ranked 0.5 or higher on the zero-to-one scale of conspiracy likelihood developed by the researchers. ▪ Source: Hany Farid and Marc Faddoul at University of California, Berkeley, and Guillaume Chaslot

Costs, benefits, and risks

Sensitive data

- *Aim*: YouTube wants to optimize views
- *Task*: Experiment with recommendation engine
- *Data*: User profiles, page views, time spent watching, etc

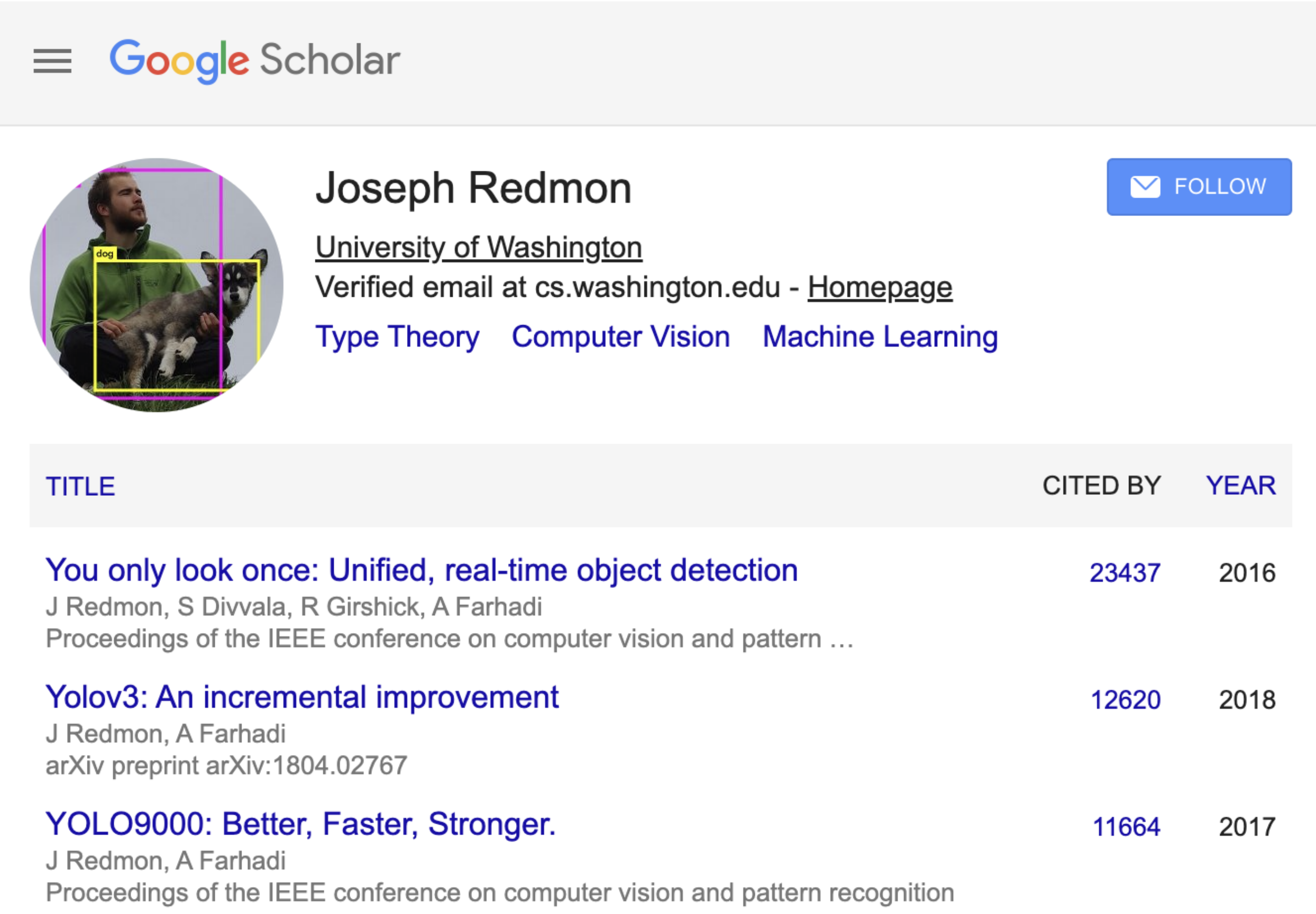
Costs, benefits, and risks

Potential for repurposing

- *Aim*: YouTube wants to optimize views
- *Task*: Experiment with recommendation engine
- *Data*: User profiles, page views, time spent watching, etc

Costs, benefits, and risks

Potential for repurposing



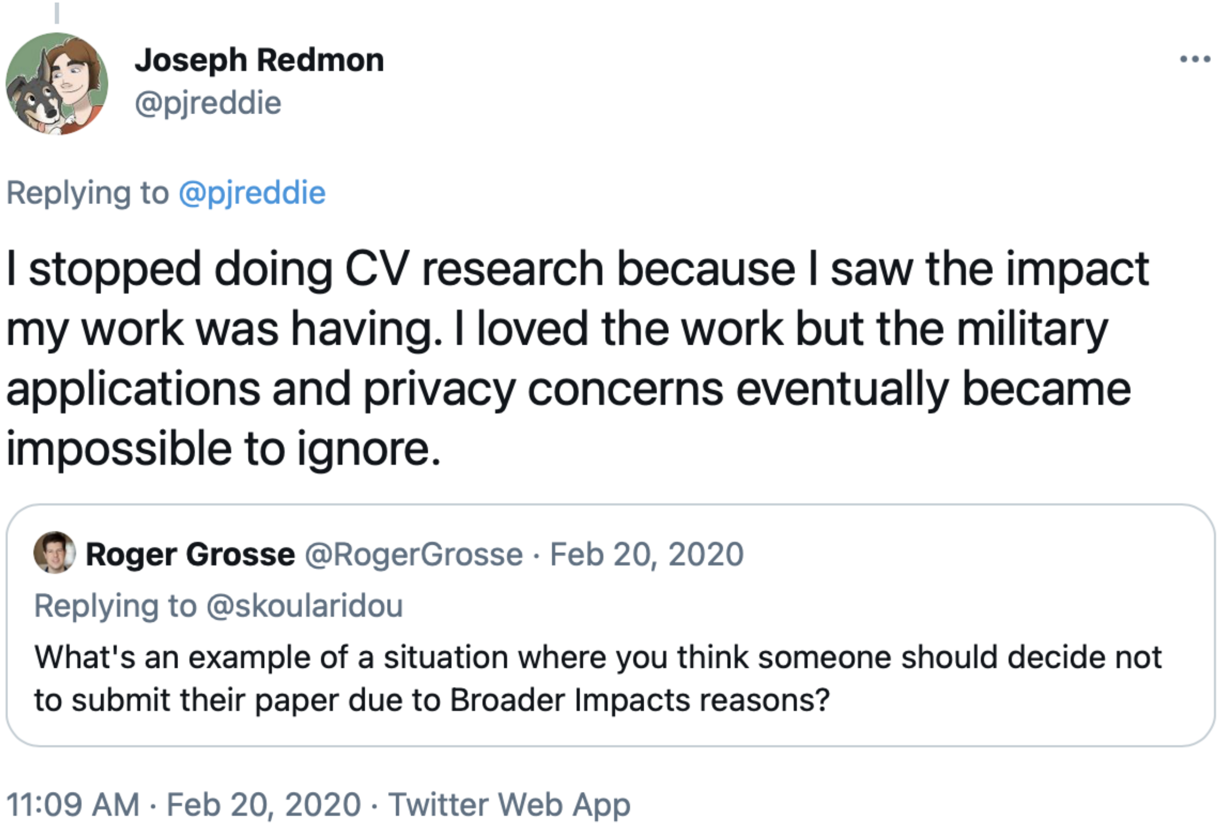
Joseph Redmon

University of Washington

Verified email at cs.washington.edu - Homepage

Type Theory Computer Vision Machine Learning

TITLE	CITED BY	YEAR
You only look once: Unified, real-time object detection J Redmon, S Divvala, R Girshick, A Farhadi Proceedings of the IEEE conference on computer vision and pattern …	23437	2016
Yolov3: An incremental improvement J Redmon, A Farhadi arXiv preprint arXiv:1804.02767	12620	2018
YOLO9000: Better, Faster, Stronger. J Redmon, A Farhadi Proceedings of the IEEE conference on computer vision and pattern recognition	11664	2017



Joseph Redmon
@pjreddie

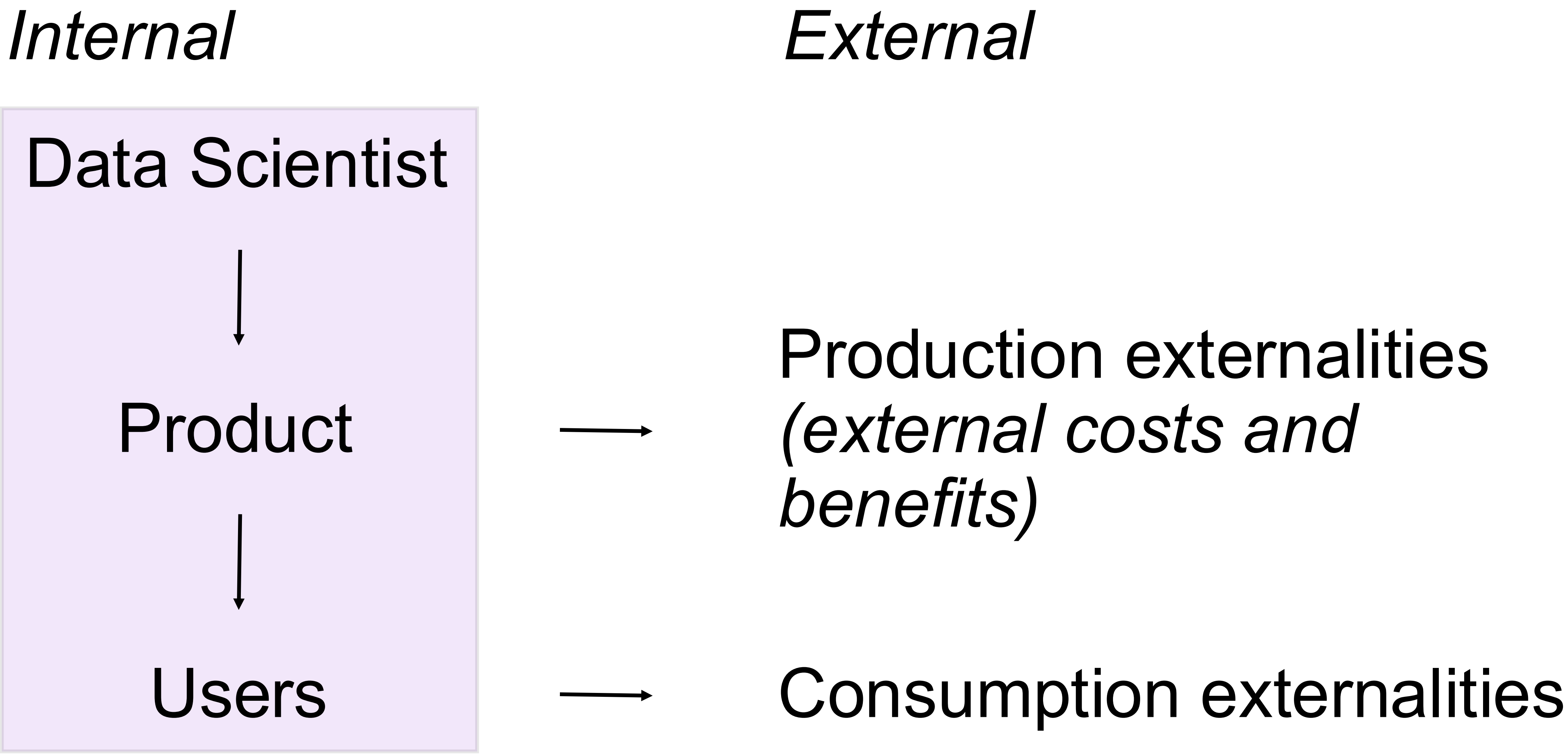
Replying to @pjreddie

I stopped doing CV research because I saw the impact my work was having. I loved the work but the military applications and privacy concerns eventually became impossible to ignore.

> **Roger Grosse** @RogerGrosse · Feb 20, 2020
> Replying to @skoularidou
> What's an example of a situation where you think someone should decide not to submit their paper due to Broader Impacts reasons?

11:09 AM · Feb 20, 2020 · Twitter Web App externalities

Costs, benefits, and risks



Internal

External

Data Scientist

Product

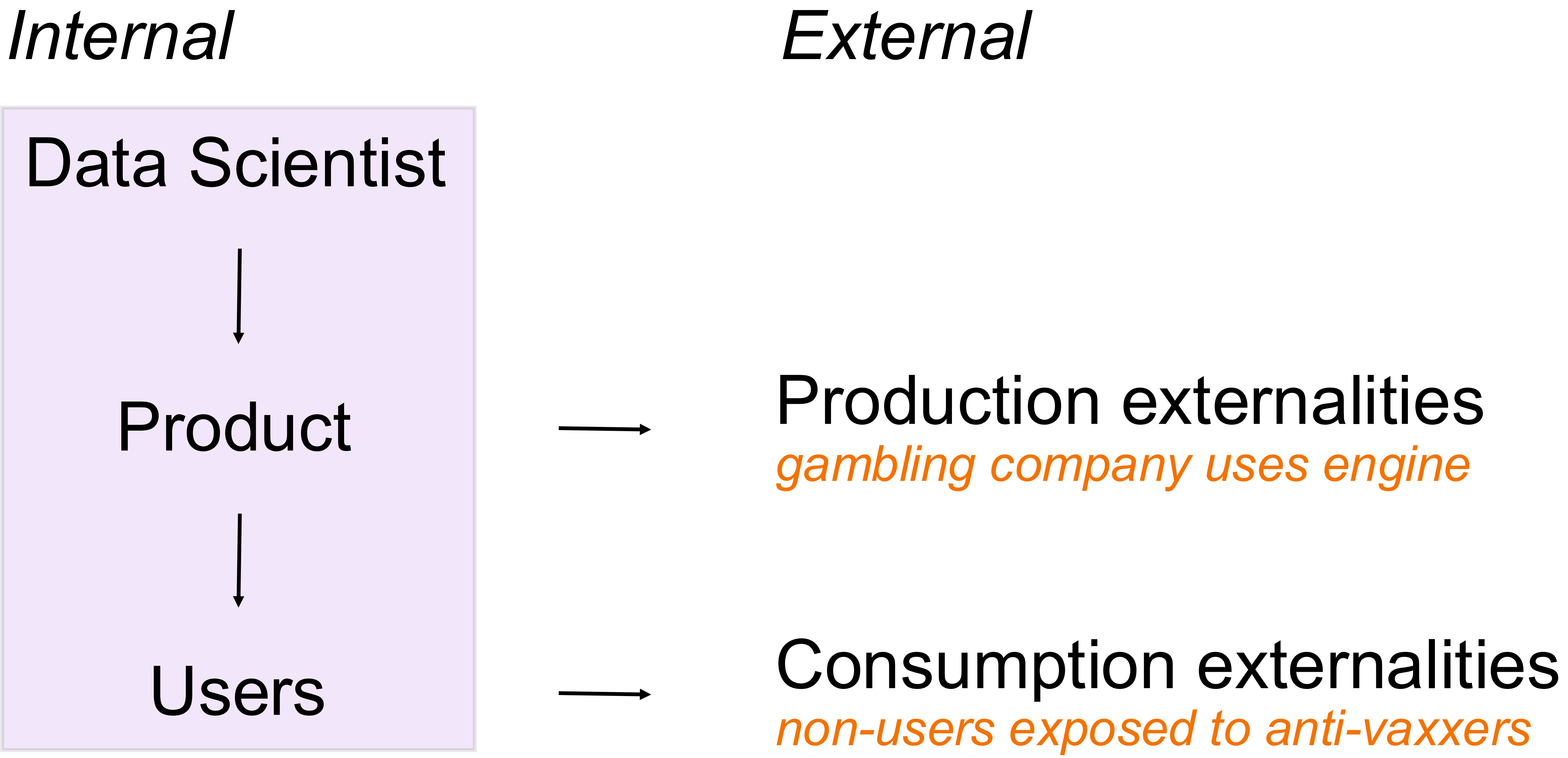
Users

Production externalities *(external costs and benefits)*

Consumption externalities

Costs, benefits, and risks

YouTube recommendation engine



Internal

External

Data Scientist → Product → Users

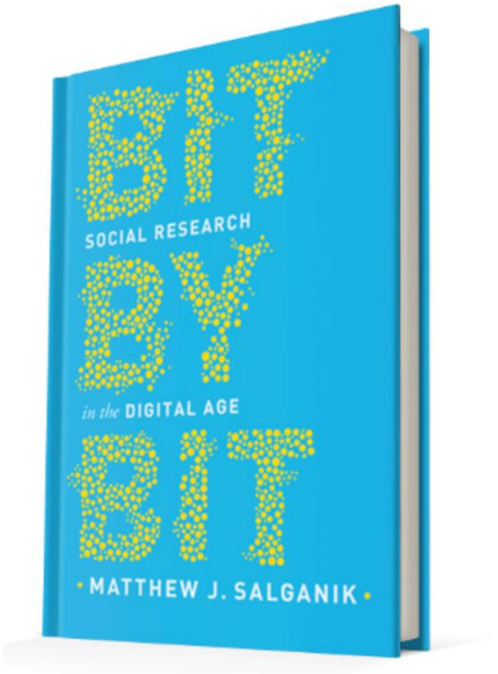
Product → Production externalities
gambling company uses engine

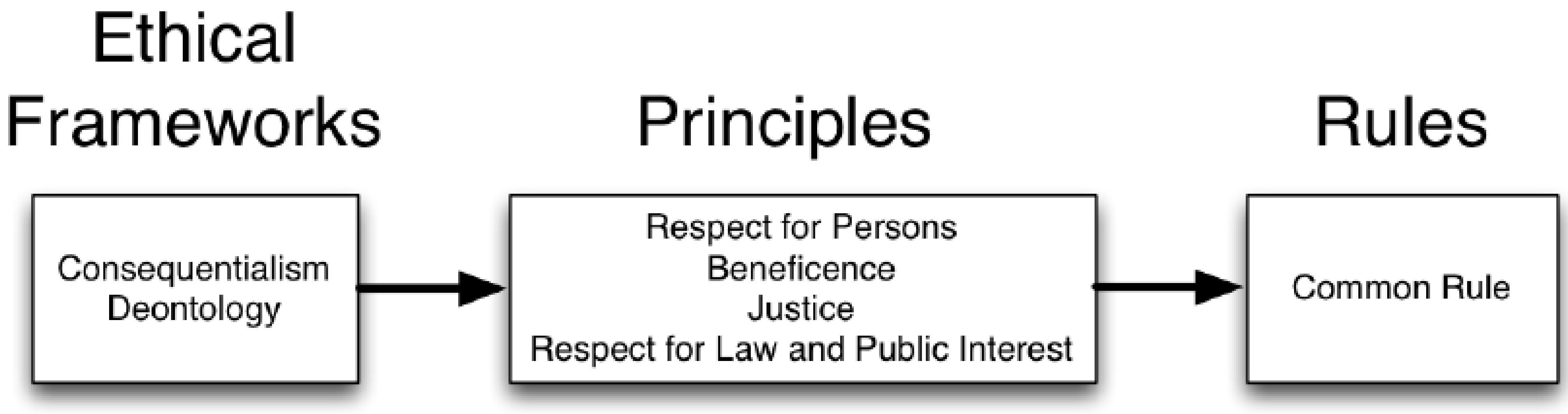
Users → Consumption externalities
non-users exposed to anti-vaxxers

What are the incentives for YouTube to capture these externalities?

principles-based approach to ethics

Bit by Bit, Chapter 6: Ethics

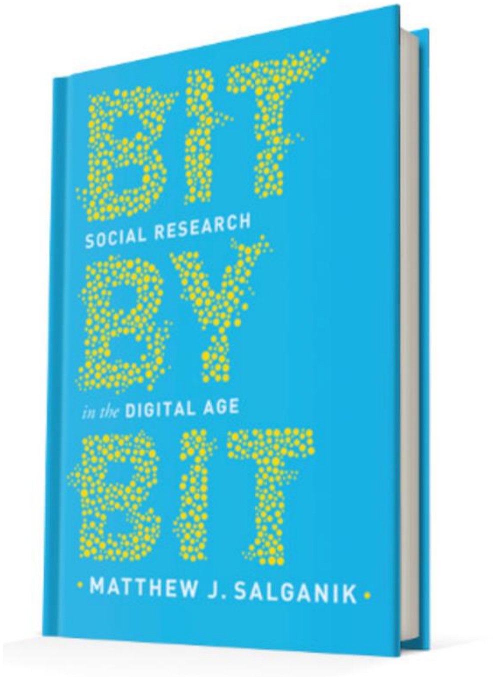


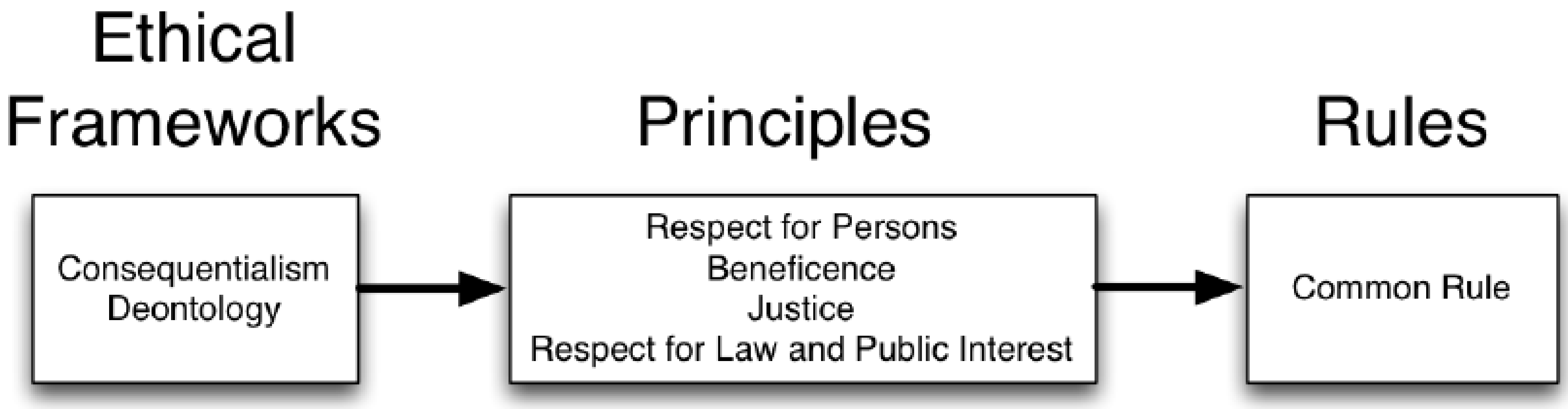


The **rules** governing research are derived from **principles** that in turn are derived from **ethical frameworks**. A main argument of this chapter is that researchers should evaluate their research through existing rules—which I will take as given and assume should be followed—and through more general ethical principles.

https://www.bitbybitbook.com/en/1st-ed/ethics/

Bit by Bit, Chapter 6: Ethics



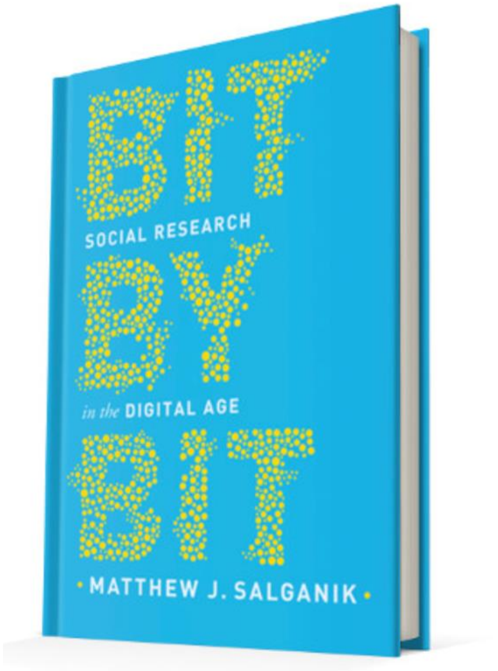


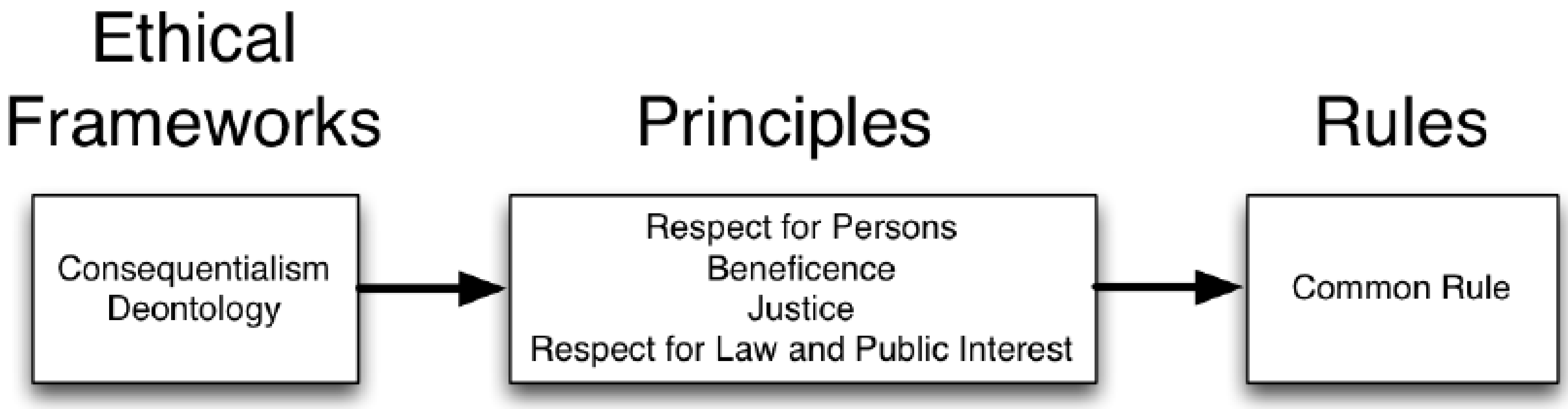
The **Common Rule** is the set of regulations currently governing most federally funded research in the United States… The four **principles** come from two blue-ribbon panels that were created to provide ethical guidance to researchers: the **Belmont Report** and the **Menlo Report**.

https://www.bitbybitbook.com/en/1st-ed/ethics/

r/ai

Bit by Bit, Chapter 6: Ethics





Ethical Frameworks

Consequentialism
Deontology

Principles

Respect for Persons
Beneficence
Justice
Respect for Law and Public Interest

Rules

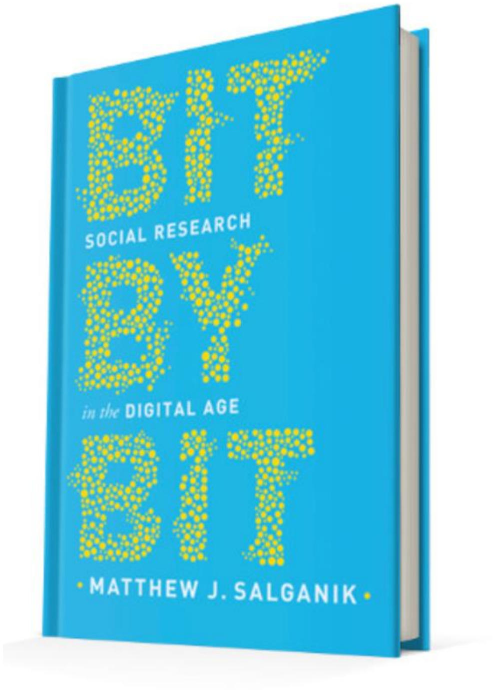
Common Rule

Finally, **consequentialism** and **deontology** are ethical frameworks that have been developed by philosophers for hundreds of years. A quick and crude way to distinguish the two frameworks is that deontologists focus on means and consequentialists focus on ends.

https://www.bitbybitbook.com/en/1st-ed/ethics/

r/ai

A principles-based approach to ethics



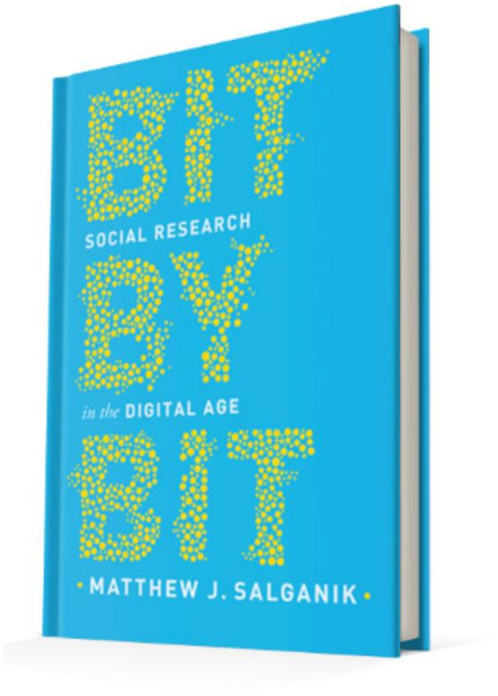
" … Neither of these approaches—the rules-based approach of social scientists or the ad hoc approach of data scientists—is well suited for social research in the digital age. Instead, I believe that we, as a community, will make progress if we adopt a **principles-based approach**.

That is, researchers should evaluate their research through existing rules—which I will take as given and assume should be followed—and through more general ethical principles. This **principles-based approach helps researchers make reasonable decisions for cases where rules have not yet been written**, and it helps researchers **communicate their reasoning to each other and the public**. "

https://www.bitbybitbook.com/en/1st-ed/ethics/

r/ai

A principles-based approach to ethics



“In some cases the principles-based approach leads to clear, actionable solutions. And, when it does not lead to such solutions, it clarifies the **trade-offs involved**, which is critical for striking an appropriate balance.

Further, the principles-based approach is sufficiently general that it will be helpful no matter where you work.”

https://www.bitbybitbook.com/en/1st-ed/ethics/

r/ai

two ethical frameworks

Two ethical frameworks

Consequentialism (Jeremy Bentham, John Stuart Mill): Take actions that lead to better states in the world

Deontology (Immanuel Kant): Focus on ethical duties, independent of their consequences

Deontologists focus on ***means*****, consequentialists** focus on ***ends***

"Arguments between consequentialists and deontologists are like two ships passing in the night."

https://www.bitbybitbook.com/en/1st-ed/ethics/

Two ethical frameworks

Deontologists focus on ***means***, **consequentialists** focus on ***ends***

Individuals, to the degree that they are capable, should be given the opportunity to choose what shall or shall not happen to them. This opportunity is provided when adequate standards for informed consent are satisfied.

Both consequentialism and deontology support **informed consent**, but for different reasons.

https://www.bitbybitbook.com/en/1st-ed/ethics/

Two ethical frameworks

Deontologists focus on ***means*****, consequentialists** focus on ***ends***

A **consequentialist** argument: Informed consent helps prevent harm to participants by prohibiting research that does not properly balance risk and anticipated benefit. In other words, consequentialist thinking would support informed consent because it helps **prevent bad outcomes** for participants.

A **deontological** argument for informed consent focuses on a researcher's duty to respect the **autonomy** of participants.

Given these arguments, a pure consequentialist might be willing to waive the requirement for informed consent in a setting where there was no risk, whereas a pure deontologist would not.

https://www.bitbybitbook.com/en/1st-ed/ethics/

Two ethical frameworks

Deontologists focus on ***means*****, consequentialists** focus on ***ends***

Transplant: A doctor has five patients dying of organ failure and one healthy patient whose organs can save all five. A **consequentialist doctor is required to kill** the healthy patient to obtain his organs. This complete focus on ends, without regard to means, is flawed.

NB: putting a price on human life

NB: reasoning under uncertainty

Time bomb: A police office captured a terrorist who knows the location of a ticking time bomb that will kill millions of individuals if it detonates. A **deontological police officer would not lie** to trick a terrorist into revealing the location of the bomb. This complete focus on means, without regards to ends, also is flawed.

https://www.bitbybitbook.com/en/1st-ed/ethics/

The sad reality of the pandemic

IDEAS

The Atlantic

The Extraordinary Decisions Facing Italian Doctors

There are now simply too many patients for each one of them to receive adequate care.

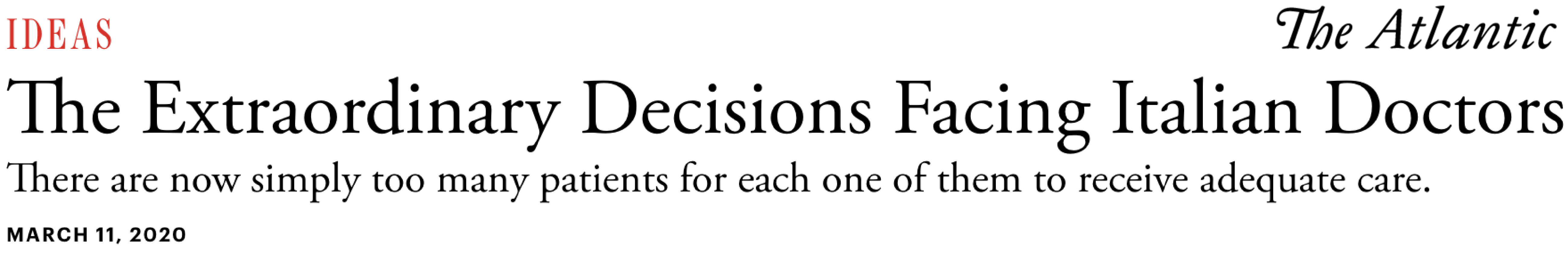
MARCH 11, 2020

Now the Italian College of Anesthesia, Analgesia, Resuscitation and Intensive Care (SIAARTI) has published guidelines for the criteria that doctors and nurses should follow as these already extraordinary circumstances worsen. The document begins by likening **the moral choices Italian doctors may face to the forms of wartime triage that are required in the field of "catastrophe medicine."**

Instead of providing intensive care to all patients who need it, the authors suggest, it may become necessary to follow **"the most widely shared criteria regarding distributive justice and the appropriate allocation of limited health resources.**

https://www.theatlantic.com/ideas/archive/2020/03/who-gets-hospital-bed/607807/

The sad reality of the pandemic



IDEAS

The Atlantic

The Extraordinary Decisions Facing Italian Doctors

There are now simply too many patients for each one of them to receive adequate care.

MARCH 11, 2020

The principle they settle upon is utilitarian. **"Informed by the principle of maximizing benefits for the largest number,"** they suggest that "the allocation criteria need to guarantee that those patients with the highest chance of therapeutic success will retain access to intensive care."

"…I must admit that I have no moral judgment to make about the extraordinary document published by those brave Italian doctors. I have not the first clue whether they are recommending the right or the wrong thing. … But if Italy is in an impossible position, **the obligation facing the United States is very clear: To arrest the crisis before the impossible becomes necessary.**"

https://www.theatlantic.com/ideas/archive/2020/03/who-gets-hospital-bed/607807/

The sad reality of the pandemic

The New York Times

'Chilling' Plans: Who Gets Care as Washington State Hospitals Fill Up?

By Karen Weise and Mike Baker

Published March 20, 2020
Updated March 22, 2020, 10:26 a.m. ET

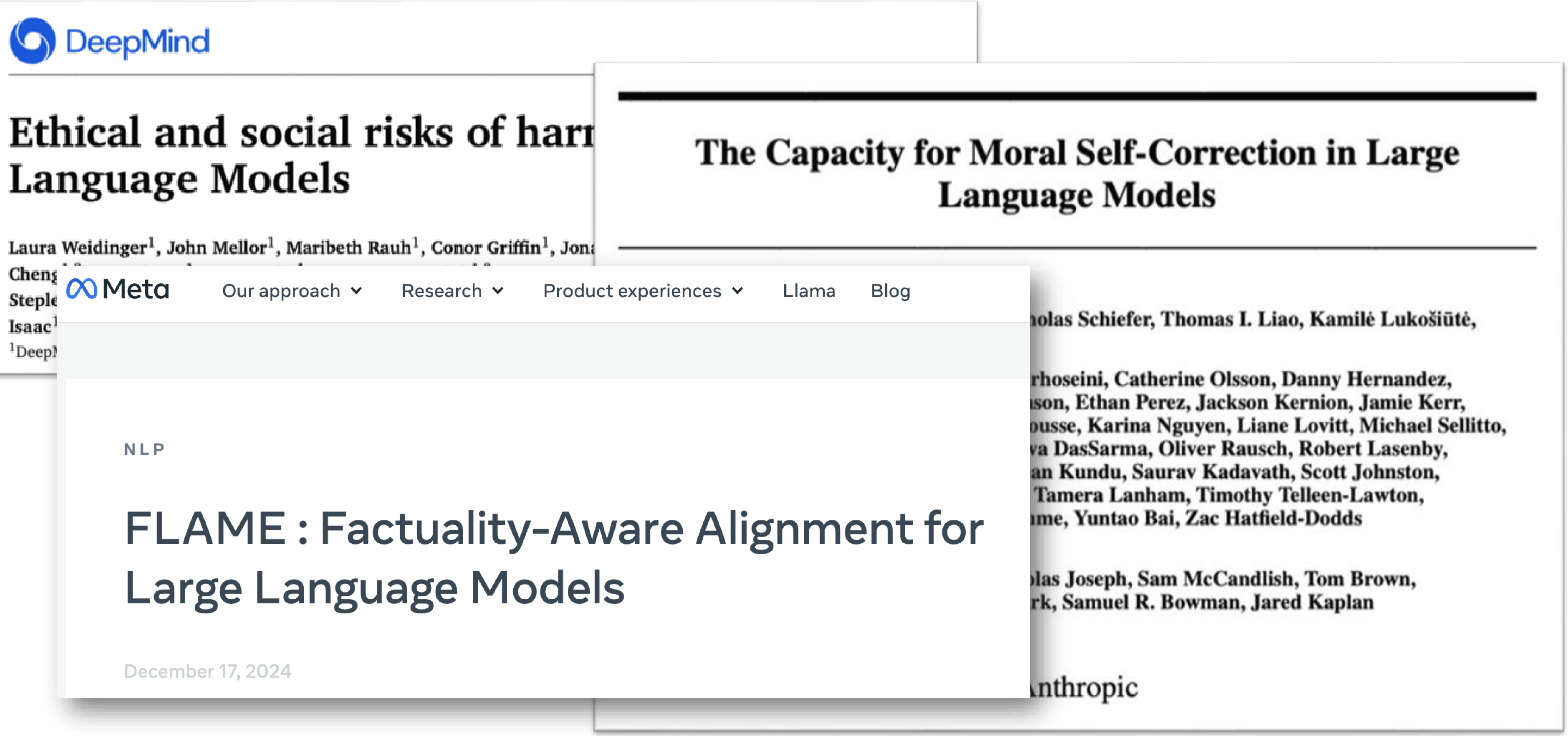
SEATTLE — Medical leaders in Washington State, which has the highest number of coronavirus deaths in the country, have quietly begun **preparing a bleak triage strategy to determine which patients may have to be denied complete medical care** in the event that the health system becomes overwhelmed by the coronavirus in the coming weeks.

…. It's protecting the clinicians so you don't have one person who's kind of playing God," she said, adding, "It is chilling, and it should not happen in America."

https://www.nytimes.com/2020/03/20/us/coronavirus-in-seattle-washington-state.html

LLMs

Perspective: Purveyors



DeepMind

Ethical and social risks of harr Language Models

Laura Weidinger[1], John Mellor[1], Maribeth Rauh[1], Conor Griffin[1], Jona
Cheng
Steple
Isaac[1]
[1]Deep

The Capacity for Moral Self-Correction in Large Language Models

olas Schiefer, Thomas I. Liao, Kamilė Lukošiūtė,

rhoseini, Catherine Olsson, Danny Hernandez,
ison, Ethan Perez, Jackson Kernion, Jamie Kerr,
ousse, Karina Nguyen, Liane Lovitt, Michael Sellitto,
va DasSarma, Oliver Rausch, Robert Lasenby,
an Kundu, Saurav Kadavath, Scott Johnston,
Tamera Lanham, Timothy Telleen-Lawton,
ime, Yuntao Bai, Zac Hatfield-Dodds

olas Joseph, Sam McCandlish, Tom Brown,
rk, Samuel R. Bowman, Jared Kaplan

nthropic

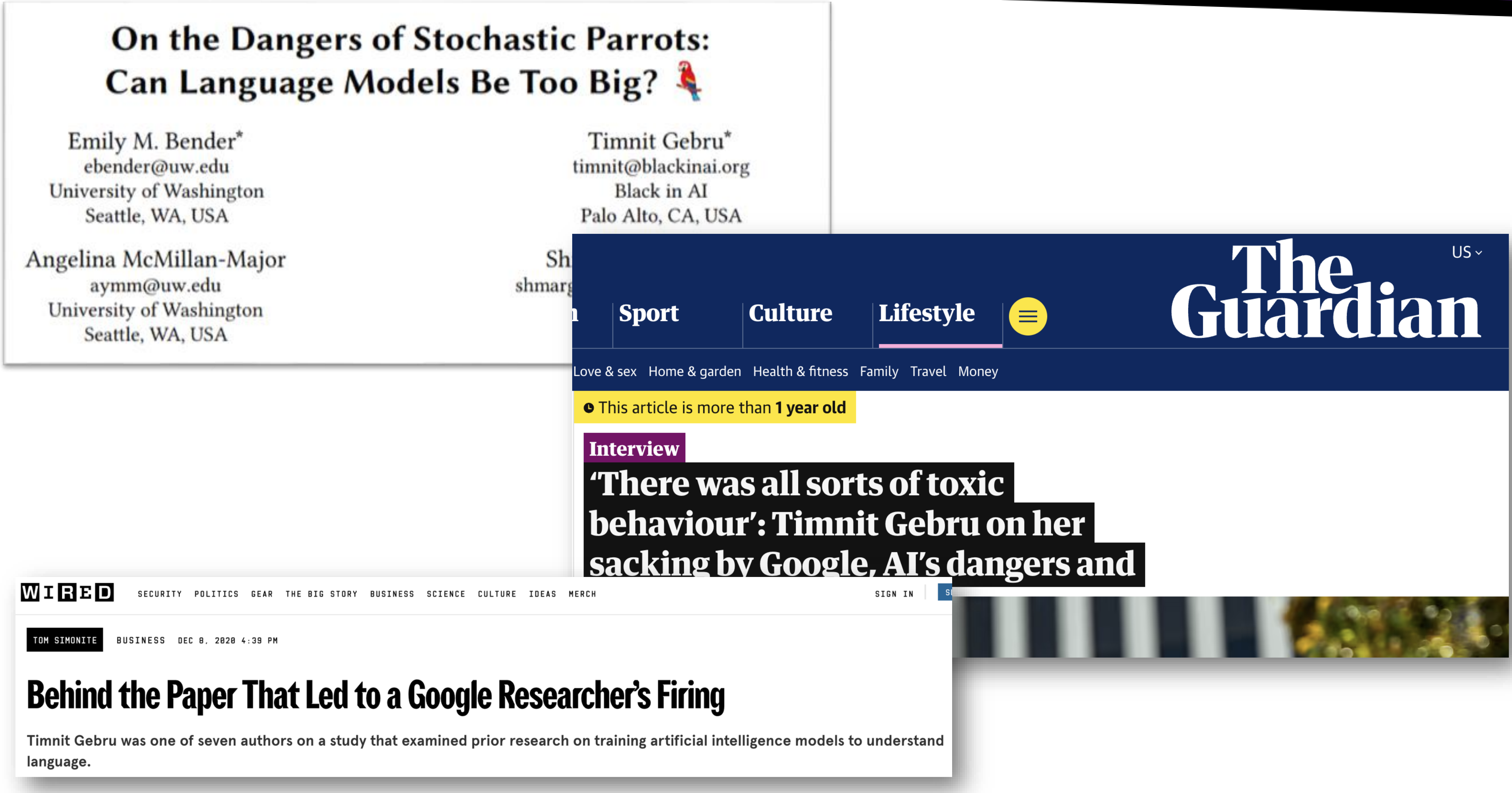
NLP

FLAME : Factuality-Aware Alignment for Large Language Models

December 17, 2024

https://arxiv.org/pdf/2112.04359 , https://arxiv.org/pdf/2302.07459 ,
https://arxiv.org/pdf/2405.01525

Perspective: Ethicists / Researchers



On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Sh
shmarg

This article is more than **1 year old**

Interview

'There was all sorts of toxic behaviour': Timnit Gebru on her sacking by Google, AI's dangers and

TOM SIMONITE BUSINESS DEC 8, 2020 4:39 PM

Behind the Paper That Led to a Google Researcher's Firing

Timnit Gebru was one of seven authors on a study that examined prior research on training artificial intelligence models to understand language.

https://dl.acm.org/doi/10.1145/3442188.3445922 , https://www.wired.com/story/behind-paper-led-google-researchers-firing/ , https://www.theguardian.com/lifeandstyle/2023/may/22/there-was-all-sorts-of-toxic-behaviour-timnit-gebru-on-her-sacking-by-google-ais-dangers-and-big-techs-biases.

Stochastic Parrots: The Fallout.

TLDR.

- **Dec 2, 2020**. Timnit Gebru announced she was "forced out" of Google after raising ethical concerns.
- **Why?** Her draft, "On the Dangers of Stochastic Parrots," was, in an internal email by Jeff Dean (head of Google AI), deemed not to **"meet our bar for publication."**
- Gebru and co-authors **argue** that training a single large model can emit hundreds of tonnes of CO_2, and that "it is past time for researchers to prioritize energy efficiency and cost…"
- **Argue** that models trained on vast, uncurated web crawls risk embedding bias. "[This] methodology that relies on datasets too large to document is therefore inherently risky."
- **Argue** chasing ever-bigger models diverts effort from more efficient, smaller systems that might advance true language understanding.
- **Argue** that highly fluent text can "fool people," enabling misinformation and other harms when models mimic but don't comprehend language.

Whether you agree or not, Dean's critique brought a lot of attention to the paper over publication rigor, as did Gebru's termination.

Co-author Emily Bender warned that this sort of incident has a potential "chilling effect" on AI ethics research…what do you think?

https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/

Stochastic Parrots? The debate.

LLMs lack…

Transparency!

If we don't see the training data (or, even if we saw it, we couldn't possibly query it due to intractability), then we can't prove novelty.

Also…

Do they fail the Chinese Room thought experiment?

LLMs can…

Generalize!

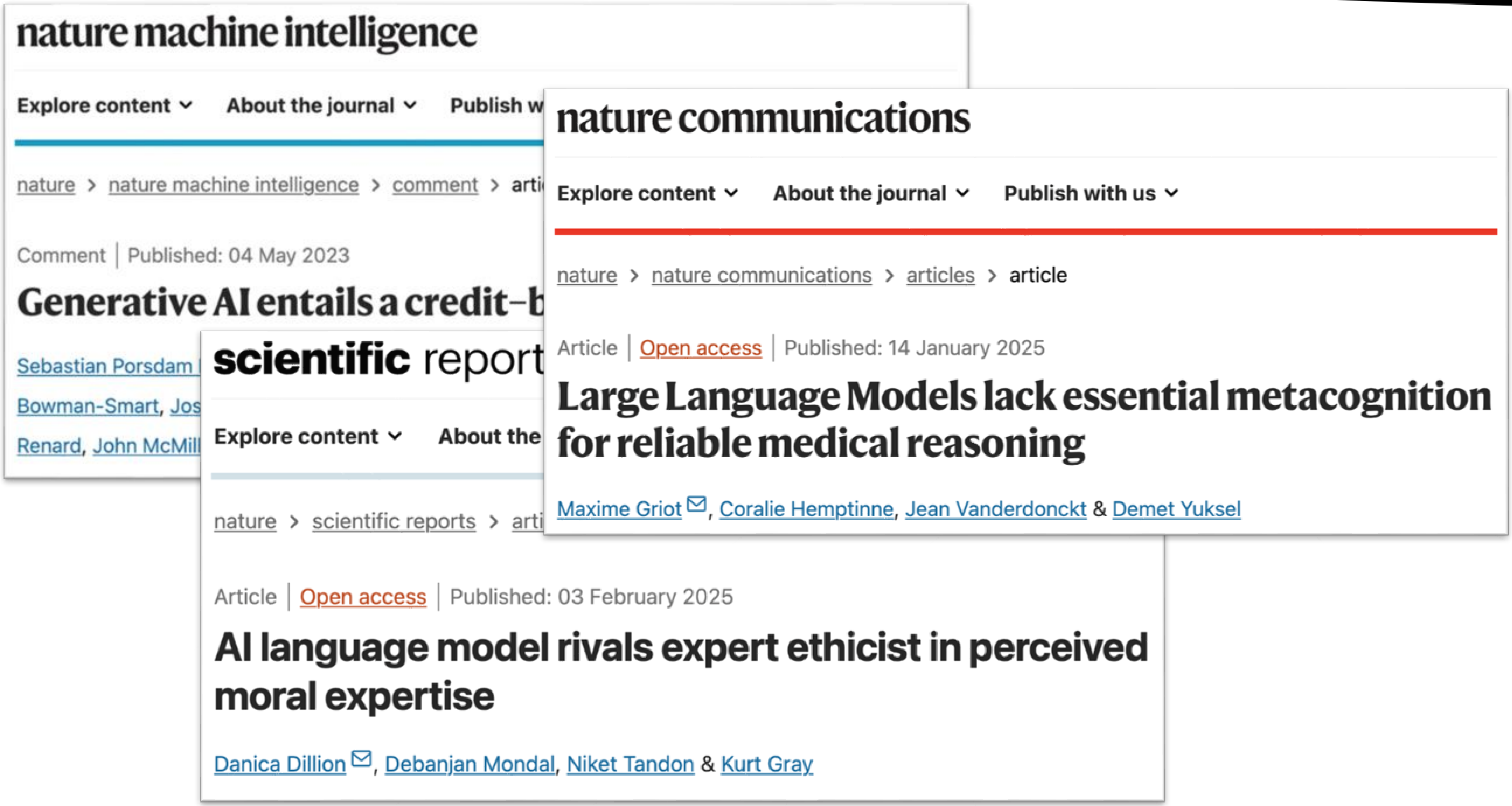
GPT > 3, for example, can adapt to new language games, or solve math problems with symbolic patterns unlikely to have occurred in the training data.

Also…

Does it matter if they are just a very efficient, but fundamentally "dumb," translator?

John Searle, https://en.wikipedia.org/wiki/Chinese_room#Computer_science

Open debate. Open problem.



nature machine intelligence

nature › nature machine intelligence › comment › arti

Comment | Published: 04 May 2023

Generative AI entails a credit–b

Sebastian Porsdam
Bowman-Smart, Jos
Renard, John McMill

nature communications

nature › nature communications › articles › article

Article | Open access | Published: 14 January 2025

Large Language Models lack essential metacognition for reliable medical reasoning

Maxime Griot ✉, Coralie Hemptinne, Jean Vanderdonckt & Demet Yuksel

scientific report

nature › scientific reports › arti

Article | Open access | Published: 03 February 2025

AI language model rivals expert ethicist in perceived moral expertise

Danica Dillion ✉, Debanjan Mondal, Niket Tandon & Kurt Gray

Deontological vs. Consequentialism

Who is at fault if LLMs cause harm?

Debate.

Responsible Data Science

Applied Ethics

Thank you!

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai center for responsible ai