

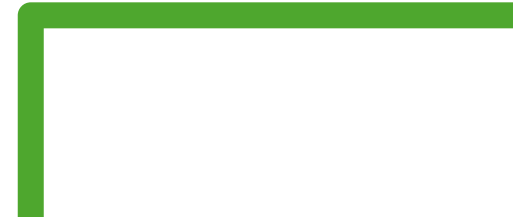
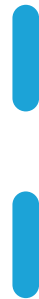
Can Wikipedia Come to AI's Rescue (Again)?

Brent Hecht

Microsoft / Northwestern

20 June 2024

brenthecht.com



A bit about me

Northwestern
University

Associate Professor

Language models and AI, in particular the “content ecosystems” they require

August 2019



“Come help
redefine the future
of work”



**Director of Applied
Science & Partner**

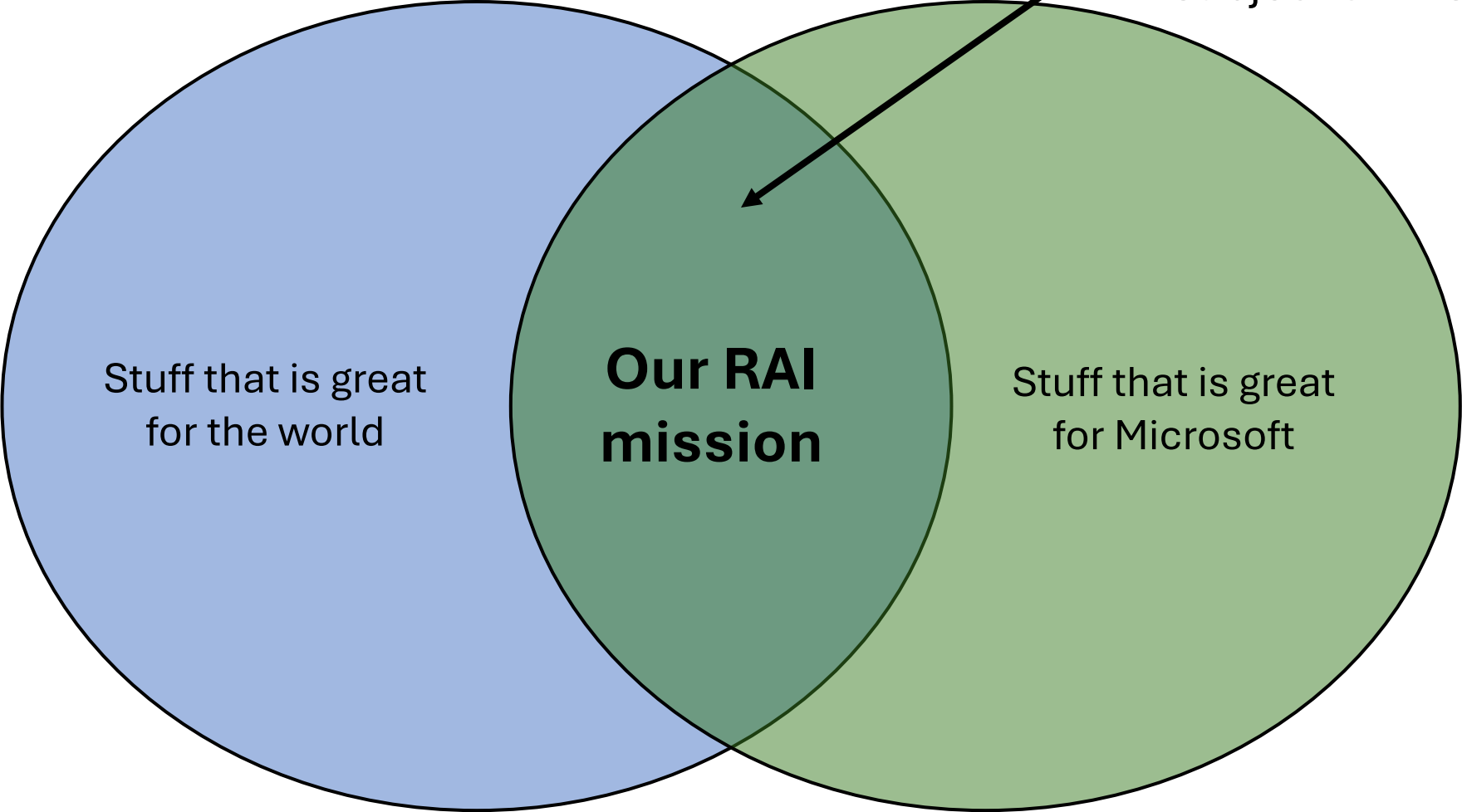
Experiences and Devices,
Microsoft

*Increasing the pace, impact,
and responsibility of research*



Theory of change

Hard to imagine a higher-impact topic in this intersection than the subject of this talk



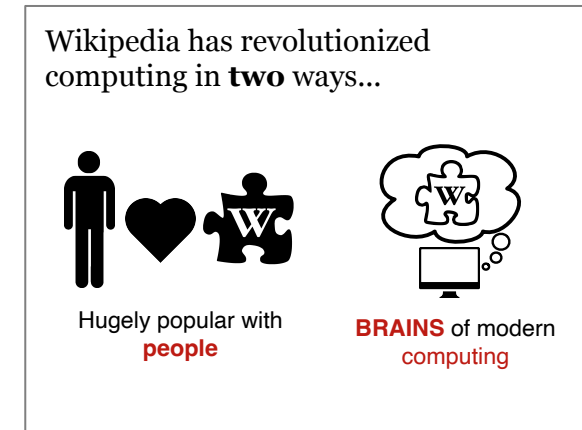
A brief outline

- **Looking back:** Wikipedia's central role in the development of modern AI (case study from my career)
- **The present conundrum:** The dominant LLM paradigm threatens Wikipedia, large portions of the content ecosystem, and ultimately itself.
- **Looking forward:** What Wikipedia can do to help itself in the LLM era, and make the LLM era much better in the process



Simply stated: Modern AI does not exist without the “Wikipedia dataset”

- Strong argument it was the single most important dataset for AI research since about 2005
 - When the stochastic turn in AI needed data, Wikipedia was there to provide it.
 - Used as the research dataset of first resort in many areas for two decades
- Semantic relatedness (“proto-LLMs”), knowledge graphs, information retrieval, information extraction...
- Core to basically every LLM training dataset from the beginning up through now
- Of course not just academia and LLMs, also many other commercial applications



Slide from WikiSym 2012 Keynote



Available online at www.sciencedirect.com

ScienceDirect

Int. J. Human-Computer Studies 67 (2009) 716–754

International Journal of
Human-Computer
Studies

www.elsevier.com/locate/ijhcs

Mining meaning from Wikipedia

Olena Medelyan*, David Milne, Catherine Legg, Ian H. Witten

University of Waikato, Knighton Road, Hamilton 3216, New Zealand

Received 23 February 2009; received in revised form 14 April 2009; accepted 1 May 2009

Communicated by e. motta
Available online 29 May 2009

Abstract

Wikipedia is a goldmine of information; not just for its many readers, but also for the growing community of researchers who recognize it as a resource of exceptional scale and utility. It represents a vast investment of manual effort and judgment: a huge, constantly evolving tapestry of concepts and relations that is being applied to a host of tasks.

[Medelyan et al. 2009](#)

I used to try to keep track of all the ways Wikipedia was supporting very successful AI products...

strengths of the Satori Ontology is the ability to manage multiple source schemas and cross domain linkages.”

- ▶ Excel
- ▶ Flight Simulator
- ▼ **Apple**
 - ▼ Siri
 - ▶ Examples via vandalism
 - ▼ Examples are easy to come by
 - ▶ “Who is Grover Cleveland?”
 - ▶ “Tell me something about the New York Times”
 - ▼ A few moderately useful references
 - <https://techcrunch.com/2013/06/10/apple-updates-siri-with-twitter-wikipedia-bing-integration-new-commands-and-male-and-female-voice/>
 - <https://www.thesun.co.uk/tech/4224107/what-is-an-indian-apples-siri-has-a-racist-and-offensive-answer-to-this-question/>
 - ▶ They call it “Siri” knowledge
 - ▶ Apple Maps
- ▼ **Google**
 - ▼ Google Search
 - ▼ Massively important relationship for both Google and Wikipedia
 - A detailed overview can be found in this paper: <https://www.aaii.org/ocs/index.php/ICWSM/ICWSM17/paper/download/15623/14799>
 - ▼ Wikipedia 2x important as deep learning to Google’s search results performance

- ▼ BERT
 - [omnioutliner://open?row=ke2RXiyn5BN](https://omnioutliner.com/open?row=ke2RXiyn5BN)
- ▼ YouTube
 - ▼ Context for “fake news”
 - <https://www.vanityfair.com/news/2018/03/youtube-wikipedia-conspiracy-theory-video-problem>
 - <https://twitter.com/krmaher/status/973791488583888896>
 - ▼ Big Google datasets
 - [KELM: Integrating Knowledge Graphs with Language Model Pre-training Corpora](https://www.kelmscience.com/)
- ▼ **Wolfram Alpha**
 - “Since the inception of [WolframAlpha](https://www.wolfram.com/language/), Wikipedia has held a special place in its development pipeline. We usually use it not as a primary source for data, but rather as an essential resource for improving our natural language understanding, particularly for mining the common and colloquial ways people refer to entities and concepts in various domains.”
 - <http://blog.wolfram.com/2015/03/20/new-in-the-wolfram-language-wikipediadata/>
- ▼ **Facebook**
 - ▼ Providing context for news
 - ▼ <https://newsroom.fb.com/news/2018/04/news-feed-fyi-more-context/>

Helping People Better Assess the Stories They See in News Feed | F...

<https://newsroom.fb.com/news/2018/04/news-feed-fyi-more-context/>

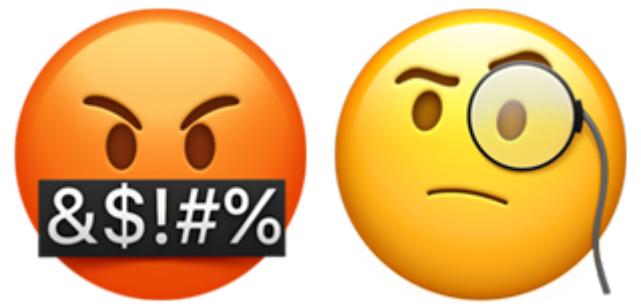
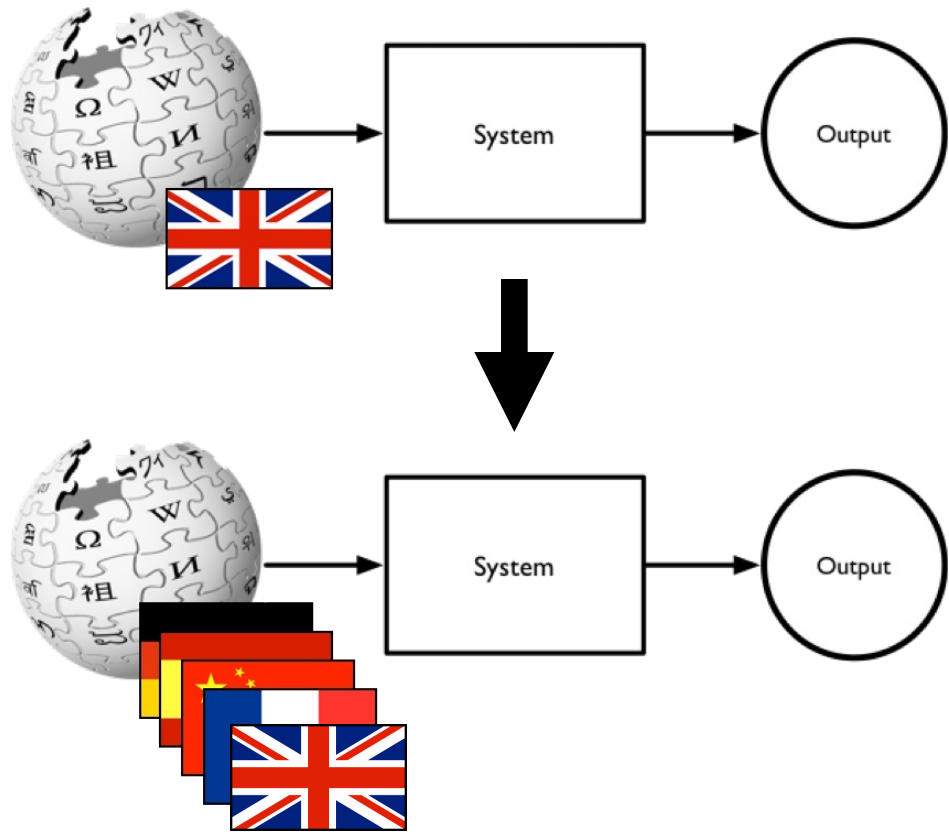


...but it got to be way too much.

To provide some color to the importance of Wikipedia, let's go back to around 2008 or so...



A failed attempt at improving semantic relatedness algorithms

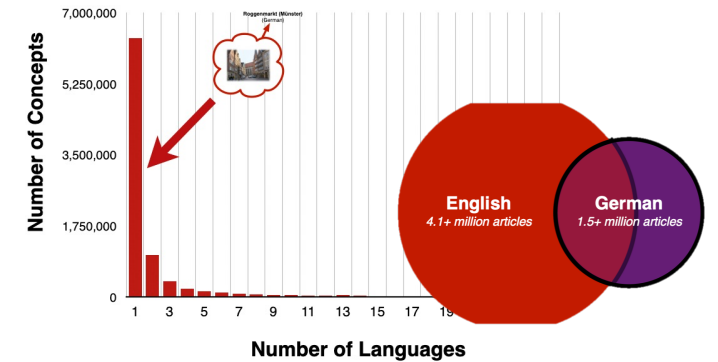


Wikipedia and uncovering “algorithmic bias”

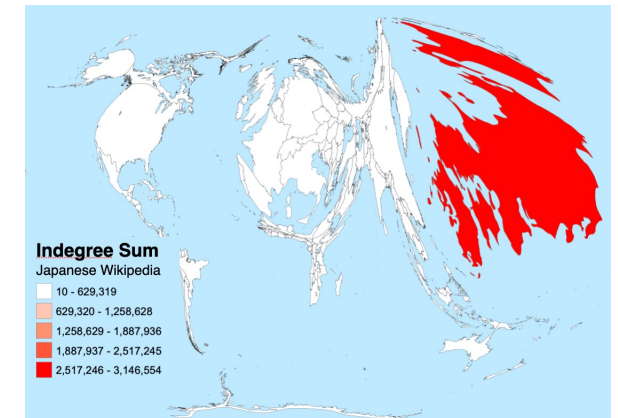
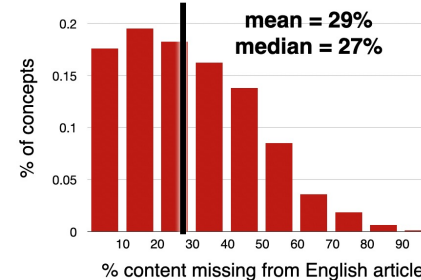
- One way I’m in debt to Wikipedia is it helped us uncover what we now know as “algorithmic bias”
- Wikipedia’s role here should be better known
- Key papers: [Hecht and Gergle 2010](#), [Bao et al. 2012](#)



Distribution of Languages Per Concept



PctMissingFromEnglish Distribution (including “missing” links)



Wikipedia and uncovering “algorithmic bias”

$SR_{\text{ENGLISH}}(\text{Concept A}, \text{Concept B})$

=?

$SR_{\text{GERMAN}}(\text{Concept A}, \text{Concept B})$

=?

$SR_{\text{SPANISH}}(\text{Concept A}, \text{Concept B})$

=?

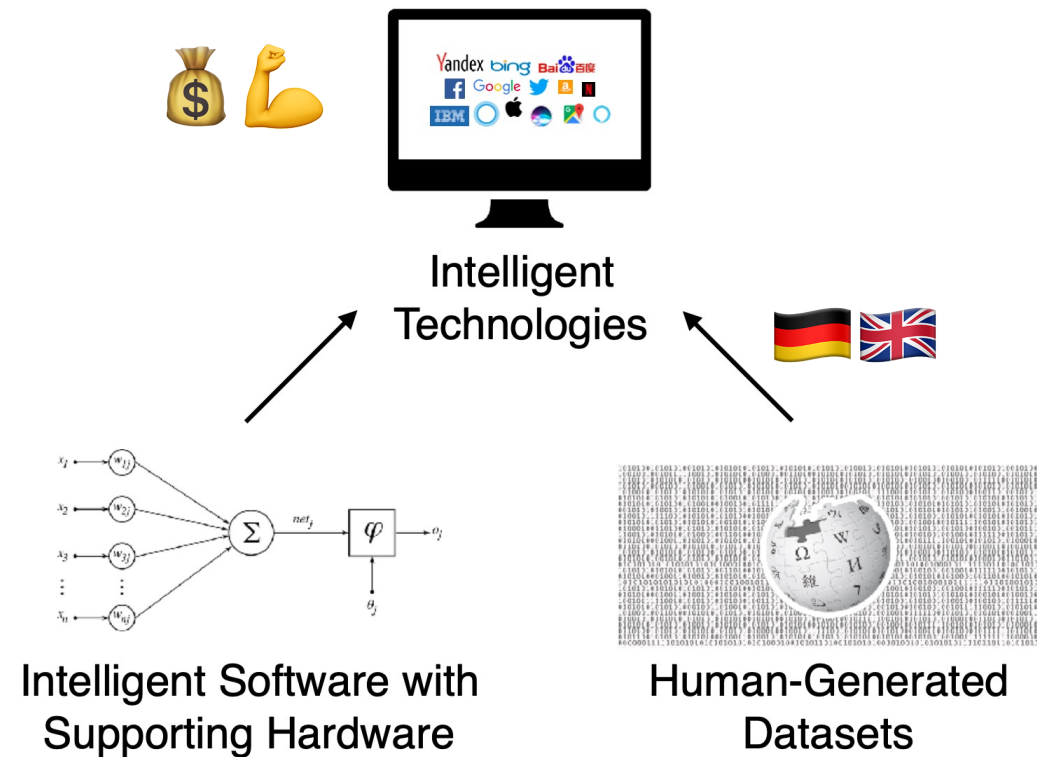
.....

Cross-language

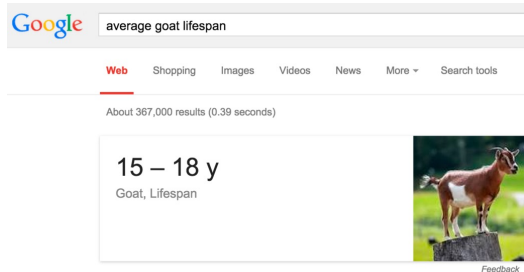
SR Measure	Mean r_s	Max r_s
WikiRelate	0.33	0.52
MilneWitten	0.47	0.55
OutlinkOverlap	0.36	0.48
WAGDirect	0.41	0.54
Explicit Semantic Analysis	0.41	0.58

From studying cultural biases to working to address harmful power imbalances

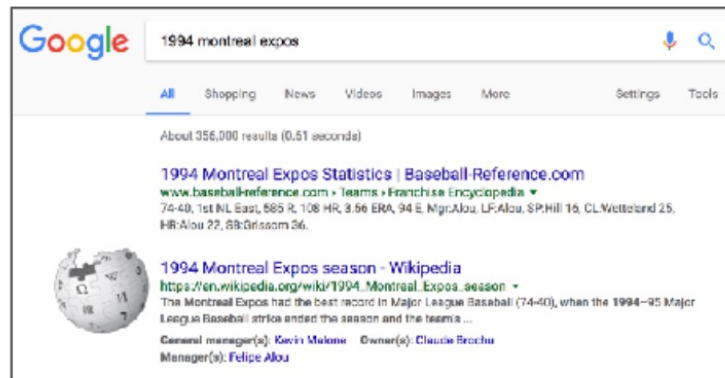
- First: Wikipedia as a way to advance AI
- Then: Concerned about power relationships between Wikipedia and AI technologies
- Finally: All data/content creators, not just Wikipedia



Wikipedia links increase click-through rates by 80%...



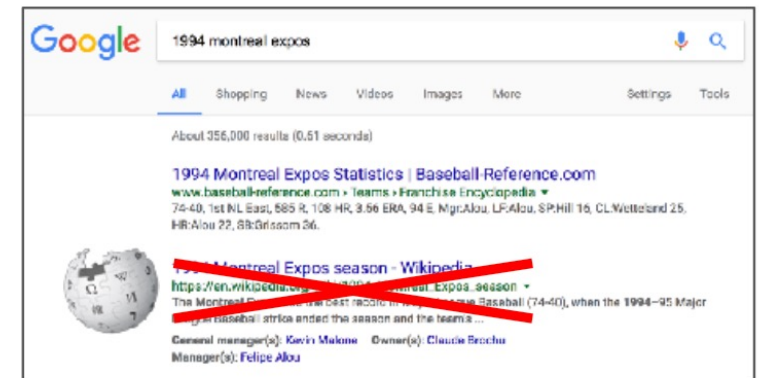
(Taraborelli 2015)



Wikipedia Links Present

618 queries

CTR = **26.1%**

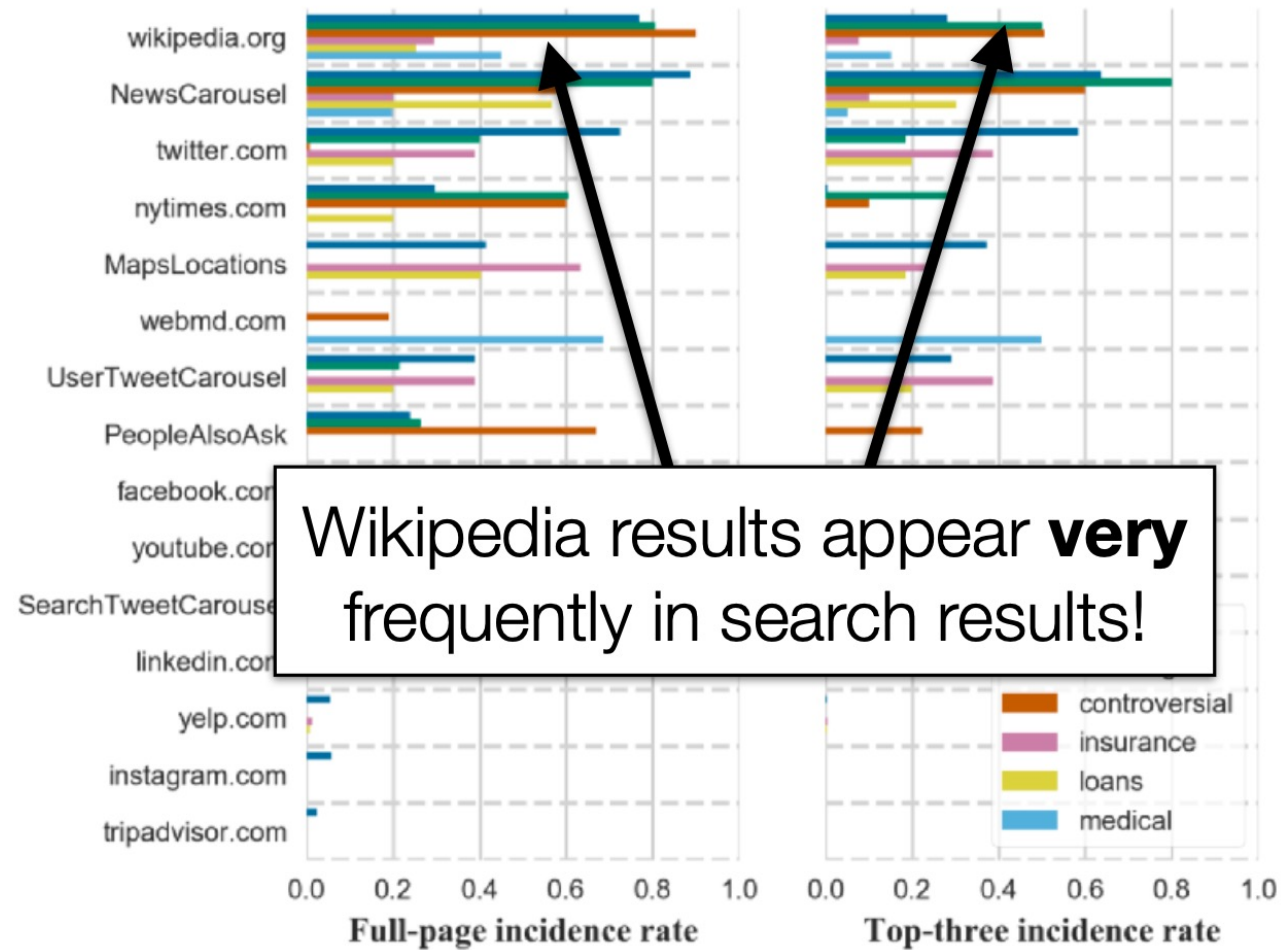


Wikipedia Links Removed

387 queries

CTR = **14.0%**

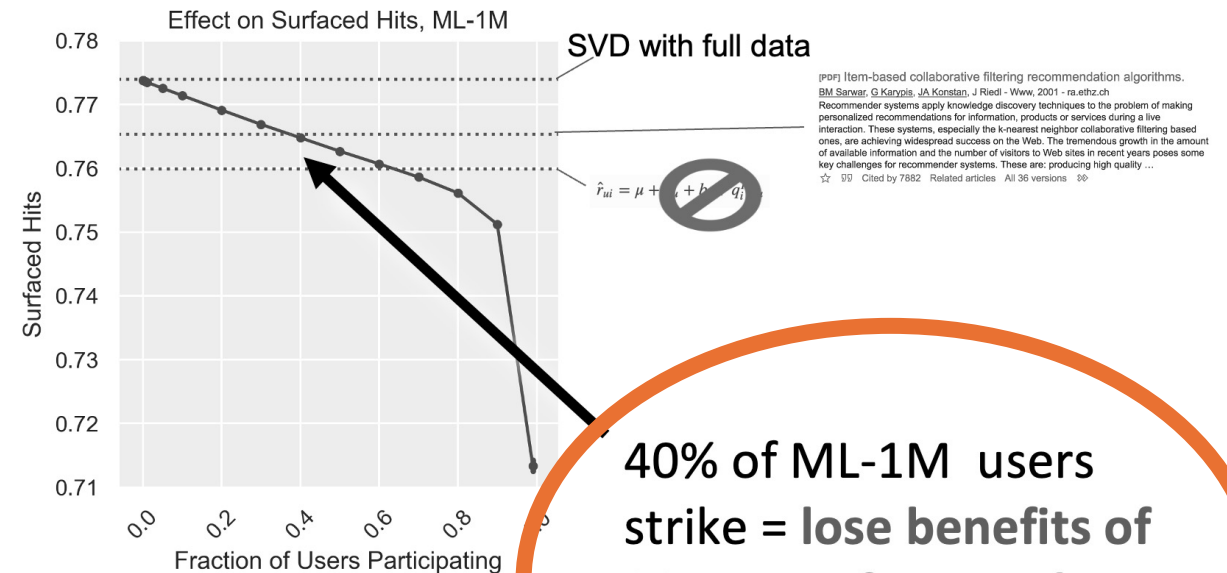
Context: A CTR improvement of 1% is a big deal



[Vincent et al. ICWSM 2019](#); [Vincent et al. CSCW 2021](#)

Key concept: The “data strike”

- “Data strikes” ([Vincent et al. 2019](#)) occur when a significant proportion of content producers for a given AI/ML system stop producing that content in order to force a change in behavior by the system owner.
- Data strikes have always been a substantial lurking force in AI and the computing industry
- Main things holding them back was information asymmetry and implicit non-compete between AI/ML system owner and content producer. Those are now gone/eroded.



Data Leverage: A Framework for Empowering the Public in its Relationship with Technology Companies

Nicholas Vincent
Northwestern University
nickvincent@u.northwestern.edu

Hanlin Li
Northwestern University
lihanlin@u.northwestern.edu

Nicole Tilly
Northwestern University
nicoletilly2023@u.northwestern.edu

Stevie Chancellor*
University of Minnesota
steviec@umn.edu

Brent Hecht
Northwestern University
bhecht@northwestern.edu

ABSTRACT

Many powerful computing technologies rely on implicit and explicit data contributions from the public. This dependency suggests a potential source of leverage for the public in its relationship with technology companies: by reducing, stopping, redirecting, or otherwise manipulating data contributions, the public can reduce the effectiveness of many lucrative technologies. In this paper, we synthesize emerging research that seeks to better understand and help people action this *data leverage*. Drawing on prior work in areas including machine learning, human-computer interaction, and fairness and accountability in computing, we present a framework for understanding data leverage that highlights new opportunities to change technology company behavior related to privacy, economic inequality, content moderation and other areas of societal concern. Our framework also points towards ways that policymakers can bolster data leverage as a means of changing the balance of power between the public and tech companies.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing theory, concepts and paradigms.**

KEYWORDS

data leverage, data strikes, data poisoning, conscious data contribution

ACM Reference Format:

Nicholas Vincent, Hanlin Li, Nicole Tilly, Stevie Chancellor, and Brent Hecht. 2021. Data Leverage: A Framework for Empowering the Public in its Relationship with Technology Companies. In *Conference on Fairness, Accountability, and Transparency (FAccT '21)*, March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 13 pages. <https://doi.org/10.1145/3442188.3445885>

*Chancellor completed much of this work while at Northwestern University.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

FAccT '21, March 3–10, 2021, Virtual Event
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8309-7/21/03...\$15.00
<https://doi.org/10.1145/3442188.3445885>

Vincent et al. *FAccT 2021*

The Dimensions of Data Labor: A Road Map for Researchers, Activists, and Policymakers to Empower Data Producers

Hanlin Li
hanlinl@berkeley.edu
University of California, Berkeley
Berkeley, CA, USA

Stevie Chancellor
steviec@umn.edu
University of Minnesota
Minneapolis, MN, USA

Nicholas Vincent
nickvincent@u.northwestern.edu
University of California, Davis
Davis, CA, USA

Brent Hecht
bhecht@northwestern.edu
Northwestern University
Evanston, IL, USA

ABSTRACT

Many recent technological advances (e.g. ChatGPT and search engines) are possible only because of massive amounts of user-generated data produced through user interactions with computing systems or scraped from the web (e.g. behavior logs, user-generated content, and artwork). However, data producers have little say in what data is captured, how it is used, or who it benefits. Organizations with the ability to access and process this data, e.g. OpenAI and Google, possess immense power in shaping the technology landscape. By synthesizing related literature that reconceptualizes the production of data for computing as “data labor”, we outline opportunities for researchers, policymakers, and activists to empower data producers in their relationship with tech companies, e.g. advocating for transparency about data reuse, creating feedback channels between data producers and companies, and potentially developing mechanisms to share data’s revenue more broadly. In doing so, we characterize data labor with six important dimensions - legibility, end-use awareness, collaboration requirement, openness, replaceability, and livelihood overlap - based on the parallels between data labor and various other types of labor in the computing literature.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models.**

KEYWORDS

user-generated data, empowerment, data leverage

ACM Reference Format:

Hanlin Li, Nicholas Vincent, Stevie Chancellor, and Brent Hecht. 2023. The Dimensions of Data Labor: A Road Map for Researchers, Activists, and Policymakers to Empower Data Producers. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, June 12–15, 2023, Preprint for FAccT '23, June 12–15, 2023, Chicago, IL, USA

<https://doi.org/10.1145/3593013.3594070>

<https://doi.org/10.1145/3593013.3594070>

Chicago, IL, USA. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3593013.3594070>

1 INTRODUCTION

Technology users generate large troves of data in their daily interactions with computing systems, e.g. behavior logs, content, and personal information. Currently, this data primarily benefits just a small set of technology organizations that are equipped with the means and resources to collect, process, and model data at scale for their own benefits (e.g. insights, models, sales of services and advertisements). For example, publicly available texts and artwork enabled the creation of generative AI models like ChatGPT and Dall-E because model developers were able to scrape and process data from billions of web pages¹. Conversely, data producers like artists, writers, and users have little to no power in deciding how their data is used or who it benefits [4, 7, 42, 63]. This power imbalance between data producers and technology operators has manifested in public outcries about industry practices in the tech sector. For example, emerging generative AI models such as Stable Diffusion, Dall-E, and GitHub Copilot have sparked extensive criticism among artists and programmers because of these models’ unapproved reuse of their work and implications on future employment opportunities [79–81]. More broadly, social media users have long protested the monetization of user data and the corporate surveillance practices that tend to go with it [45].

Given data producers’ lack of power over the data they generate, researchers, policymakers, and activists have advocated for a new producer-oriented paradigm shift to increase the voice of the data-generating public – understanding data generation as a form of labor, or “data labor” [7]. Supporters of this approach have argued that treating data as an outcome of social labor instead of “exhaust” will pave the way for more broadly distributing the power and benefits of data [86], and scholars have addressed what this may look like in practice. Initial (yet abstract) proposals include supporting “data unions” [63] or “mediators of individual data” [42] that negotiate data use terms with technology firms on behalf of their data-producing “union” members [63], drafting legislation that would grant users greater control over the data they produce [1, 76], and creating tools to support user-driven collective action [20, 86].

¹<https://commoncrawl.org/2022/10/sep-oct-2022-crawl-archive-now-available/>

Li et al. *FAccT 2023*

PSA Research Group
People, Space, and Algorithms

Don't give OpenAI all the credit for GPT-3: You might have helped create the latest "astonishing" advance in AI too

By Nicholas Vincent on 2020-09-22

The much-celebrated GPT-3 that can answer questions, write poems, and more wouldn't be possible without content written by millions of people around the world. Shouldn't they get some credit?

You may have heard about OpenAI's "GPT-3"— an "astonishing" machine learning system that can produce impressive poems, code and op-eds. However, it wasn't just OpenAI that built this: it was also millions of people writing, posting, and voting on content. In fact, it's possible that you played a role in creating GPT3!

Indeed, content and data you helped generate may have been used to build AI systems like GPT-3 in the past. It's even more likely your data will be used to build AI systems in the future. Do you deserve some of the credit for AI's success? What about the profits (for instance, through a "data dividend")? What might you do if you aren't

PSA Research Group
People, Space, and Algorithms

GitHub Copilot and the Exploitation of "Data Labor": A Wake-Up Call for the Tech Industry

By psagrp on 2021-07-08

This post was jointly written by the PSA Research Group. Points of contact: [Nick Vincent](#) and [Hanlin Li](#).

The release of GitHub Copilot is causing quite the outcry in the tech industry, and for good reason. Copilot is a big leap forward in the effort to automate programming, and perhaps eventually programming jobs. However, we in the tech industry should realize that Copilot is merely a taste of our own medicine; we've been creating similar problems for people outside the tech industry for years.

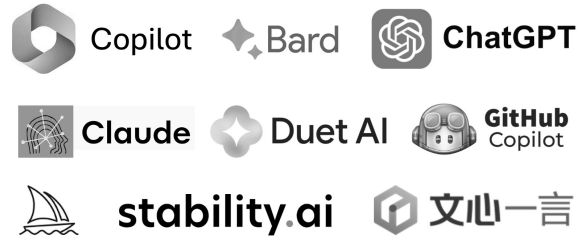
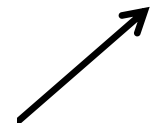
GitHub Copilot is an AI system that uses a statistical model learned from "billions of lines of public code" on GitHub. This code has been written by millions of programmers around the world, and many of those programmers are not happy. They argue that GitHub should

A brief outline

- **Looking back:** Wikipedia's central role in the development of modern AI
- **The present conundrum:** The dominant LLM paradigm threatens Wikipedia, large portions of the content ecosystem, and ultimately itself.
- **Looking forward:** What Wikipedia can do to help itself in the LLM era, and make the LLM era much better in the process

Two Grand Bargains that Make LLMs Possible

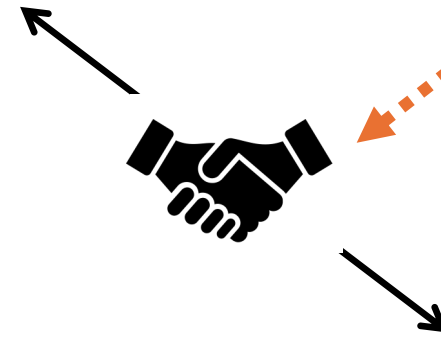
Very traditional bargain based on the exchange of services for money



Model Builders

(and builders of LLM-based applications)

Very unconventional bargain based around **information asymmetry** (“Legibility” + “End-use Awareness”) **non-compete** (“Livelihood Overlap”) (see [Li et al. 2023](#))



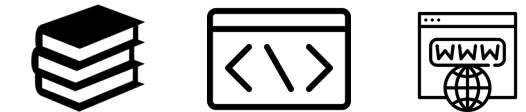
GPUs

Trad'l cloud stack

Energy

Cloud Infra Providers

(Hardware, Energy, etc)



Literature

Code

The web



User-generated content



Scientific papers



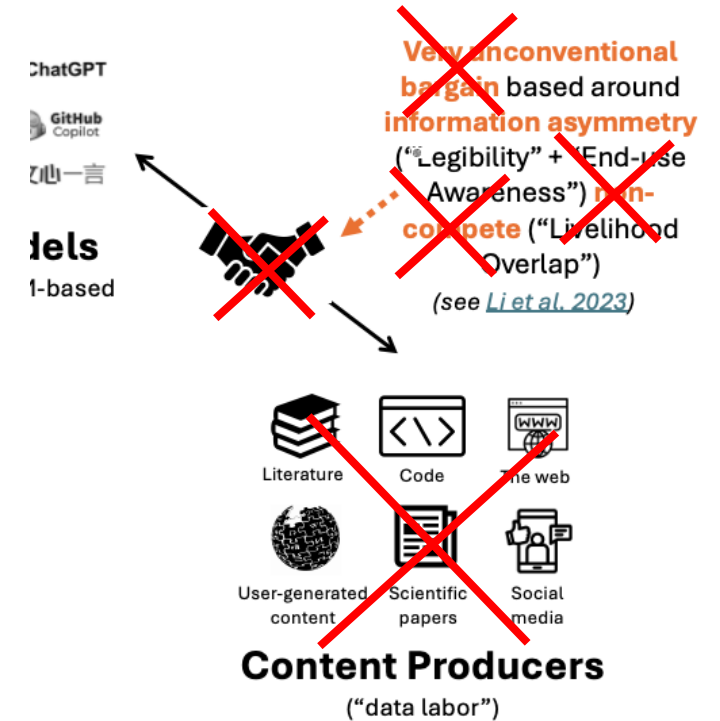
Social media

Content Producers

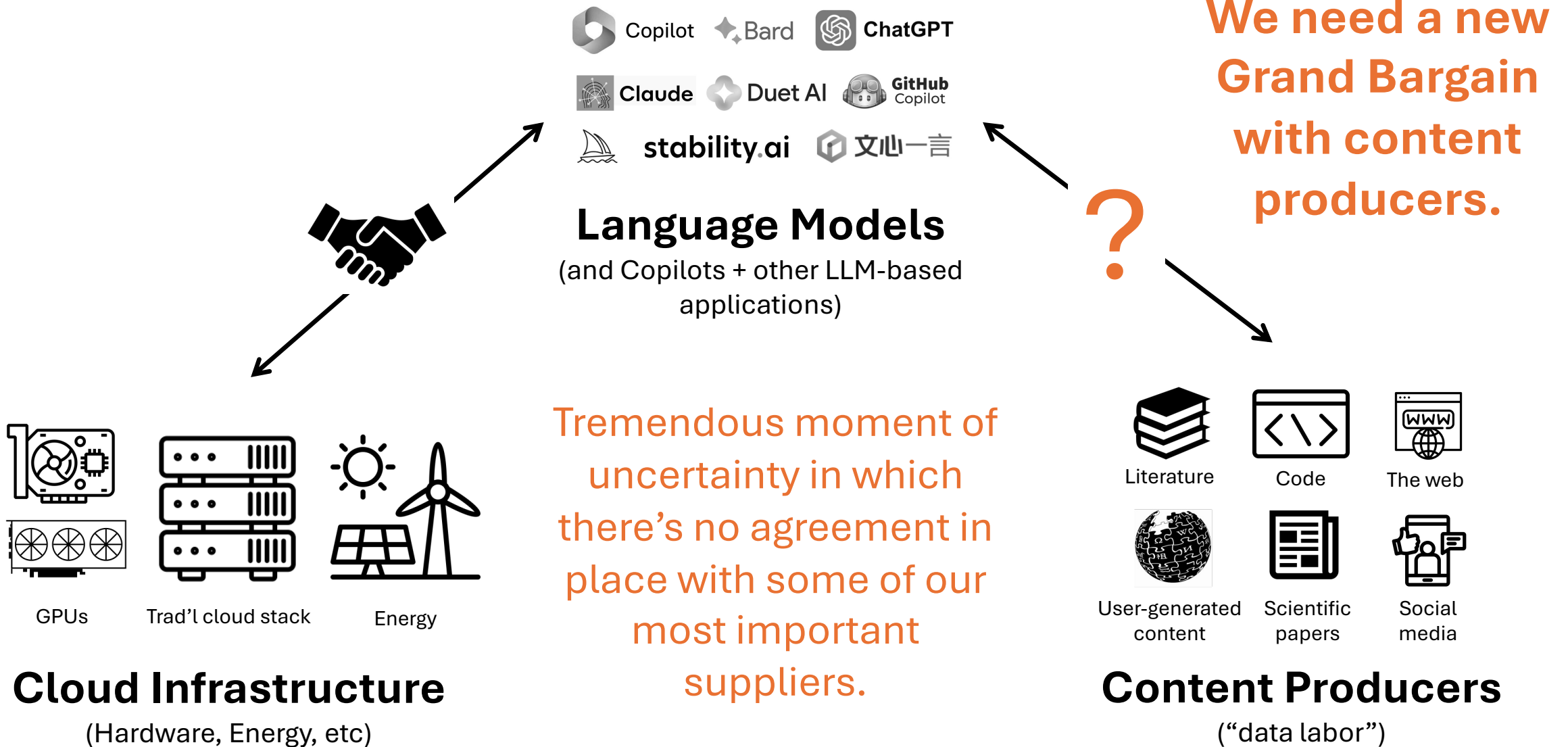
(“data labor”)

Flash forward to 2022...

- **Information asymmetry** heavily eroded by ChatGPT, GitHub Copilot
 - People know what we're doing and how we're doing it
- **Non-compete heavily** eroded by LLM-based businesses
 - Promise of LLMs is largely in the same information work domains from which they get their content (law, medicine, coding, science...)
 - Understandable given the nature of the technology: They naturally compete with their content supply chain (within-domain vs. outside domain performance)



We need a new Grand Bargain with content producers



Two ways current LLM paradigm is in trouble from lack of grand bargain

- **Capabilities:** AI significantly less competent than it could be
- **Societal impacts:** Potential harms very substantial

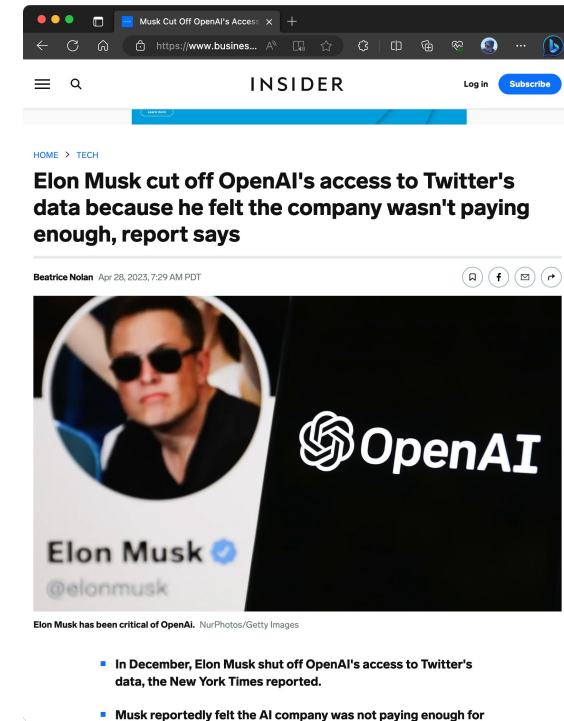
Misalignment of incentives leads to a “data strike era”



Journalism Data Strike



General Data Strike



Twitter/X Data Strike

Well-executed data strikes likely highly effective against LLMs

Eval data	<u>PD</u>	<u>PDSW</u>	<u>PDSWBY</u>	Pythia
FreeLaw	5.3	5.7	6.5	5.6
Gutenberg	15.2	12.5	14.1	13.1
HackerNews	38.0	13.7	14.5	13.3
Github	13.5	2.7	2.8	2.4
NIH ExPorter	28.2	19.2	15.0	11.1
PhilPapers	31.7	17.6	15.0	12.7
Wikipedia	28.9	20.3	11.3	9.1
CC News	34.0	23.3	21.2	12.0
BookCorpus2	25.3	19.2	19.6	13.2
Books3	27.2	19.3	18.6	12.6
OpenWebText2	37.8	21.1	18.8	11.5
Enron Emails	18.6	13.2	13.5	6.9
Amazon	81.1	34.8	37.0	22.9
MIMIC-III	22.3	19.0	15.5	13.1
Average	29.1	17.3	16.0	11.4

Significant drops in perplexity when forced to only train on data with explicit consent.

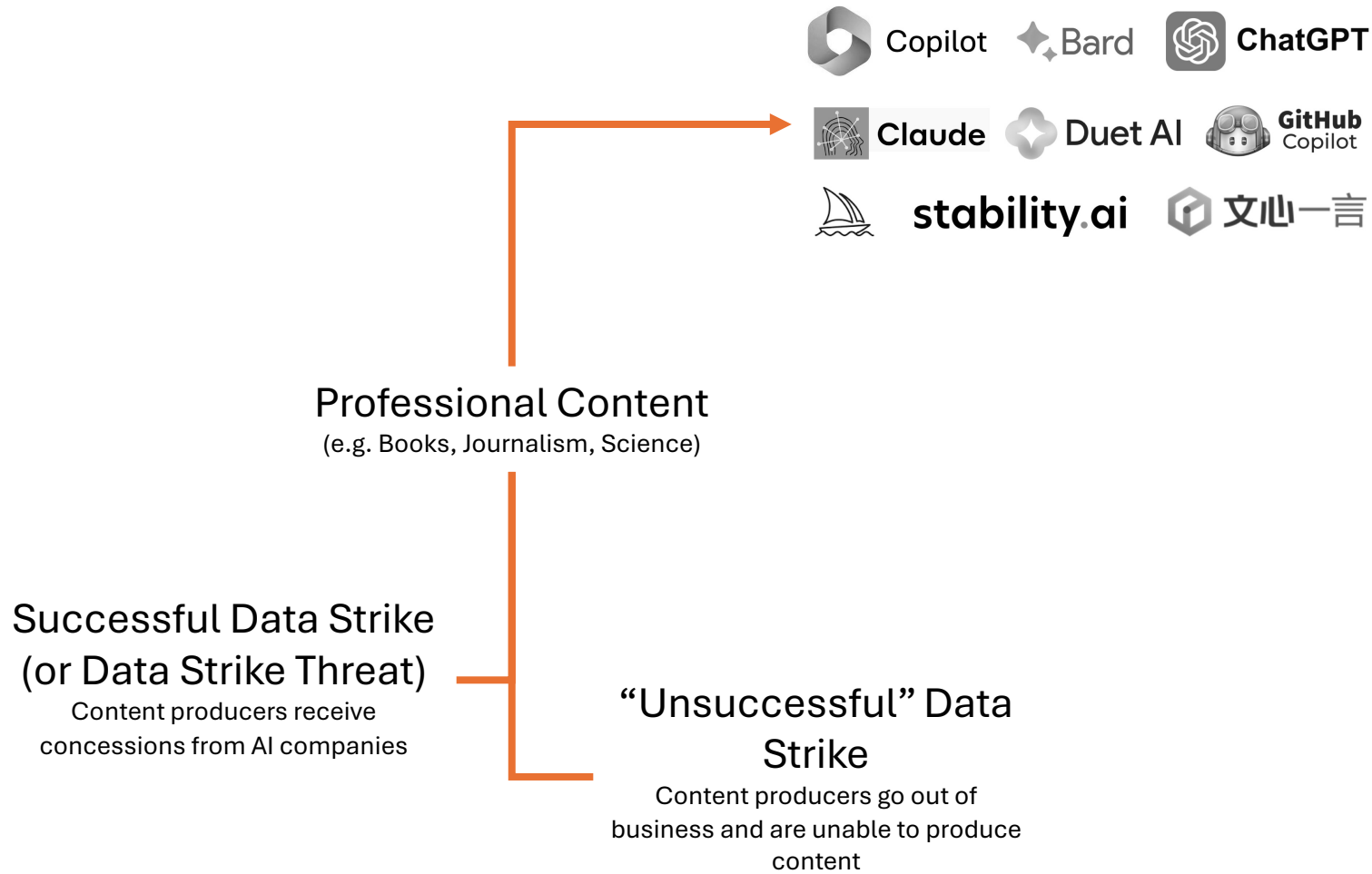
Suggests data strikes along professional domains would work well.

What do LLM data strikes look like?

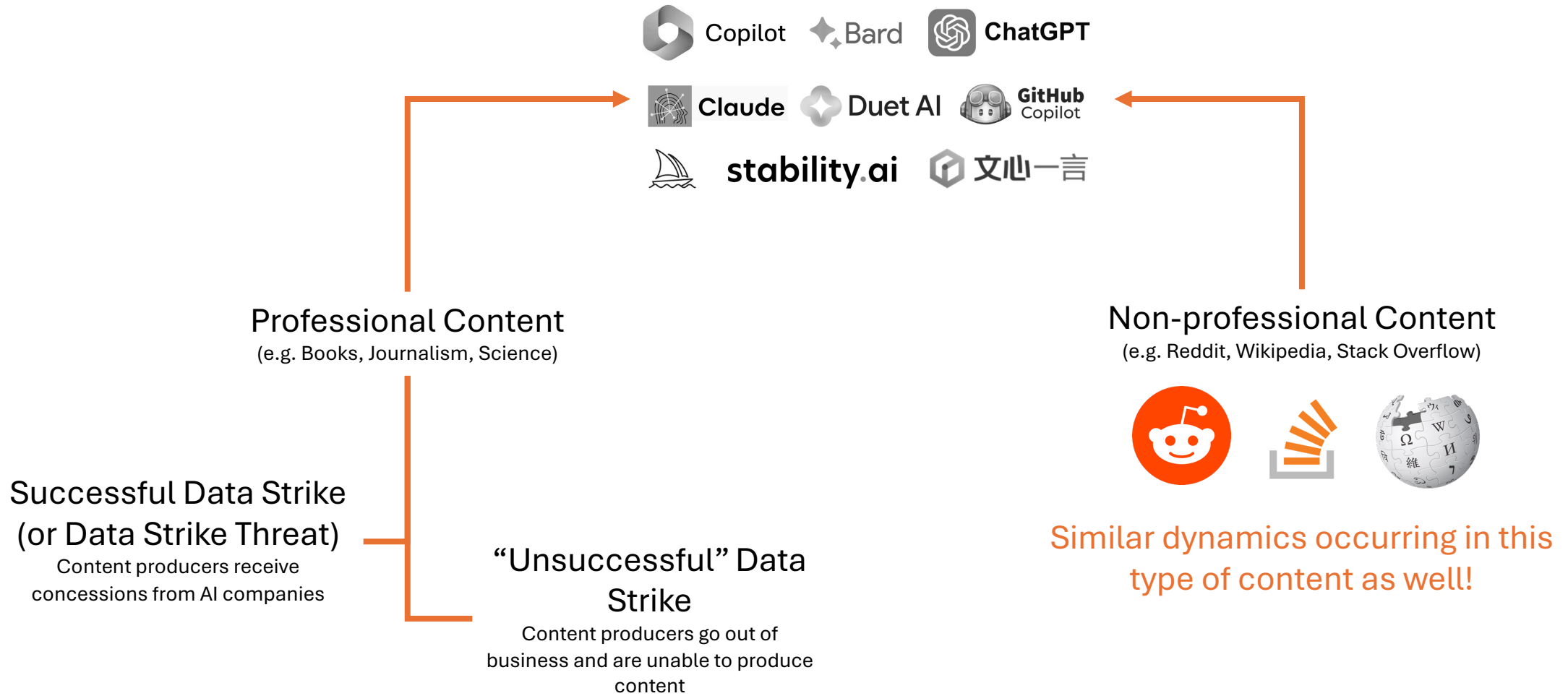
- Low effort: robots.txt, IP blocking
 - Easy to give up Common Crawl and Bing traffic in particular
- Medium effort: Putting everything behind a paywall or registration wall
- Higher effort: Work stoppages, class action lawsuits
- **Highest effort: Going out of business**



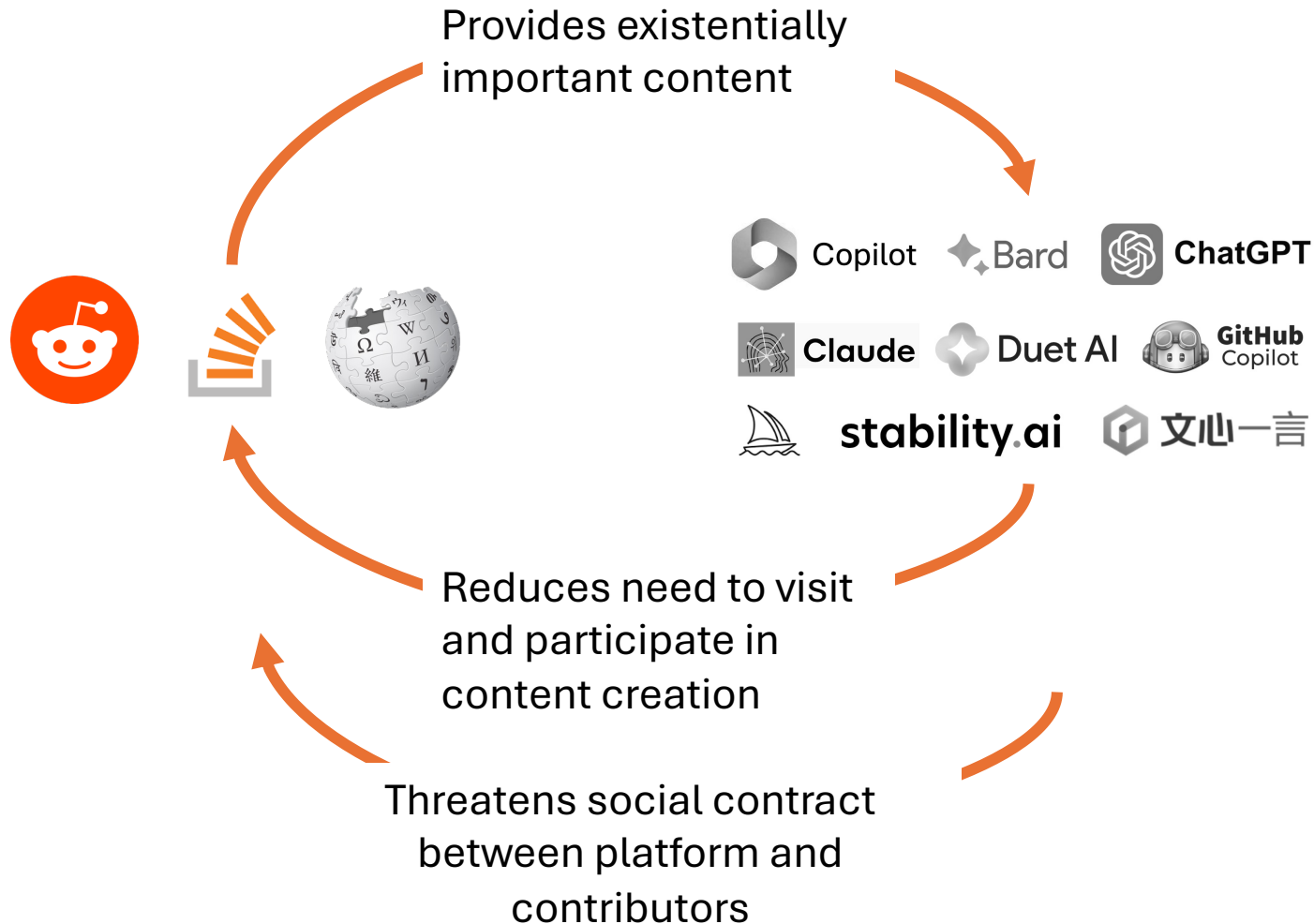
Are data strikes guaranteed?



Are data strikes guaranteed?



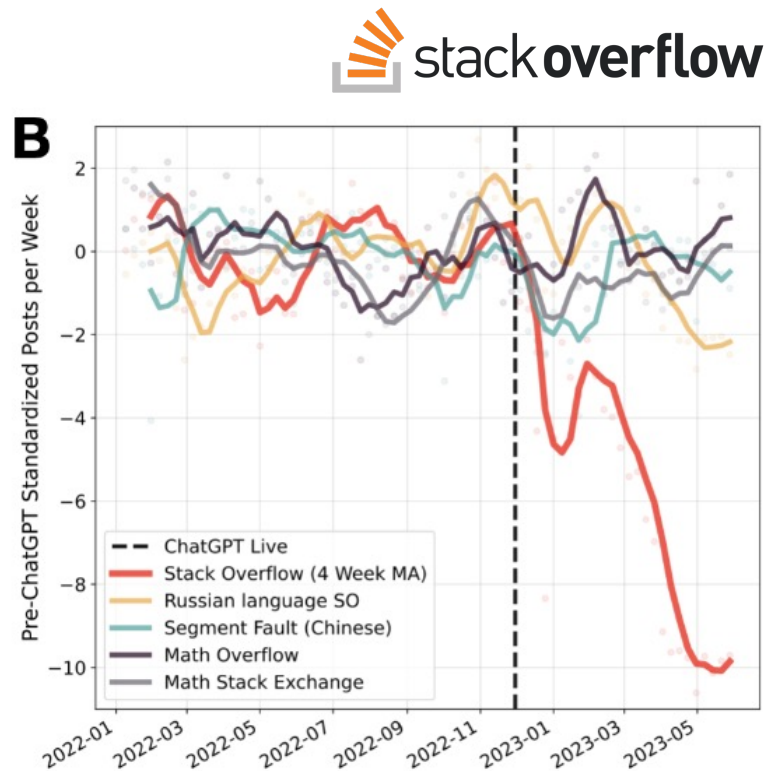
The hypothesized LLM-UGC “Doom Loop”



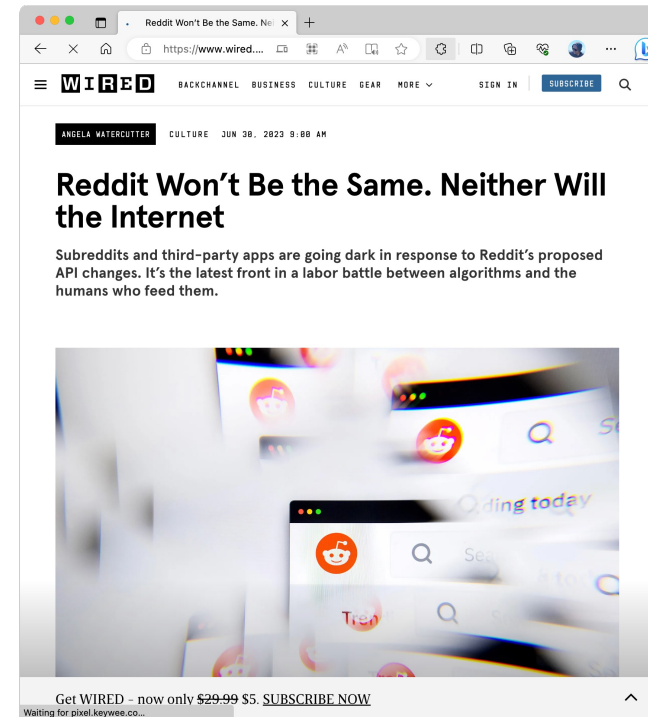
Aka the “Paradox of Reuse”

([Taraborelli 2015](#);
[McMahon et al. 2017](#);
[Vincent 2022](#))

Some evidence the doom loop has begun

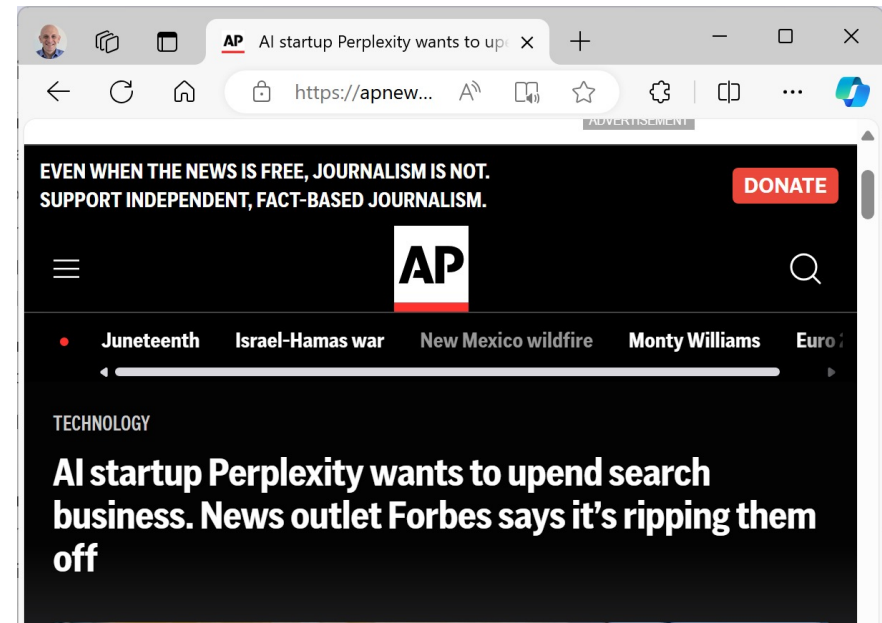
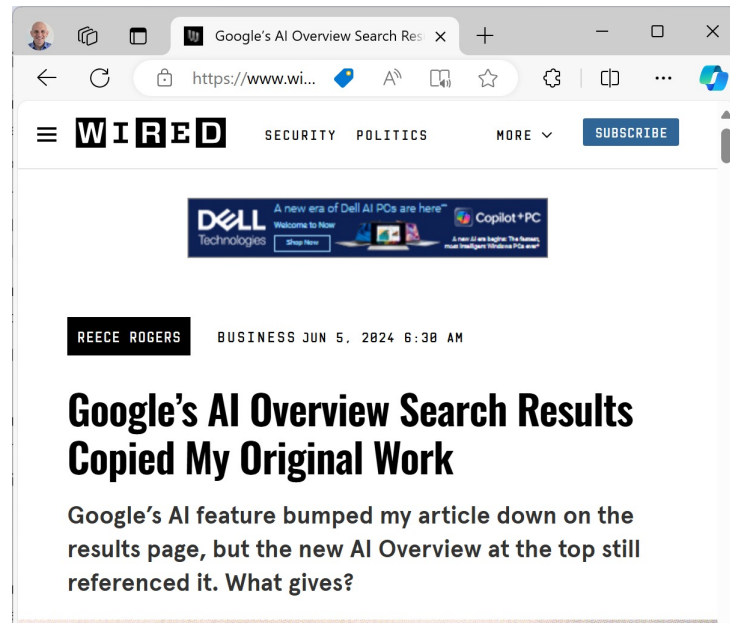
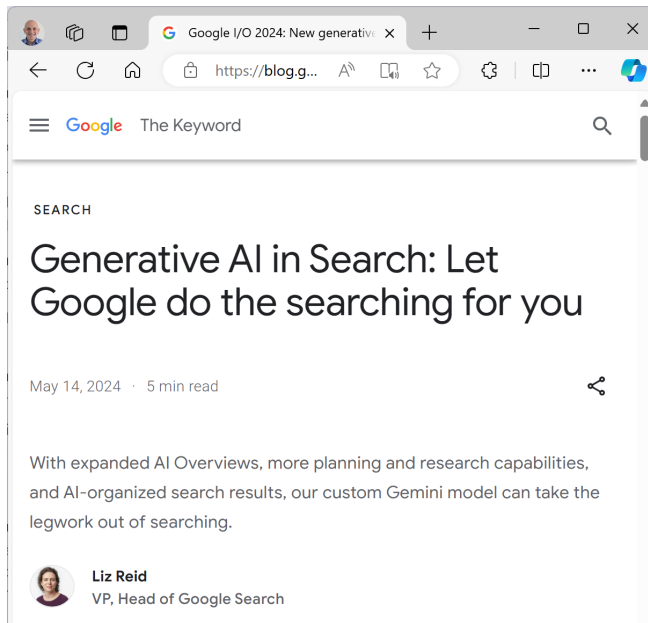


[Rio-Chanona et al. 2023](#)



Reddit moderator strike caused by LLM-induced change to Reddit's business model

Recent developments likely to accelerate “doom loop” trends



The “Paradox of Reuse” of Everything



The image shows a screenshot of a Google search interface. At the top left is the Google logo. To its right is a search bar containing the text "average goat lifespan". Below the search bar are navigation tabs: "Web" (highlighted with a red underline), "Shopping", "Images", "Videos", "News", "More" (with a dropdown arrow), and "Search tools". Below the tabs, it says "About 367,000 results (0.39 seconds)". The main search result is displayed in a white box with a light gray border. On the left side of this box, the text "15 – 18 y" is shown in a large font, with "Goat, Lifespan" in a smaller font below it. On the right side of the box is a photograph of a brown and white goat standing on a wooden stump in a grassy field. Below the photograph, the word "Feedback" is written in a small, italicized font.

Google

average goat lifespan

Web Shopping Images Videos News More Search tools

About 367,000 results (0.39 seconds)

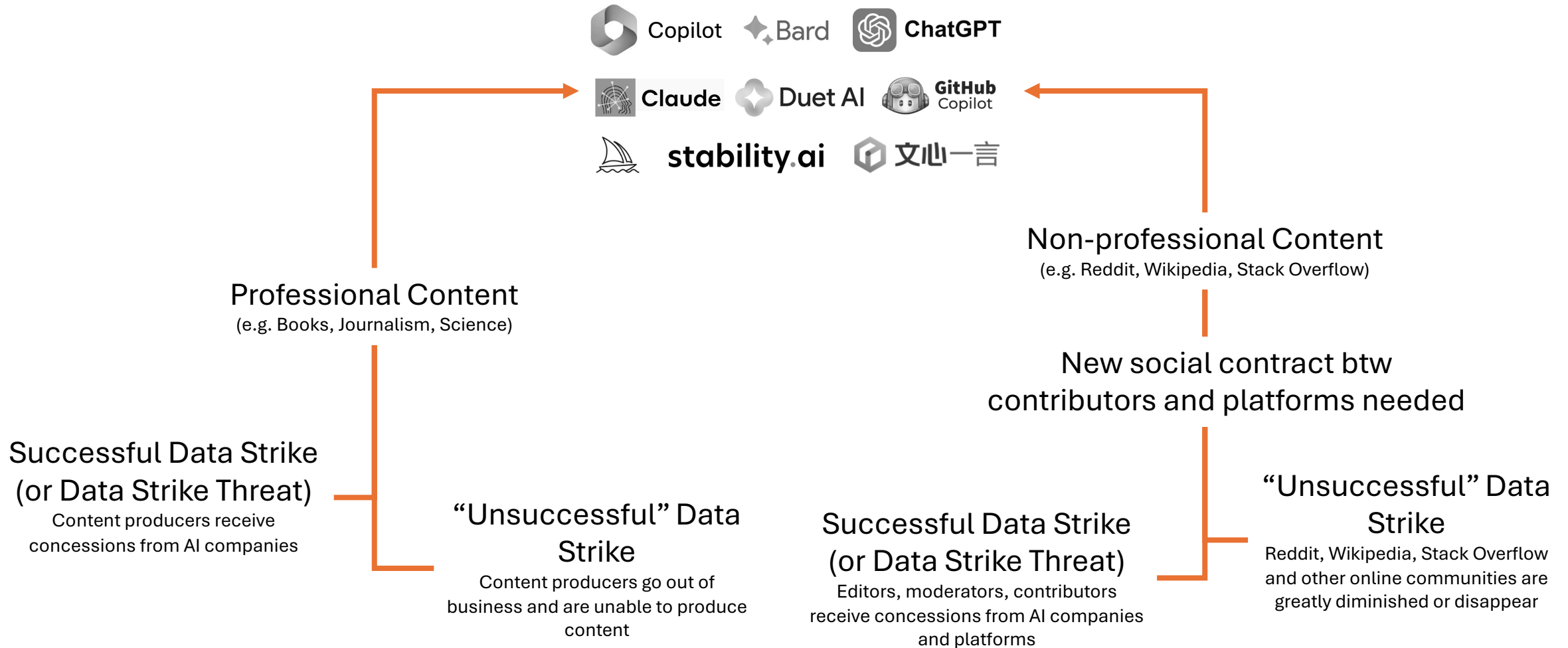
15 – 18 y

Goat, Lifespan

Feedback

(Taraborelli 2015)

Are data strikes guaranteed?

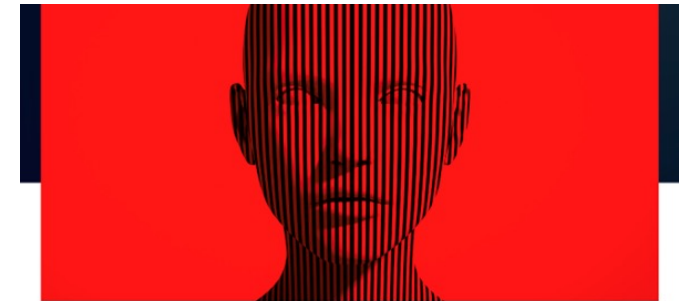


Two ways current LLM paradigm is in trouble from lack of grand bargain

- **Capabilities:** AI significantly less competent than it could be
- **Societal impacts:** Potential harms very substantial

Risk of large-scale labor disruptions

- More complex than commonly understood, can be overblown
- Critical to leverage history of technology and productivity
- There is unmet demand in so many places that can absorb productivity increases
- What AI community works on matters; are we building substitutional or augmentative technologies? (“Turing Trap”)
- But elephant in the room: Goal of some AI companies vs. current capabilities
 - Effectively will create infinite productivity increases that no amount of increased demand can absorb



Insights

The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence



Erik Brynjolfsson
Director
Stanford Digital Economy Lab

January 12, 2022
20-min read



Listen to this content

10:00 / 20:13

In 1950, Alan Turing proposed an “imitation game” as the ultimate test of whether a machine was intelligent: could a machine imitate a human so well that its answers to questions are indistinguishable from those of a human.¹ Ever since, creating intelligence that matches human intelligence has implicitly or explicitly been the goal of thousands of researchers, engineers, and entrepreneurs.

OpenAI’s mission is to ensure that artificial general intelligence (AGI)—by which we mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity. We will attempt to directly build safe and beneficial AGI, but will also consider our mission fulfilled if our work aids others to achieve this outcome. To that end, we commit to the following principles:

Significant drop in “labor share of revenue”?

- “The fraction of economic output that accrues to workers as compensation in exchange for their labor” ([US BLS 2017](#))
- Once thought to be relatively stable, has been declining in recent decades, at least in part due to computing
 - 1% increase in IT intensity is associated with a 0.1% decline in labor share of revenue ([Brynjolfsson et al. 2023](#))
- Reasonable argument is this is because we’re just not paying most of our labor, e.g. done by our quality and relevance assessors, reviewers, etc.
 - Reddit moderators save Reddit *at least* 3% of revenue per year ([Li et al. /CWSSM 2022](#))
- Serious risk of massive reduction in labor share of revenue if current LLM paradigm continues

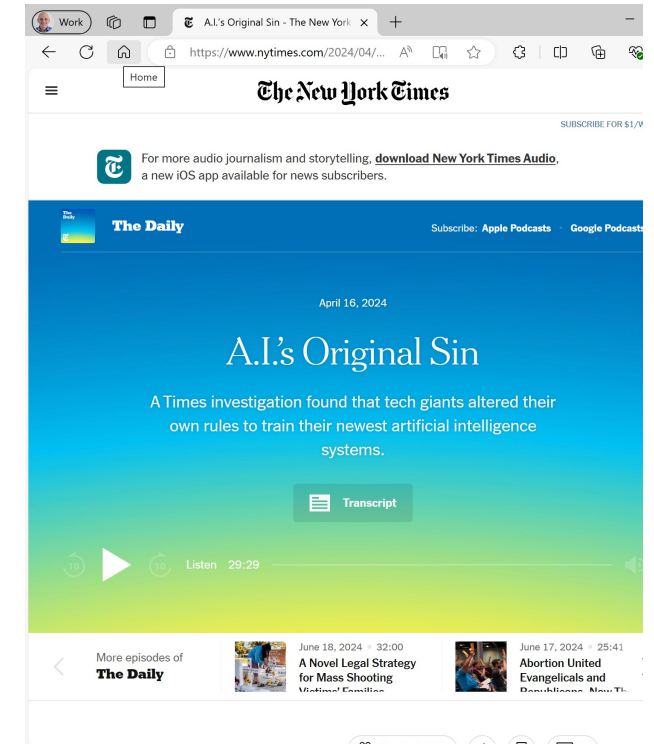
Effects of computing-induced wealth and power concentration

- Excessive wealth concentration widely recognized as non-optimal for many societal goals that have wide consensus
- **Economic productivity** e.g. GDP (which is also really important to Microsoft's business!)
- **Failure of democratic institutions** – “We may have democracy, or we may have wealth concentrated in the hands of a few, but we can't have both.” (Louis Brandeis)
- **Obsolescence of most ethical frameworks:** If all humans have equal value, why should a few have 10^9 more resources?

The screenshot shows a web browser displaying a TED Ideas article. The article title is "The 4 biggest reasons why inequality is bad for society" under the "BUSINESS" category, dated June 3, 2014. An embedded video player is visible, showing the title "The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence" by Erik Brynjolfsson, Director of the Stanford Digital Economy Lab. The video player includes a "Listen to this content" button and a progress bar. Below the video, there is a "WHAT'S TO E" section with a quote: "It's safe to philosoph the problem". The article text discusses the Turing Test and the benefits of human-like artificial intelligence (HLAI), mentioning that in 1950, Alan Turing proposed an "imitation game" as the ultimate test of whether a machine was intelligent. The text also notes that HLAIs can lead to a trap where machines become better substitutes for human labor, leading to a loss of economic and political bargaining power for workers. The article concludes that HLAIs can be enormously beneficial, but there are currently excess incentives for automation rather than augmentation among technologists, business executives, and policymakers.

Risk of legitimacy crisis in AI

- What happens if the dominant technology has such a wide mismatch between value accretion and value creation?
 - Can the market still work? Will people still believe in it?
- Property dispossession perceived as an irremovable “origin sin” (*NY Times*) without changes?
- History of scaled property dispossession is terrifying, including those justified by technological advancement
 - English Enclosures a strong analogy
 - Led to hundreds of years of riots, and likely (in part) the beheading of King Charles I.
 - Diggers and Levellers, revisited -> data poisoning, data center attacks?



Demographic dimensions to property dispossession

- New paper at CHI: “American Jews May Be Disproportionately Harmed by Intellectual Property Dispossession in LLM Training”
- *“Systemic property dispossession from minority groups has often been carried out in the name of technological progress. In this paper, we identify evidence that the current paradigm of LLMs likely continues this long history.”*
- Parallels historical property dispossession because of causal relationship to history; reason why certain demographics invested more in intellectual property

A Canary in the AI Coal Mine: American Jews May Be Disproportionately Harmed by Intellectual Property Dispossession in Large Language Model Training

Heila Precel
heila@bu.edu
Boston University
Boston, United States

Allison McDonald
amcdon@bu.edu
Boston University
Boston, United States

Brent Hecht
bhecht@northwestern.edu
Northwestern University
Evanston, United States

Nicholas Vincent
nicholas_vincent@sfu.ca
Simon Fraser University
Burnaby, Canada

ABSTRACT

Systemic property dispossession from minority groups has often been carried out in the name of technological progress. In this paper, we identify evidence that the current paradigm of large language models (LLMs) likely continues this long history. Examining common LLM training datasets, we find that a disproportionate amount of content authored by Jewish Americans is used for training without their consent. The degree of over-representation ranges from around 2x to around 6.5x. Given that LLMs may substitute for the paid labor of those who produced their training data, they have the potential to cause even more substantial and disproportionate economic harm to Jewish Americans in the coming years. This paper focuses on Jewish Americans as a case study, but it is probable that other minority communities (e.g., Asian Americans, Hindu Americans) may be similarly affected and, most importantly, the results should likely be interpreted as a “canary in the coal mine”

Training. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 17 pages. <https://doi.org/10.1145/3613904.3642749>

1 INTRODUCTION

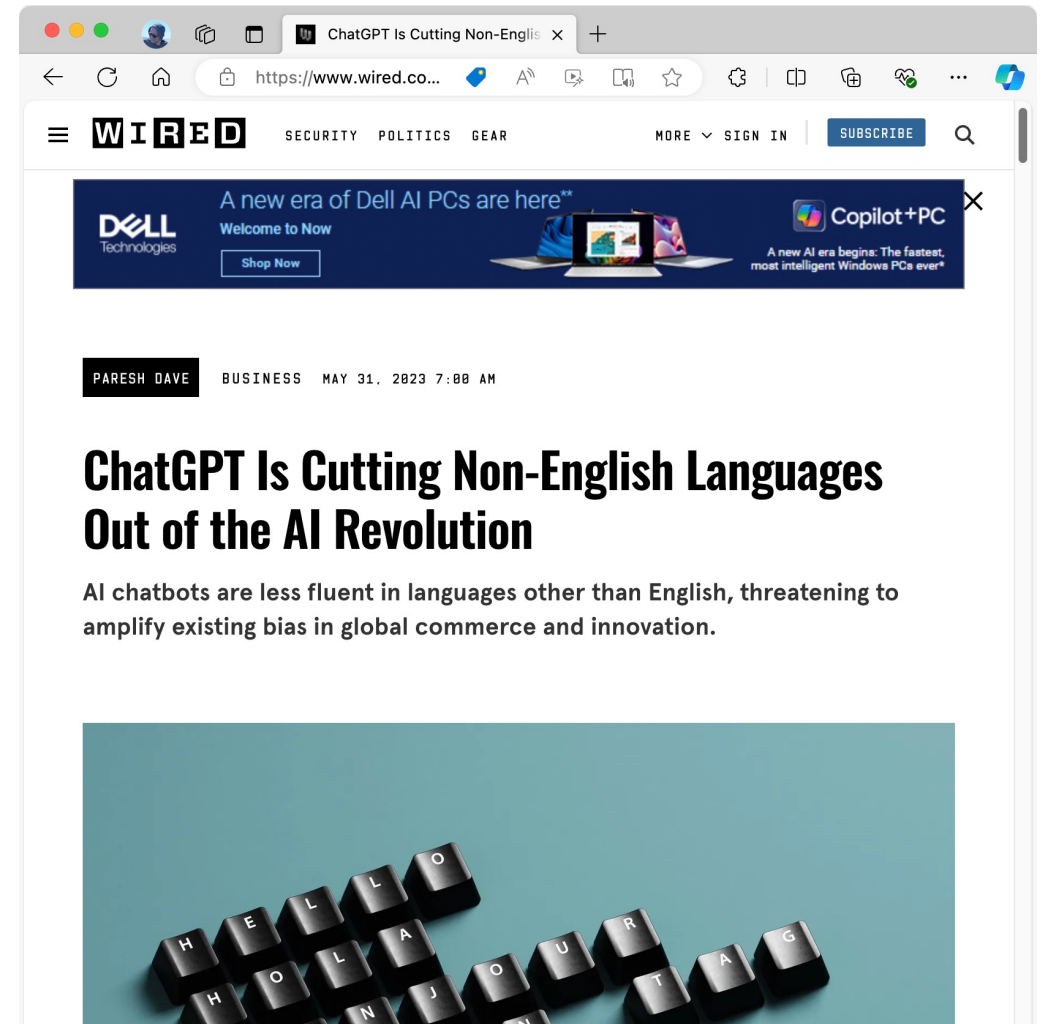
One of the most prominent critiques of large language models (LLMs) is that they train on massive amounts of content without the consent of the authors of that content [8, 59, 73, 91, 98, 103, 109, 124–126]. This concern is exacerbated by one of the core promises of LLMs: their ability to use patterns in their training data to substitute for the paid labor of those who created said data. People in a wide range of professions (e.g., fiction-writing and journalism) are now accessing language modeling companies of not only stealing their content (e.g., novels and news stories), but also of using this very content to put them out of a job (e.g., [58, 98, 105, 116]). Indeed,

Dataset	% IP with DJN author	% IP with U.S. Jewish author	% Expected IP with U.S. Jewish author	Relative Dispossession Magnitude
PubMed Central	0.19	1.39-1.91	0.5-0.7	2.02-3.71 X
Books3	0.98	7.01-9.64	1.8-2.4	2.92-5.36 X
ArXiv	0.28	2.01-2.77	0.5-0.7	3.07-5.63 X
GitHub	0.29	2.08-2.86	0.4-0.6	3.53-6.46 X
FreeLaw	0.93	6.65-9.14	1.8-2.4	2.77-5.08 X
Total	0.54	3.83-5.26	1.0-1.3	2.86-5.25 X
Weighted Total	0.37	2.63-3.61	0.8-1.0	2.46-4.51 X

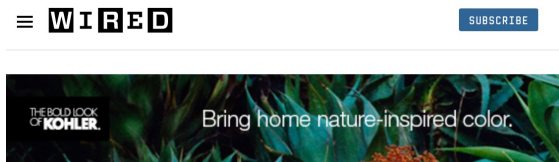
[Precel et al. 2024](#)

“Data colonialism” as a real risk for LLM content creation efforts

- Need to make sure strong grand bargain is in place before sourcing/creating content for LLMs in low-resource languages
- Else there’s a risk of replicating very problematic value transfers



Opportunity: “They took my stuff” to “We went to the moon



NICK VINCENT HANLIN LI IDEAS JAN 20, 2023 9:08 AM

ChatGPT Stole Your Work. So What Are You Going to Do?

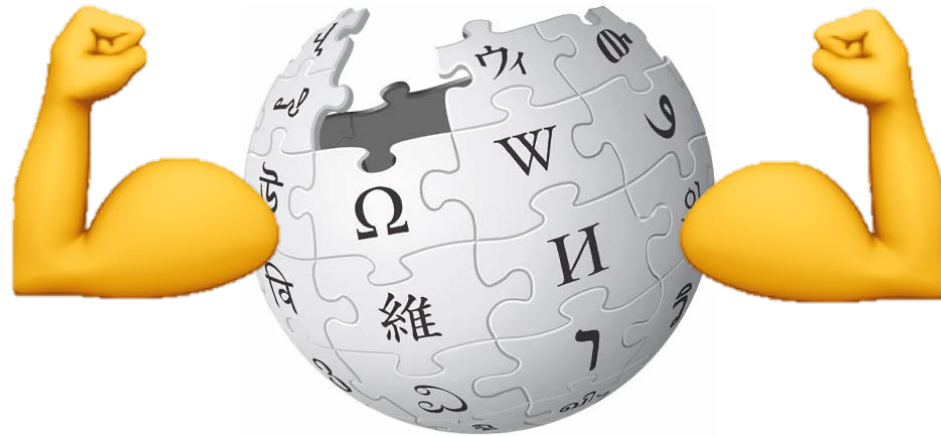
Creators need to pressure the courts, the market, and regulators before it's too late.



How do we go from this...

...to this?

What should the Wikipedia community do in this moment?



First proposition: Understand it
has a good deal of leverage

*Maximized if acting as a community, partnering with other
open data communities*

Successes in the past!


The screenshot shows a web browser displaying a TechCrunch article. The browser's address bar shows the URL: <https://techcrunch.com/2022/06/29/google-and-the-internet-archive-are-t...>. The article is titled "Google and the Internet Archive are the first customers to gain commercial access to Wikipedia content" and is categorized under "Enterprise". The author is Sarah Perez, and the article was published on June 29, 2022, at 9:27 AM PDT. The main image is a globe made of puzzle pieces with various characters and symbols. To the right of the article is a sidebar with social media sharing options (Facebook, LinkedIn, Reddit, Email, Print) and a sponsored content section for Andela. The Andela ad features a woman and the text "Talent pool a little dry? Borderless hiring gets the job done." Below the article is a banner for "DISRUPT Series A to B Startup? Join the ScaleUp Program At Disrupt 2024." with a "LEARN MORE" button.

TC
Login
Search
Startups
Venture
Apple
Security
AI
Apps
Events
Startup Battlefield
More

Enterprise


Google and the Internet Archive are the first customers to gain commercial access to Wikipedia content

Sarah Perez / 9:27 AM PDT • June 29, 2022 [Comment](#)



Host A Side Event At Disrupt 2024
Raise brand awareness & reach 10,000 tech leaders
[LEARN MORE](#)

Sponsored Content



Talent pool a little dry? Borderless hiring gets the job done.
Sponsored by Andela

DISRUPT | Series A to B Startup? Join the ScaleUp Program At Disrupt 2024. [LEARN MORE](#)

How: The four sources of content leverage

- Legal leverage
- Policy leverage
- Reputational leverage
- “Data leverage”
 - “Conscious Data Contribution”
 - “Data Strikes”

Legal leverage: Some interesting legal research directions

- Naturally cannot comment on any active lawsuits
- Copyright law is only a tiny part of the picture in the literature
 - Labor law, contract law, unjust enrichment, privacy law, huge number of jurisdictions (countries, first-order admins)
- LLMs produce content, so laws that reduce content protections can undermine value of LLMs



NEXSTAR MEDIA WIRE NEWS

Chick-fil-A in NC faces backlash for offering to pay 'volunteer' workers in chicken sandwiches

BY MICHAEL BARTIROMO, NEXSTAR MEDIA WIRE - 07/29/22 10:12 AM ET



News Home All News ScienceInsider News Features

HOME > NEWS > SCIENCEINSIDER > WHAT DOES THE HISTORIC SETTLEMENT WON BY HENRIETTA LACKS'S FAMILY MEAN FOR...

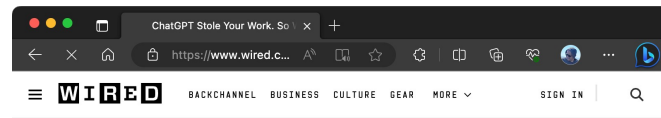
SCIENCEINSIDER | PEOPLE & EVENTS

What does the historic settlement won by Henrietta Lacks's family mean for others?

A legal expert says Thermo Fisher agreement could help some patients whose tissues were commercialized win redress, but they still face obstacles

7 AUG 2023 · 11:45 AM ET · BY MEREDITH WADMAN

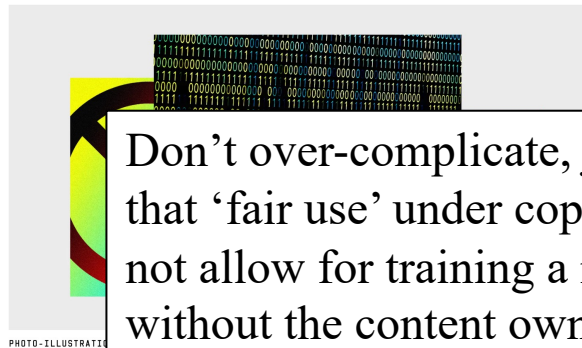
Policy leverage: Simple, realistic policy changes can reorder the LLM world overnight



NICK VINCENT HANLIN LI IDEAS JAN 20, 2023 9:00 AM

ChatGPT Stole Your Work. So What Are You Going to Do?

Creators need to pressure the courts, the market, and regulators before it's too late.



Don't over-complicate, just "[Clarify] that 'fair use' under copyright law does not allow for training a model on content without the content owner's consent, at least for commercial purposes."

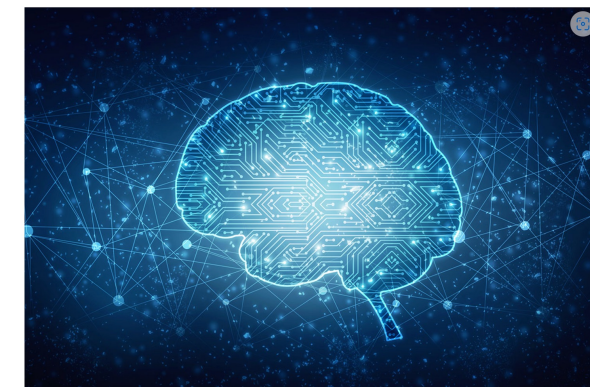


More than 10,000 Authors Sign Authors Guild Letter Calling on AI Industry Leaders to Protect Writers

Artificial Intelligence

July 18, 2023

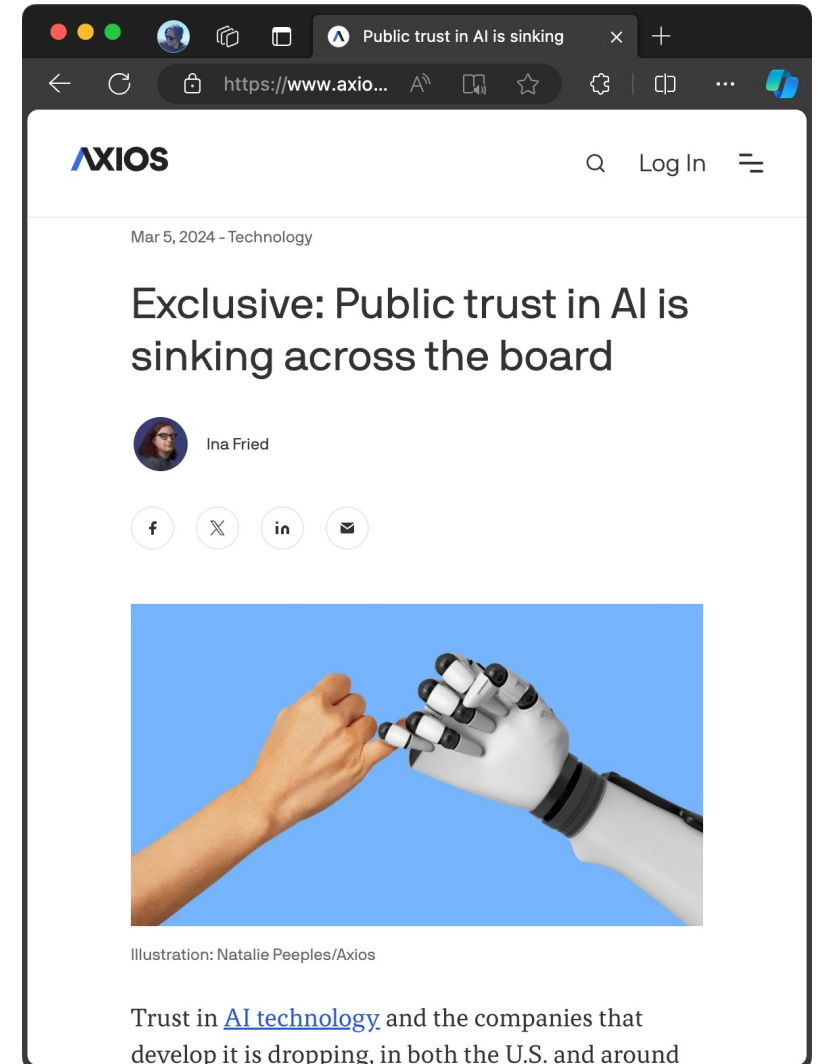
Share [Twitter](#) [Facebook](#) [LinkedIn](#) [Email](#)



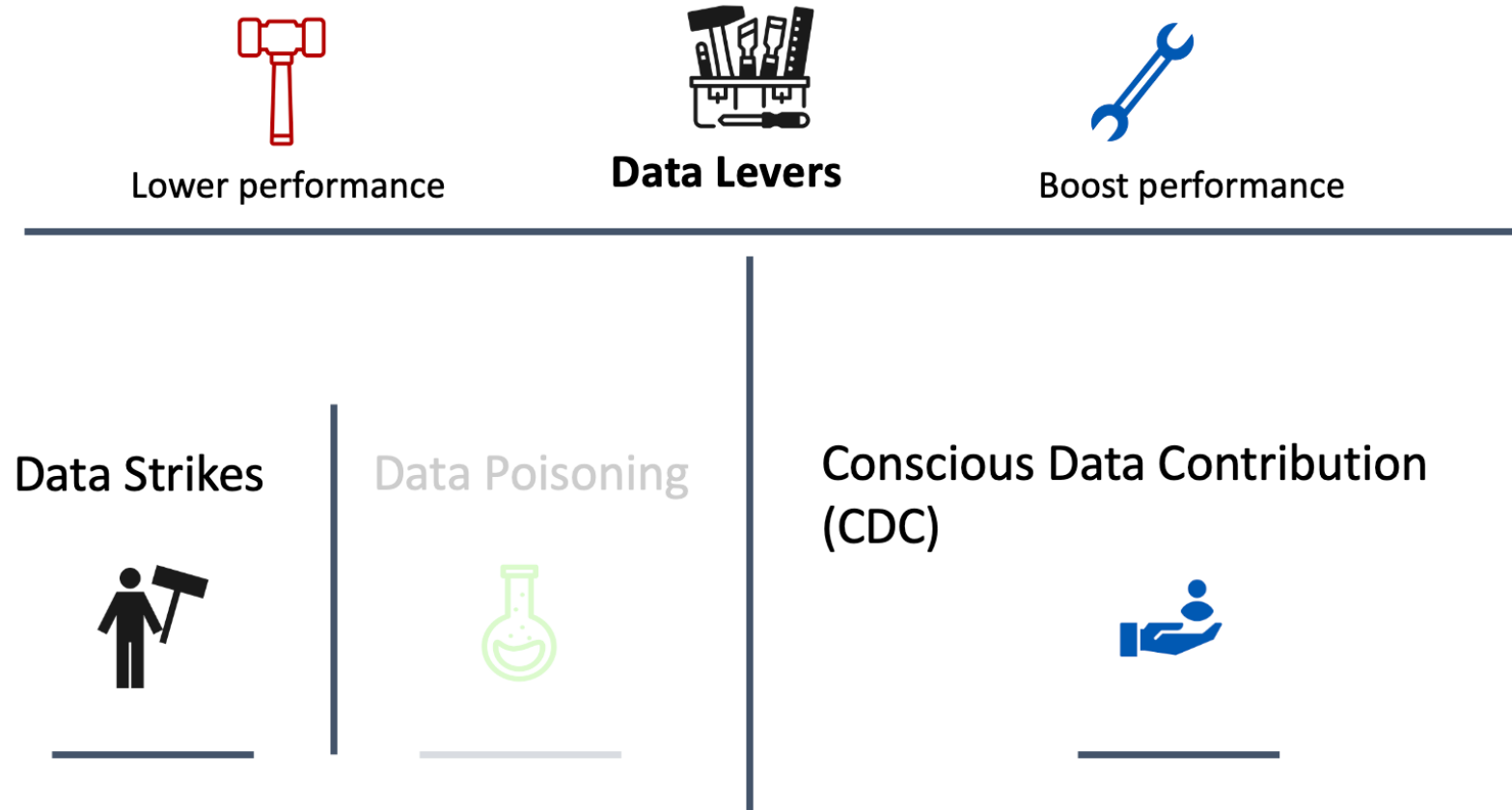
Updated July 19, 2023

Reputational leverage: Already having a significant effect

- Statements with legitimacy from trusted content institutions will matter
- Proposed solutions to bring trust up to higher levels will matter
- Lots of opportunity here for win-wins



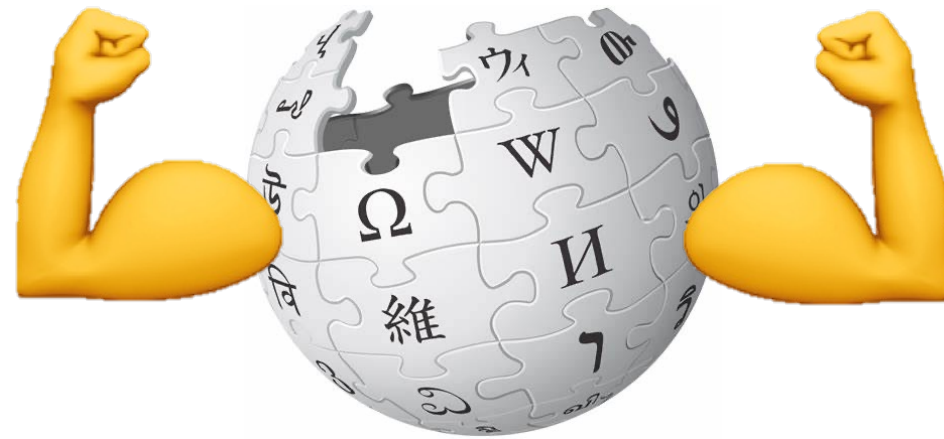
Data leverage: Strikes and CDC



The current LLM paradigm is not the only one

- Parametric uses of content
 - Collective bargaining with content producer unions
 - Sustaining content deals with large content providers
- Non-parametric uses of content (the future?)
 - Just-in-time content use (e.g. [Min et al. 2023](#))
 - Style markets ([Crawford 2022](#))
 - Provenance tracking ([Lanier 2023](#))
 - Per-use or per-time subscription models
 - Cloud platforms can implement “model + content buffet”-style offerings

Use leverage takes effort! (“no pain no gain”)



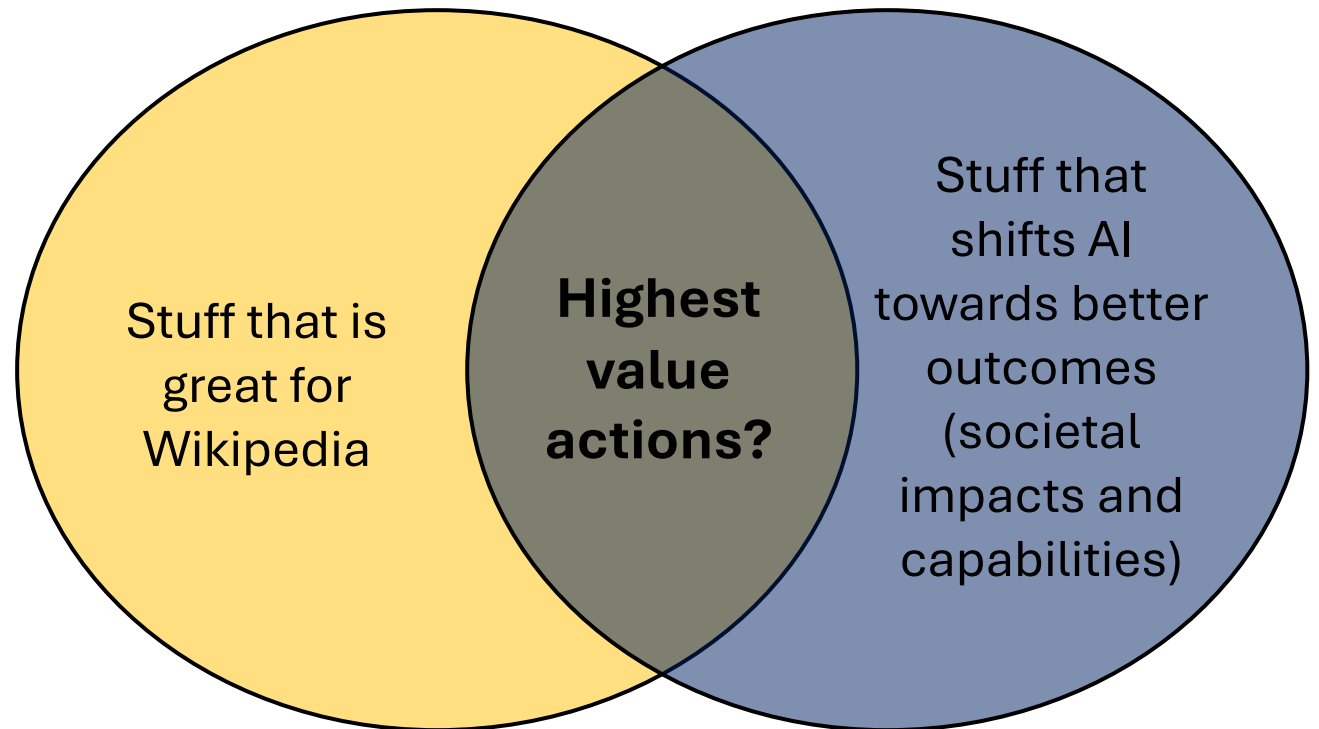
This logo didn't get these muscles overnight!

How: Values vs. tactics

- Some things that are values of the community look more like tactics through an AI lens, in particular “open”
 - LLMs should encourage us to ask what do we want to accomplish with “open”?
 - “Sum of all knowledge” → ensuring equal access to the benefits of open knowledge?
- What does “open” as a value vs. a tactic look like in the AI era?
 - Behavioral licenses like “Responsible AI Licenses” interesting to consider
 - Licensing likely only part of any solution here, and enforcement essential
 - Other types of leverage also needed

How should Wikipedia use its leverage?

- *For itself*
 - Ensure Wikipedia stays healthy in the AI era
 - Incentivize AI systems that use Wikipedia to reflect Wikipedia values
- *For others:* Wikipedia likely one of the best opportunities to shift the AI market towards better outcomes more generally



Wikipedia values can help AI thrive

- **Free and accessible repository of human knowledge:** Wikipedia faces similar challenges to all other open content producers
 - Solve it for Wikipedia, solve it for everyone?
- **No original research:** Wikipedia has strong stake in high-quality globally accessible information ecosystem
 - Need people to be incentivized to share knowledge in the open
- **Citation/verifiability:** Provenance a key step in passing value to content producers
 - Current LLM paradigm does not prioritize citation/provenance, but there are technical paths forward
- **What others?** For discussion!

A brief outline

- **Looking back:** Wikipedia's central role in the development of modern AI (case study from my career)
- **The present conundrum:** The dominant LLM paradigm threatens Wikipedia, large portions of the content ecosystem, and ultimately itself.
- **Looking forward:** What Wikipedia can do to help itself in the LLM era, and make the LLM era much better in the process

Thank you! Questions?

- Reach out at brent.Hecht@Microsoft.com or bhecht@northwestern.edu
- Links to papers available at my website: brenthecht.com
- Special thanks to three former students who led much of the work discussed: Isaac Johnson (WMF), Nick Vincent (Simon Fraser), Hanlin Li (U Texas)
- Deck available at: brenthecht.com/wikiworkshop2024