

Ethical and Societal Considerations

AI, ETHICS, AND SOCIETY



AI: Ethical and Societal Consideration

What considerations come to mind?

Ethical and Societal Considerations

Example Considerations:

- AI Safety
- AI Fairness (and incentives mismatch)
- AI Security (as a new attack surface and attack enabler)
- AI sustainability

Ethical/Societal Considerations in Safety

An Autonomy Perspective

AI Applications in Autonomy and Robotics

AI Application Examples in CPS/IoT: Last-mile Drone Delivery (~ Decade Old)

Drone deliveries will be a fast-growing application of AI and autonomy

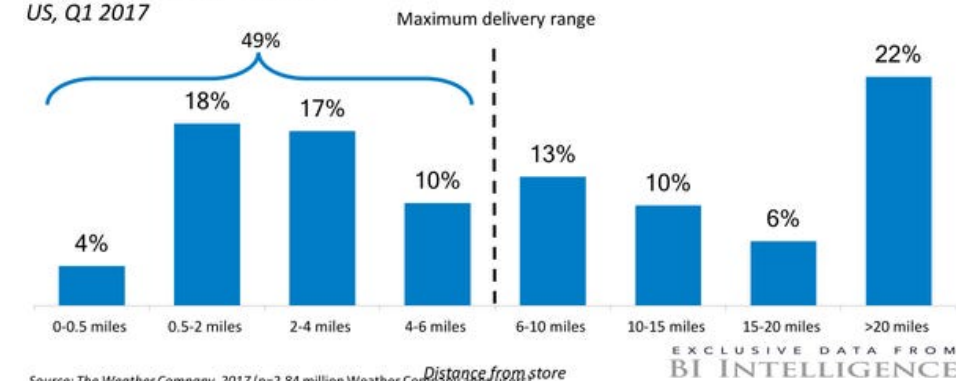
- Amazon, UPS, Domino's, and others invest in last-mile drone delivery



- Walmart is positioned to dominate drone delivery due to its large network of stores (no need for long-range delivery)

Walmart's Potential Customer Base For A Drone Delivery Service

US, Q1 2017





Estimated market size*

* Value of B2C goods delivered by drones

\$251m
[2024]

\$4bn
[2029]

\$65bn
[2034]

Cost of single delivery

\$2
[2034]



\$81
[2034]

Average order value (AOV) of delivered goods

With years of development and proven technical capabilities, drones are reshaping industries through the transformation of traditional delivery methods.

It all started from swift aerial deliveries of medical supplies to remote areas in Africa. Shortly after, it was evident that this model will also elevate commercial applications, mainly for retail and logistics where drone deliveries could bring the most value to customers.

Number of B2C drone deliveries, million



■ North America
 ■ South America
 ■ Europe
 ■ Middle East & Africa
 ■ Asia-Pacific

+66%

CAGR '24-34

63%

77%

65%

72%

64%

Autonomous Vehicles

Motor vehicle crashes caused resulted in over 40,000 fatalities in 2023; a study by the National Highway Traffic Safety Administration (NHTSA) has shown that 94% of crashes are due to *human errors*!

The global Autonomous Vehicle market currently sits at 3.36 trillion.

Autonomous driving is quickly becoming the norm

Robo-Taxi Market

The Global Robo Taxi Market is expected to grow to 450 Billion by 2033 from 1.7 Billion in 2023

It is projected to grow at a rate of 74.7%

Expected: A significant reduction in vehicle crashes (autonomous cars don't drink and drive, don't text, don't get tired and fall asleep at the wheel, and don't have emotional issues that distract them from driving)

Robotaxi Market Overview



74.7%
CAGR GROWTH

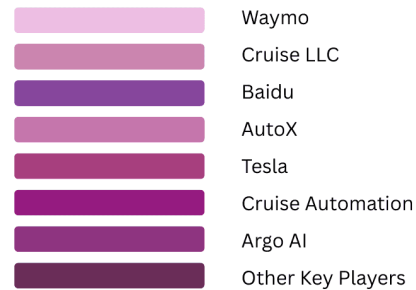


The Global Robotaxi Market size is expected to be worth around USD 450.2 Billion by 2033, from USD 1.7 Billion in 2023

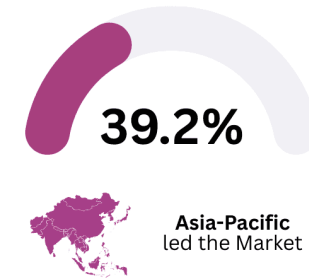
Robotaxi Market, 2023-2033 (USD Billion)



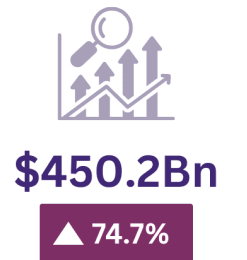
Key Companies



Leading Region

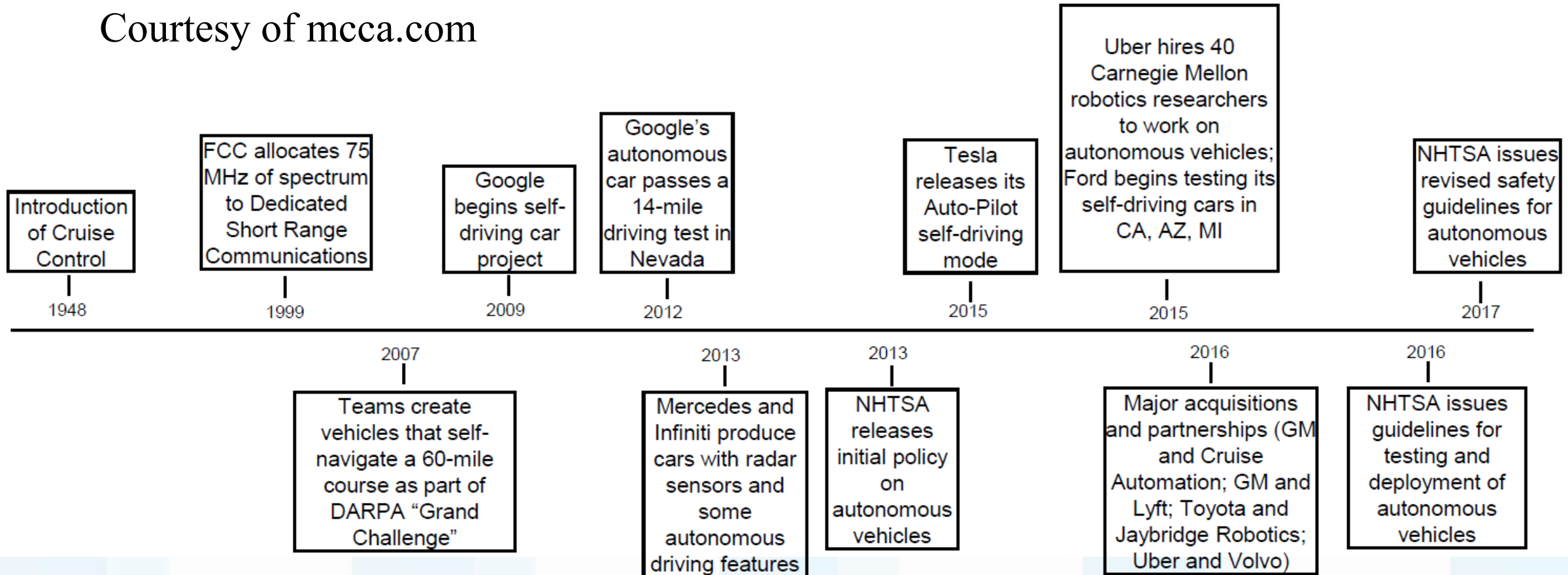


Market Size Forecast 2033



History of Autonomous Vehicles

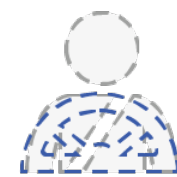
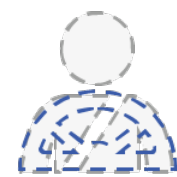
Courtesy of mcca.com



Levels of Automation

SOCIETY OF AUTOMOTIVE ENGINEERS (SAE) AUTOMATION LEVELS

Full Automation



0

No Automation

Zero autonomy; the driver performs all driving tasks.

1

Driver Assistance

Vehicle is controlled by the driver, but some driving assist features may be included in the vehicle design.

2

Partial Automation

Vehicle has combined automated functions, like acceleration and steering, but the driver must remain engaged with the driving task and monitor the environment at all times.

3

Conditional Automation

Driver is a necessity, but is not required to monitor the environment. The driver must be ready to take control of the vehicle at all times with notice.

4

High Automation

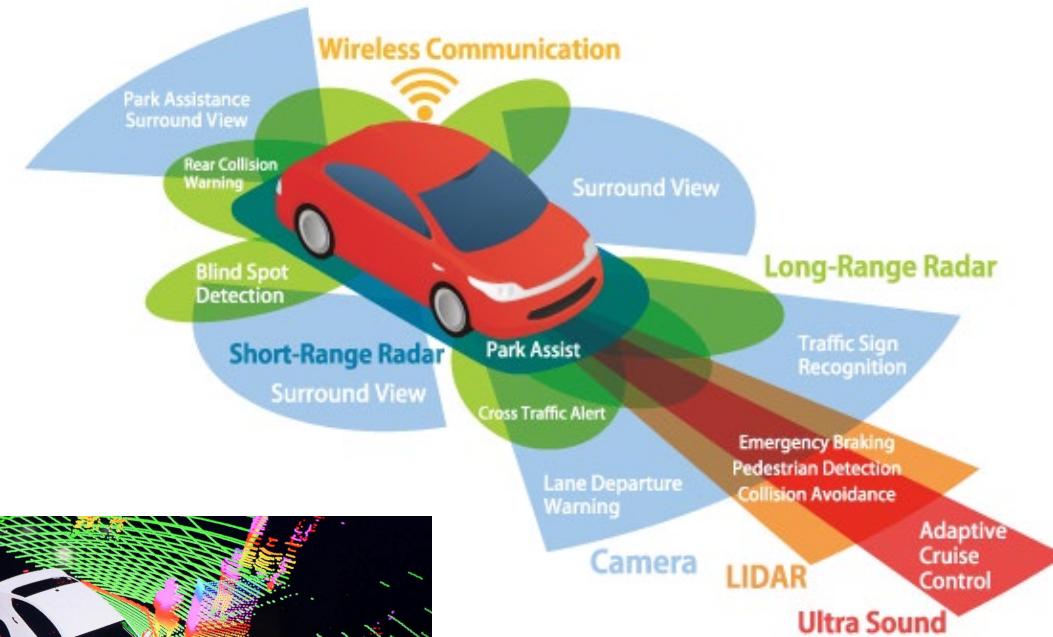
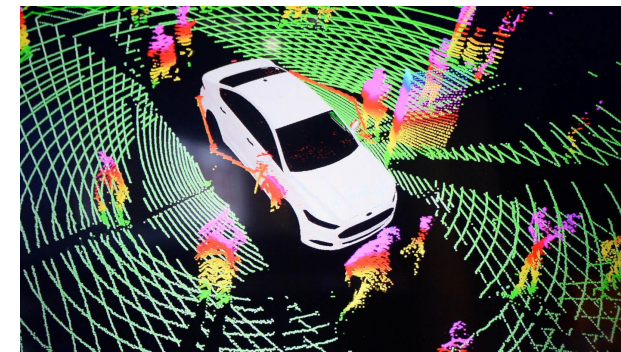
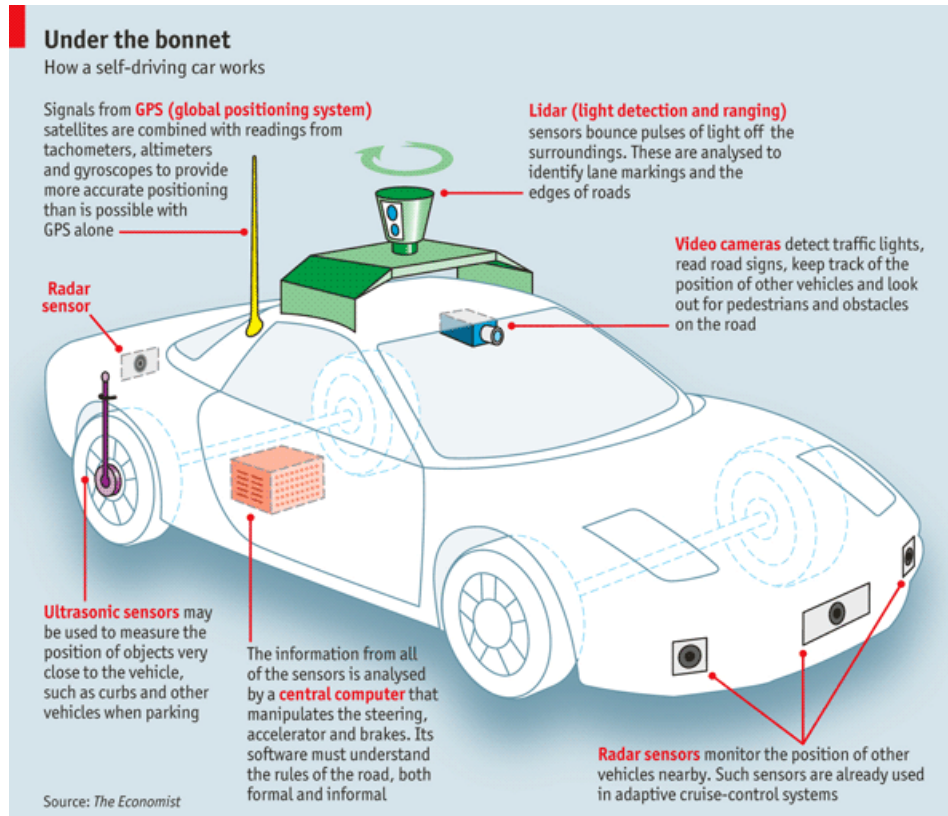
The vehicle is capable of performing all driving functions under certain conditions. The driver may have the option to control the vehicle.

5

Full Automation

The vehicle is capable of performing all driving functions under all conditions. The driver may have the option to control the vehicle.

Sensors



A Multimodal Vision-Language-Action (VLA) Model: From Perception to Actuation

Perception:

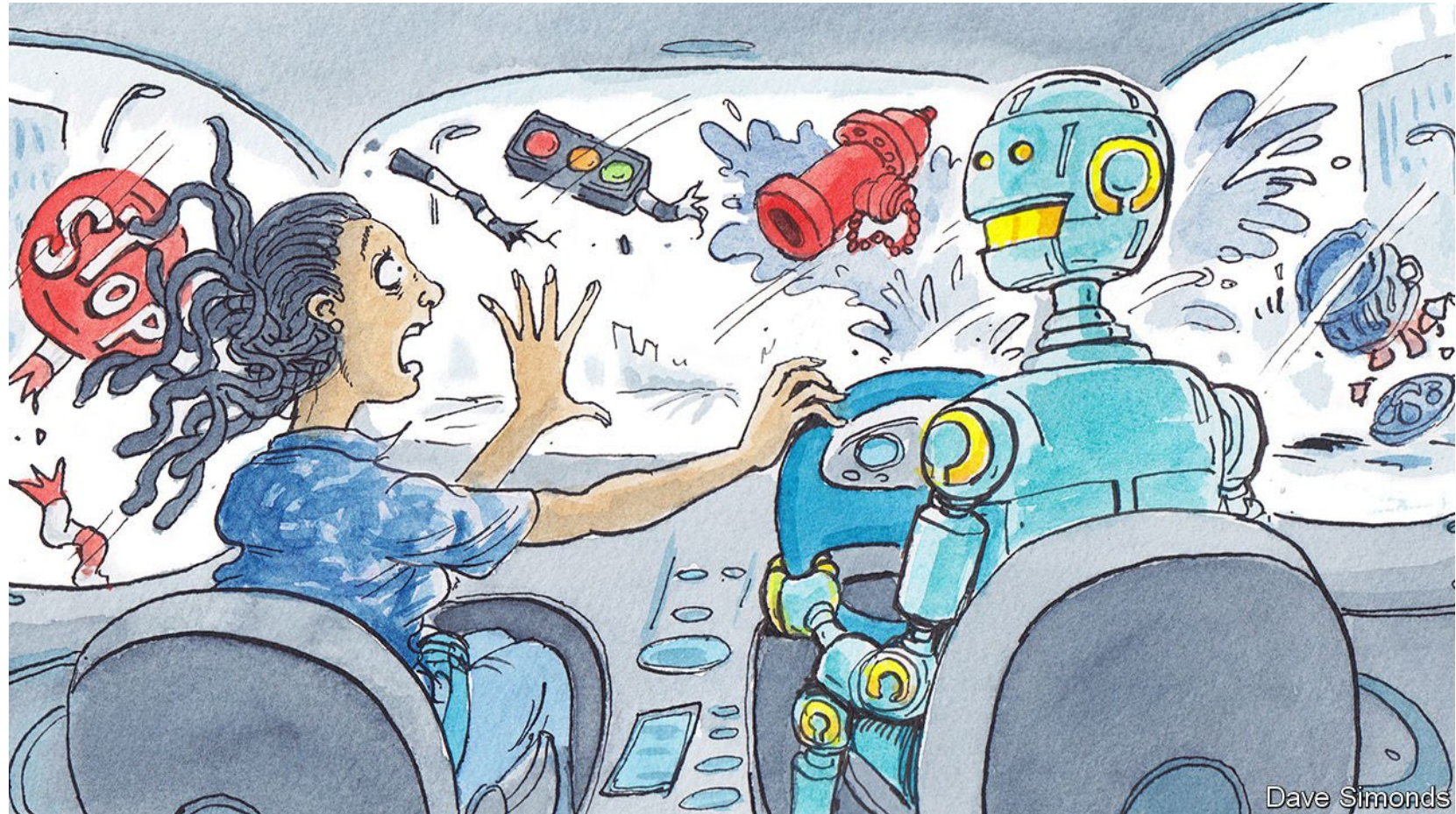
- **Cameras:** Provide real-time obstacle detection to facilitate lane departure and track roadway information (like road signs).
- **Radar:** Radio waves detect short & long-range depth.
- **LIDAR:** Measures distance by illuminating target with pulsed laser light and measuring reflected pulses with sensors to create 3-D map of area.
- **GPS:** Triangulates position of car using satellites. Current GPS technology is limited to a certain distance. Advanced GPS is in development.
- **Ultrasonic Sensors:** Uses high-frequency sound waves and bounce-back to calculate distance. Best in close range.
- **DSRC Receiver:** Communications device permitting vehicle to communicate with other vehicles (V2V) using DSRC, a wireless communication standard that enables reliable data transmission in active safety applications.

Actuation: Steering, Gas/Breaks

Context: Current Plan (e.g., drive to Chicago)

Robotic Rules of the Road

Policy issues,
formats, and
standards



An Ethical Issue: Liability

Will courts treat autonomous vehicles as drivers and apply a negligence standard or as sophisticated technology and apply a product liability standard?

How will liability be apportioned?

- Fleet Operator/Service Providers
- Vehicle manufacturers
- Technology companies/software manufacturers
- Local government's responsible for maintaining infrastructure

Ethical Considerations in Autonomous Car Design

“Human drivers may be forgiven for making an instinctive but nonetheless bad split-second decision, such as swerving into incoming traffic rather than the other way into a field. But programmers and designers of automated cars don’t have that luxury, since they do have the time to get it right and therefore bear more responsibility for bad outcomes.”

- Patrick Lin, The Atlantic

The Trolley Problem

A trolley's brakes have failed.

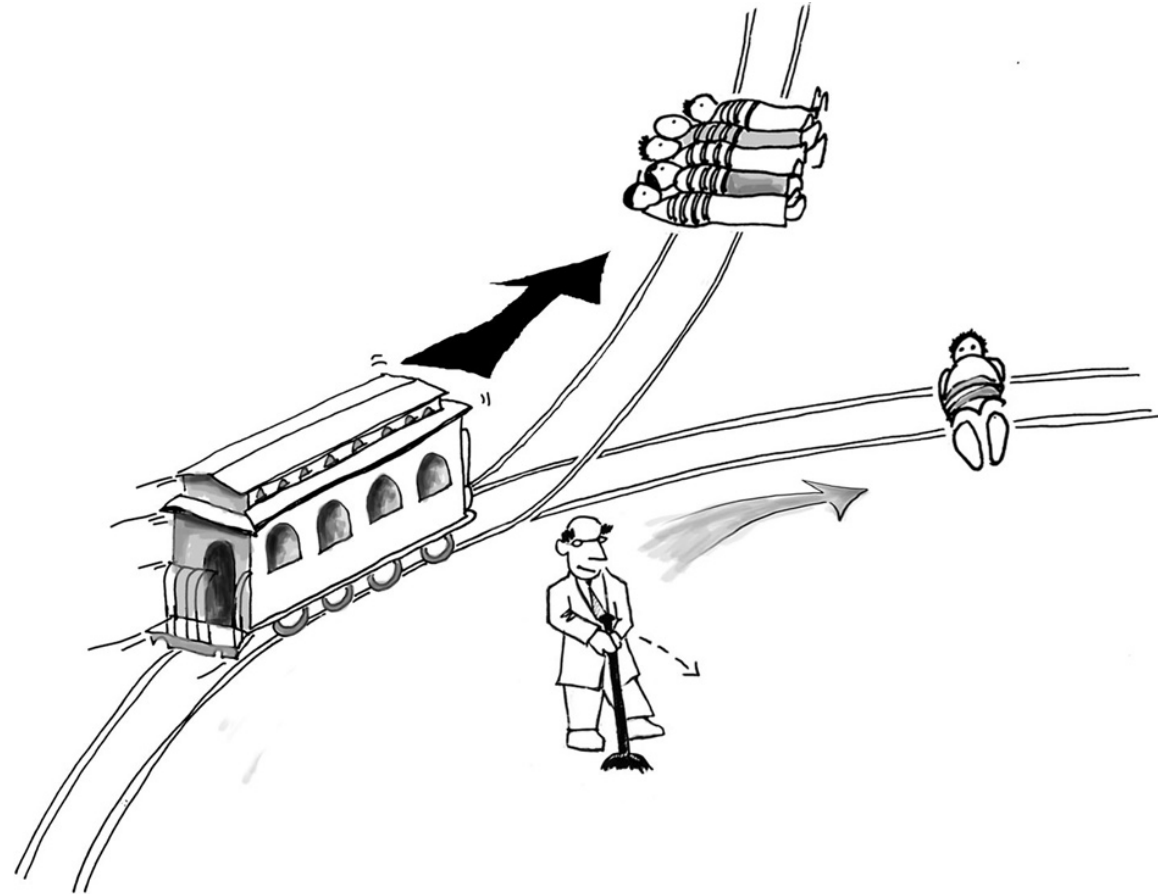
You are controlling the signal switch.

If you do nothing, five people will be killed.

If you activate the switch, only one person will be killed.

What do you choose to do?

Critical distinction: Allowing death versus causing death?



A Tesla Example

The video involves content depicting a live car crash that might be disturbing to some viewers (although no one was seriously hurt)

<https://www.facebook.com/watch/?v=1117159980371396>

Ethical/Societal Considerations in Fairness

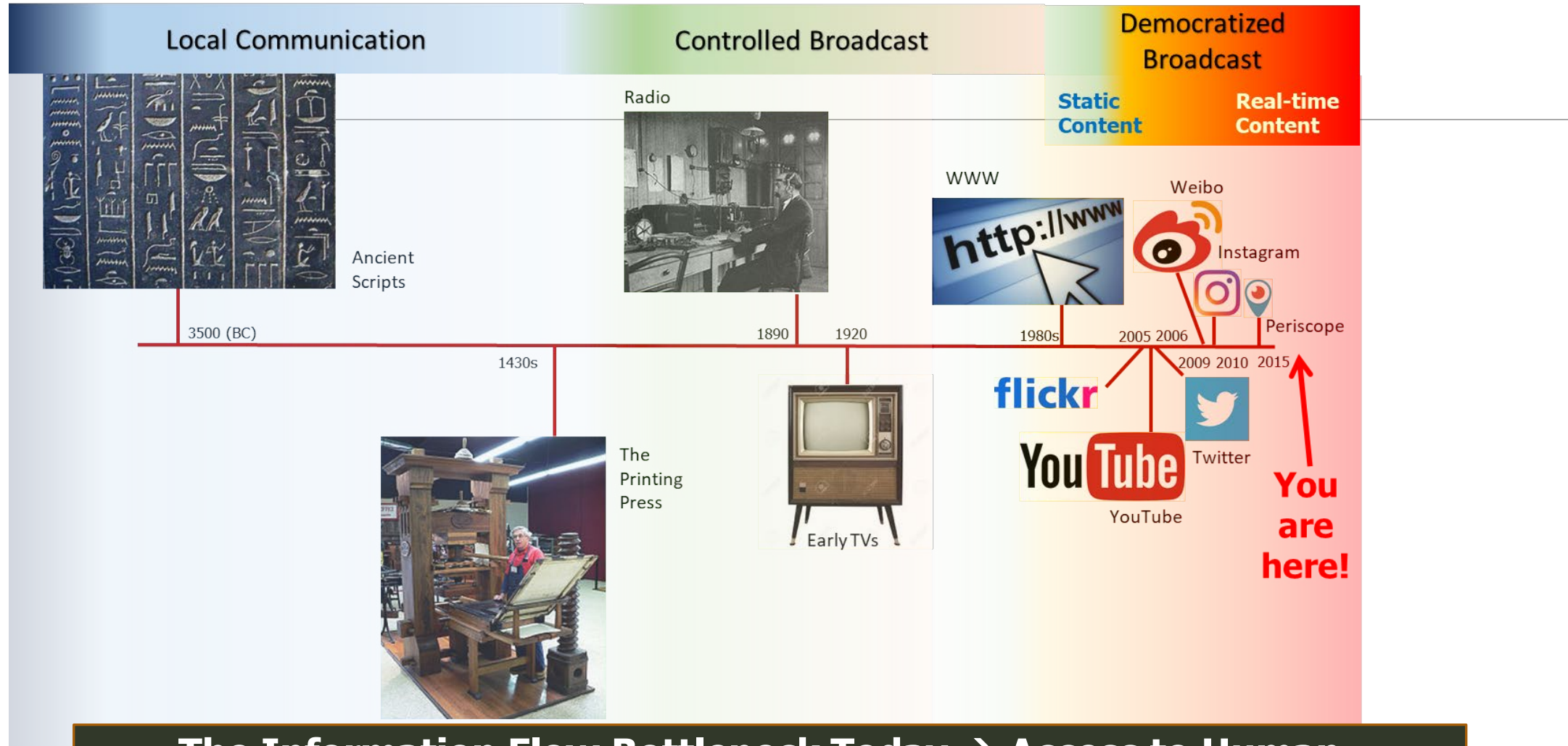
Algorithmic Fairness in AI

Algorithmic AI and the Click Economy

Algorithmic Fairness

- A key application of AI (besides autonomy) is to bring you information!
- AI allows the next level of information customization to fit each recipient.
- What are the implications in the age of Attention Economy and click-maximizing incentives?

We Are at a Turning Point in Human History



The Information Flow Bottleneck Today → Access to Human Attention, not Access to Broadcast Media

Social Dynamics in the Age of Global Information Sharing/Access

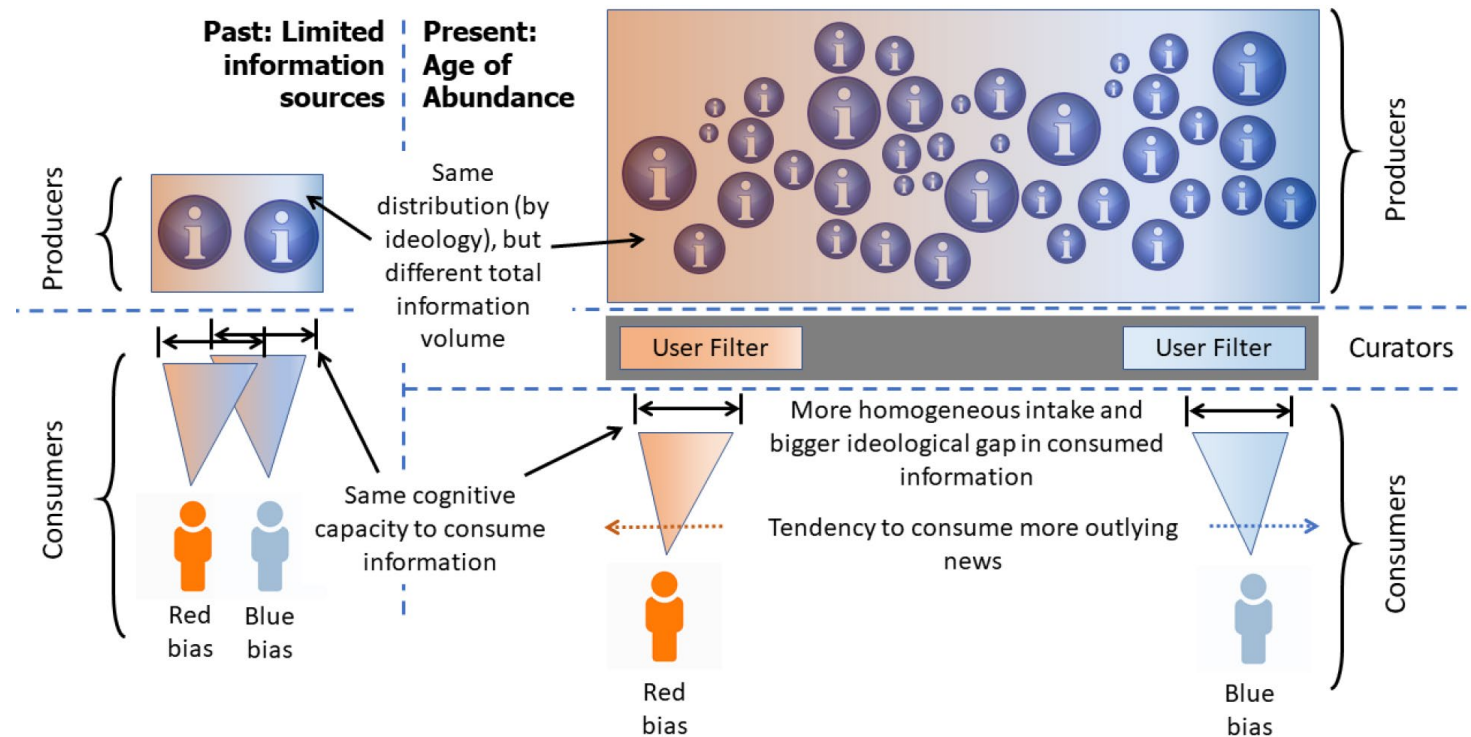
Ease of information access (at scale)

→ Increased need to prioritize content consumption

→ **Bigger role for ingrained biases (confirmation bias, negativity bias, ...)**

→ Less ideologically-diverse consumption (higher customization)

→ Increased ideological isolation and separation



Modeling Belief Dynamics in the Age of Democratized Information Sharing/Access

Key elements:

Individuals tend to favor information that is consistent with their own pre-existing beliefs

- Set of “plausible” inputs:

$$\mathcal{X}^{(i)}(t) = \{a_j \mid |x_i(t) - x_j(t)| \leq \epsilon_i\}$$

- Belief updates:

$$x_i(t + \Delta t) = (1 - \alpha_{\Delta t})x_i(t) + \alpha_{\Delta t}f(\mathcal{X}^{(i)}(t)) + \sigma\Delta W$$

Modeling Belief Dynamics in the Age of Democratized Information Sharing/Access

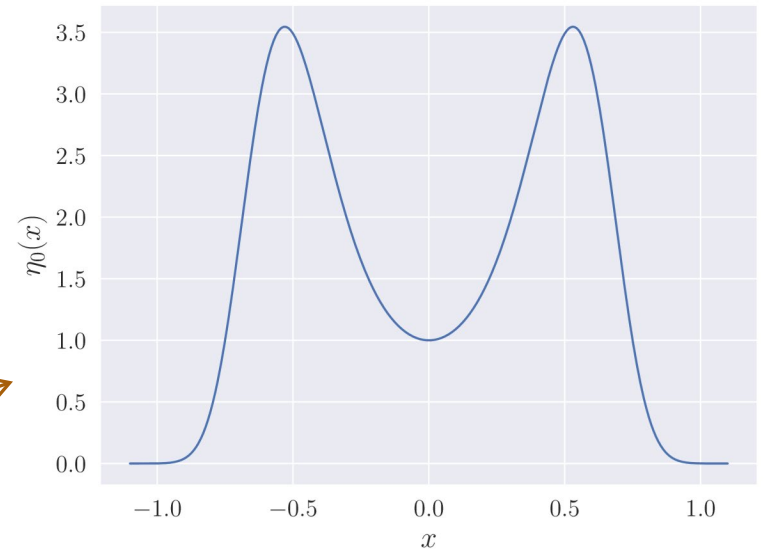
Preference for outlying content:

$$x_i(t + \Delta t) = (1 - \alpha_{\Delta t})x_i(t) + \alpha_{\Delta t}f(\mathcal{X}^{(i)}(t)) + \sigma\Delta W$$

$$f(\mathcal{X}^{(i)}(t)) = \frac{\sum_{a_j \in \mathcal{X}^{(i)}(t)} x_j(t) \eta(x_j(t), t)}{\sum_{a_j \in \mathcal{X}^{(i)}(t)} \eta(x_j(t), t)}$$

$$\eta(x, t) = \eta_0(x)e^{\kappa\rho_s(x, t)}$$

$$\mathcal{X}^{(i)}(t) = \{a_j \mid |x_i(t) - x_j(t)| \leq \epsilon_i\}$$

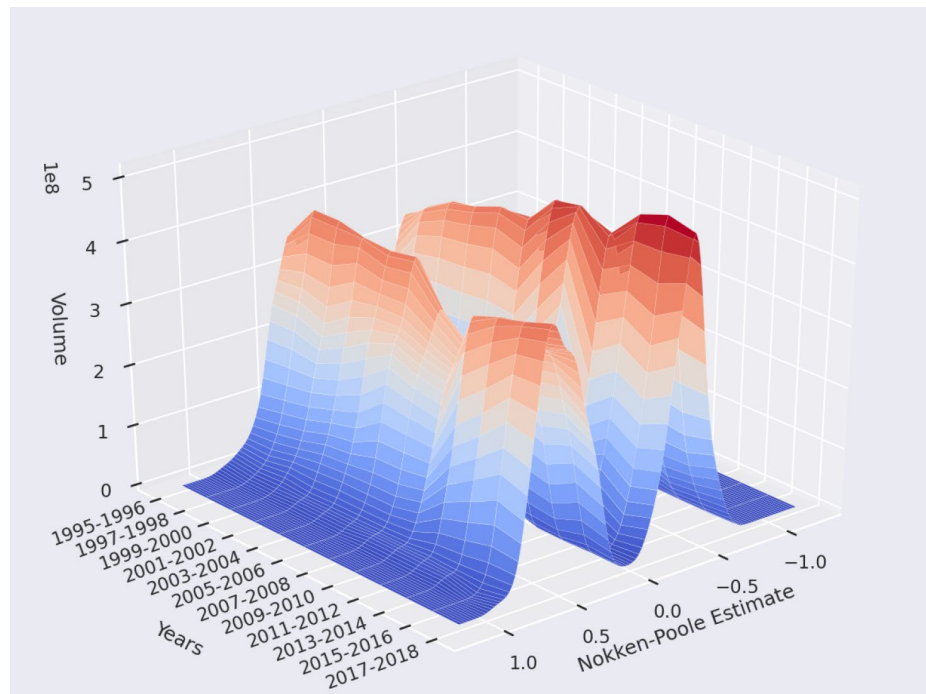


An example $\eta_0(x)$

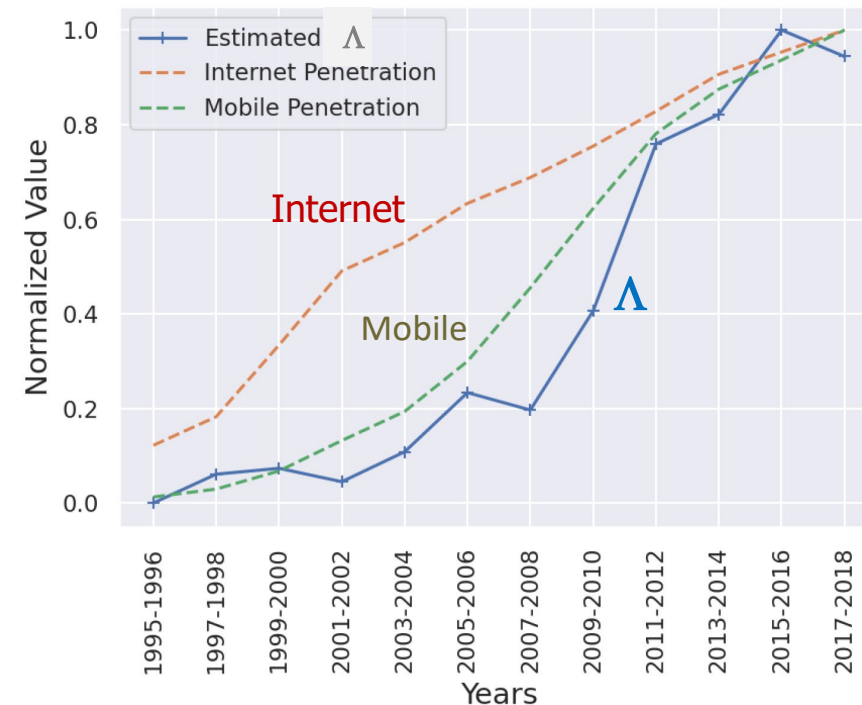
Impact of Global Information Sharing/Access on Belief Distribution

A US Congress Case Study

Actual distribution of beliefs (Nokken-Poole Estimate) of US Congress Members over the last 25 years



Comparison of (the estimated best fit) volume Δ to Internet and Cell Phone penetration



Persuasion, Morality, and AI-Based Customization: The Moral Foundation Theory

Evolution predisposed us to adopt several behaviors that historically had an evolutionary advantage; these behaviors form our basis of morality:

Care (compassion, not inflicting harm)

Loyalty (belonging to a team, being part of a group)

Fairness (equality, honesty, justice, non-discrimination, etc.)

Authority (respect for rules, tradition, or a higher structure)

Purity (sanctity, freedom from taboos)

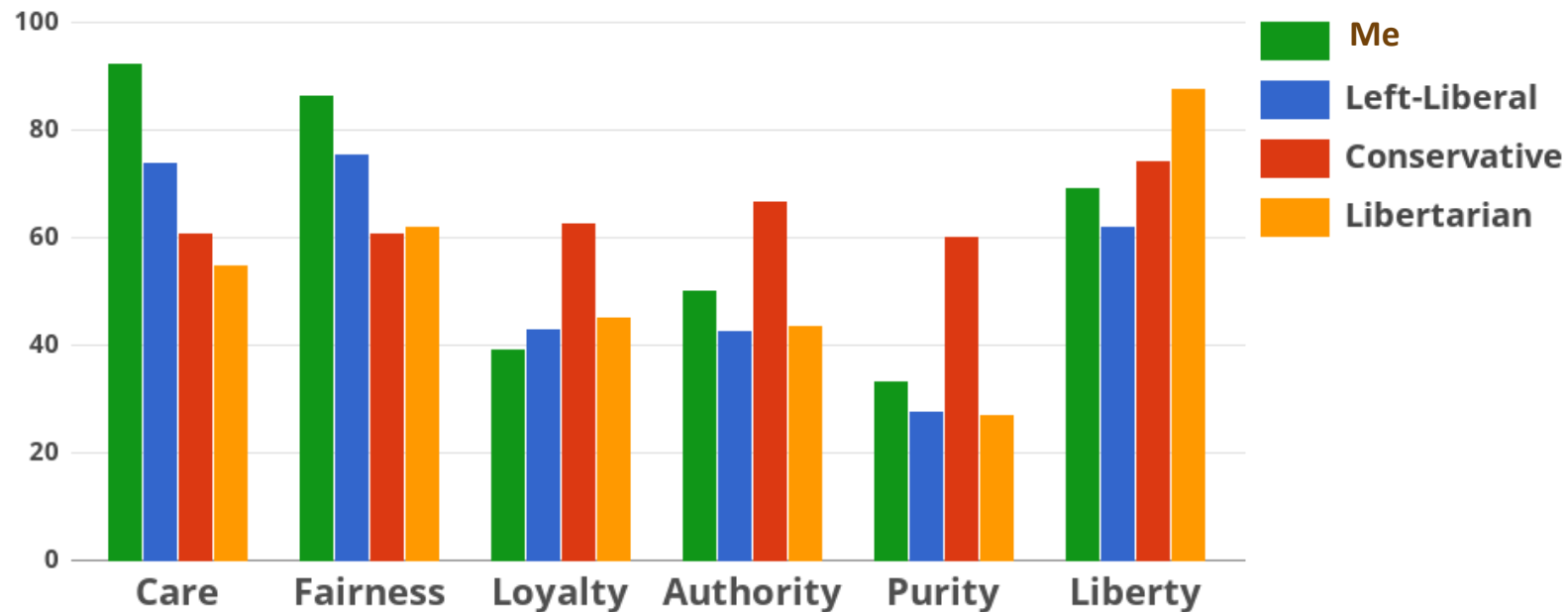
Liberty (independence, autonomy, being your own master)

Not all of the above foundations are equally reinforced in different cultures; Different cultures have a different “moral DNA”!

AI Can Perform Moral Fingerprinting (of an Individual or Community) based on Vocab

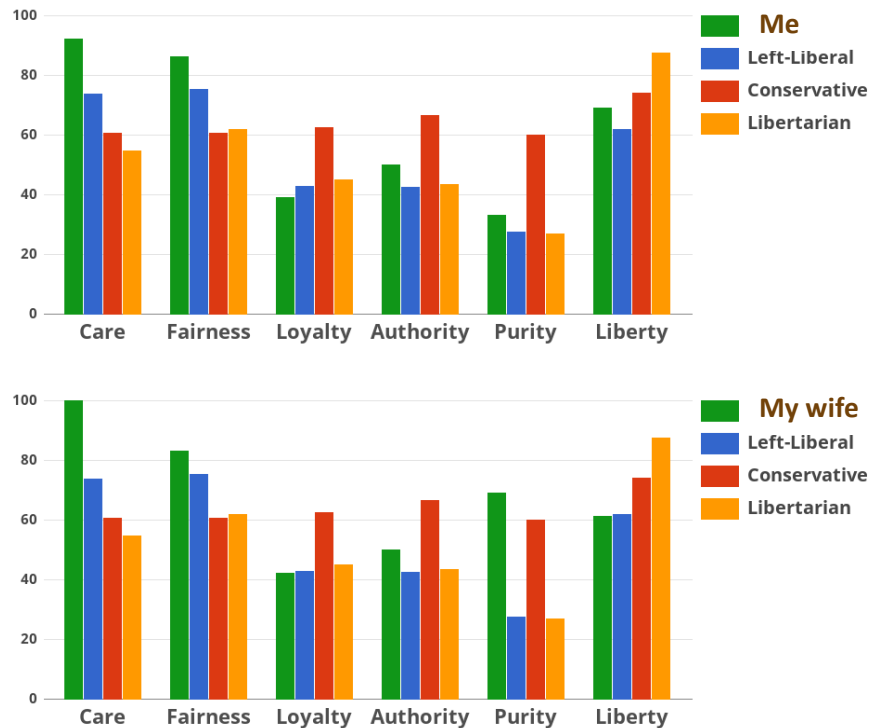
The current state of the art is formal surveys

(According to the moral foundations test at:
<https://www.idrlabs.com/morality/6/test.php>)



Application of Moral Fingerprinting: match.com++?

(According to the moral foundations test at: <https://www.idrlabs.com/morality/6/test.php>)



Hypothesis: Agreement on moral foundations leads to a better mutual understanding

Persuasion

Example: Excerpts from Barack Obama's speech* to rally US support for action against ISIL (also known as ISIS) in 2014 (by appealing to moral values of *Authority*, *Care*, *Fairness*, and *Loyalty*)

- “ISIL is not Islamic”, “ISIL is certainly not a state”
 - **Authority infraction:** No higher order or structure obligates one to identify with ISIL's cause.
- “ISIL is a terrorist organization”, “... it has no vision other than the slaughter of [innocents]”
 - **Care infraction:** They inflict harm in violation of moral values.
- “They enslave, rape, and force women into marriage”, “They threatened a religious minority with genocide”
 - **Fairness infraction:** They discriminate by gender, religion, etc.
- “They took the lives of two American journalists”
 - **Loyalty appeal:** It is patriotic (of Americans) to stand together against ISIL.

*<https://time.com/3320666/obama-isis-speech-full-text/>

AI and Persuasion

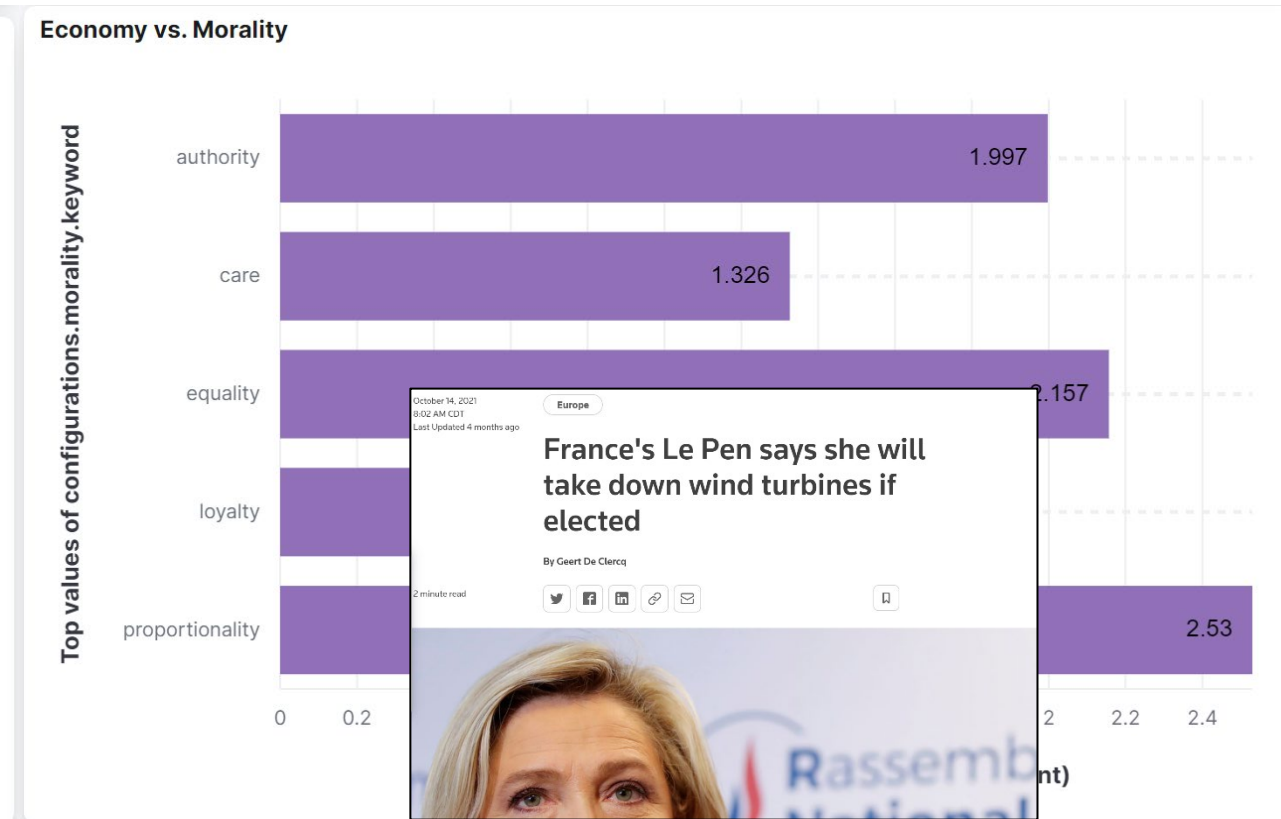
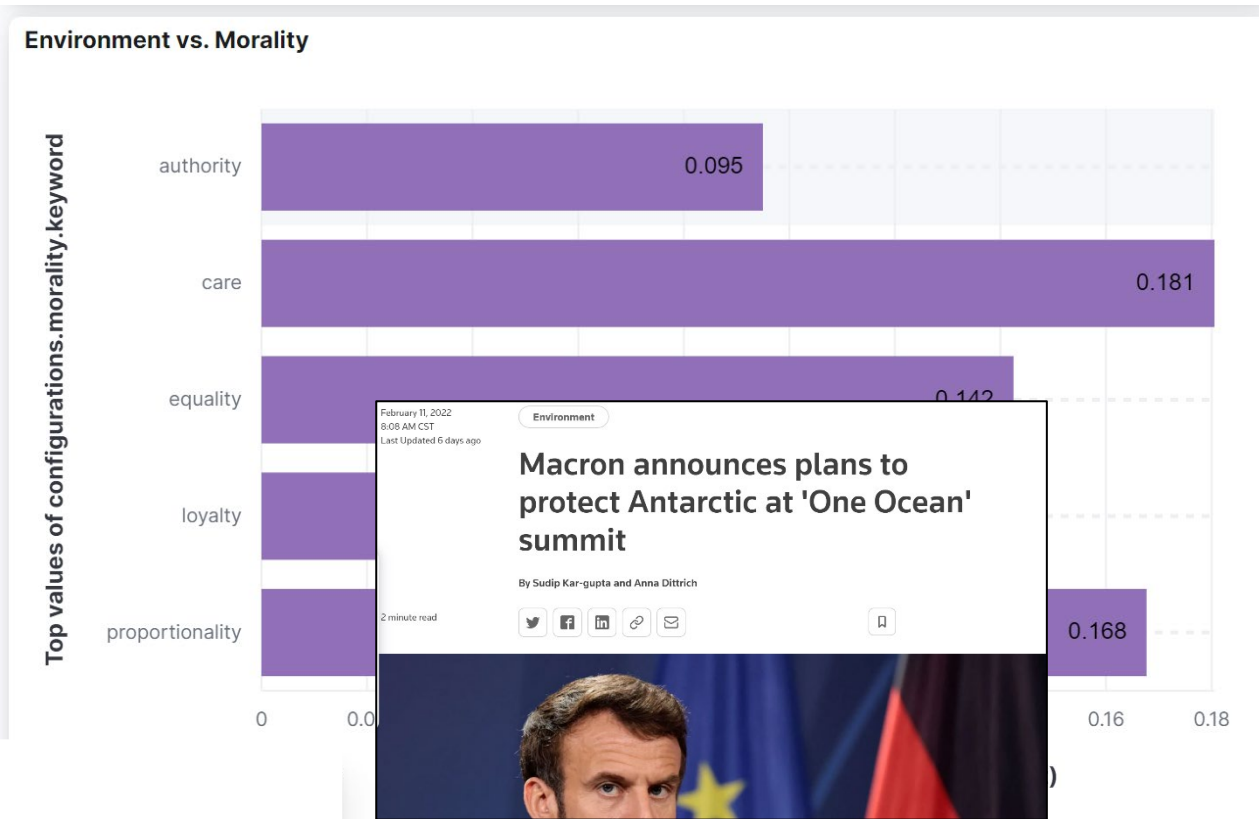
AI can craft arguments to convince someone of an act based on the recipient's moral fingerprint

Example (I used ChatGPT to offer an argument for vaccination from different moral perspectives):

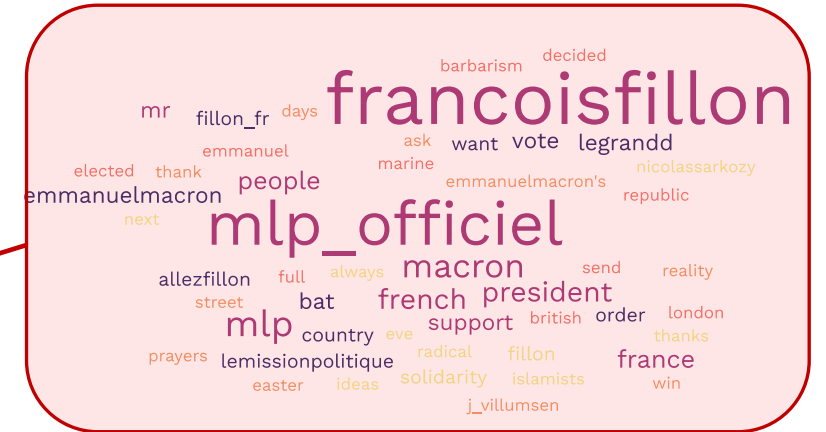
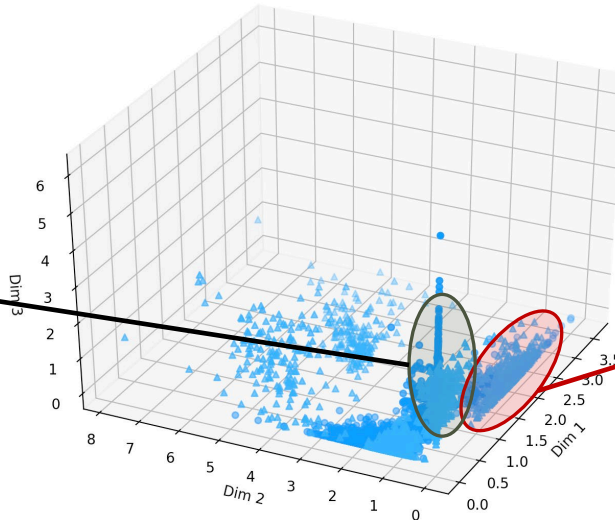
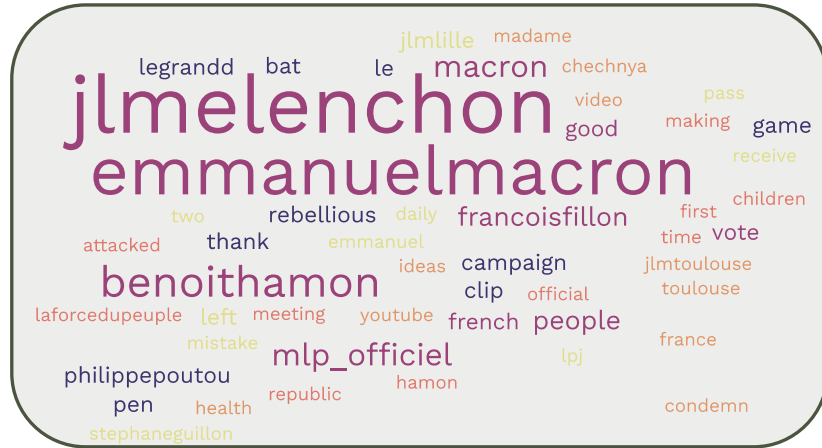
- **Fairness:** “Relying on others to keep disease low while opting out yourself shifts risk onto them—fairness means doing your part, and vaccination is one of the simplest ways to share that responsibility.”
- **Authority:** “Public health authorities and medical experts across the world consistently recommend vaccination as one of the most effective and evidence-based ways to prevent serious disease and protect community health.”

Understanding and Predicting Population Response

- Example from the French Election Data Set
 - Which (moral) demographic are most concerned about economy?
 - Which (moral) demographic are most concerned about the environment?



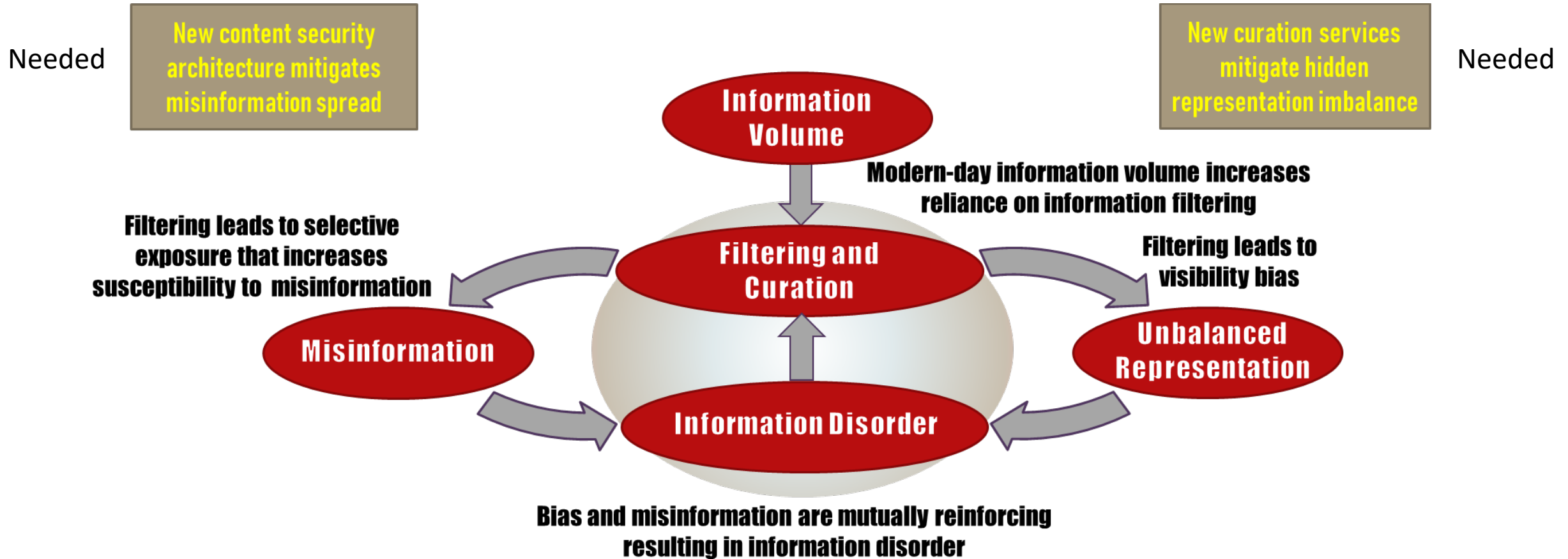
Understanding and Predicting Population Response



Embedding of Users and Messages

Conclusion: Our automated ideology-based population and message mapping (embedding) techniques can **distill the fundamental ideological divides among key population segments** and expose **their psychographic and demographic** affinities.

The Click (and Persuasion) Economy



Ethical/Societal Considerations in Security

An Agentic AI Perspective

Agentic AI threats

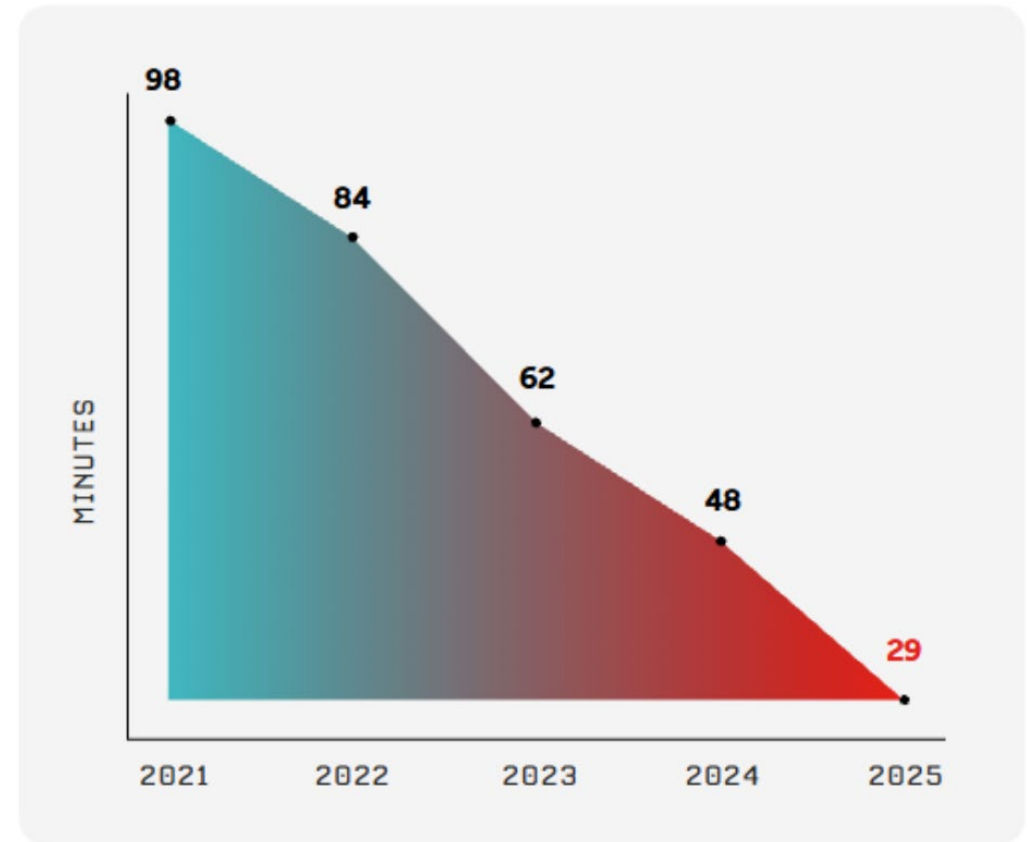
Agentic AI and the Growing Security Threat Landscape

The average eCrime breakout time fell to **29 minutes in 2025**

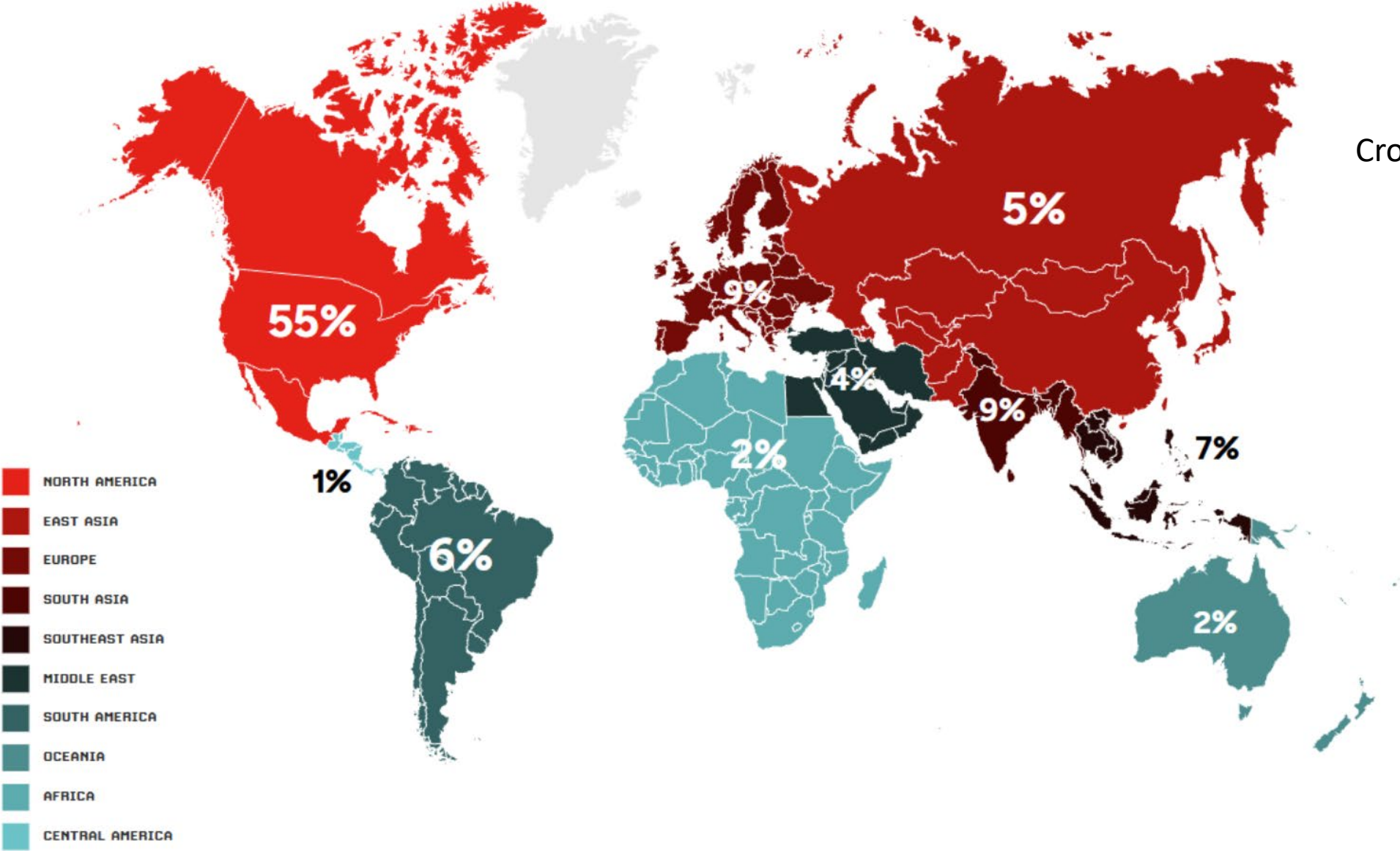
65% increase in speed from the prior year.

The fastest breakout took just 27 seconds.

In one intrusion, data exfiltration began within four minutes of initial access. The window to detect, decide, and respond has narrowed dramatically.



Interactive Intrusions by Region



CrowdStrike report

Examples of AI-powered Attacks

- OpenAI's ChatGPT image generator enables creation of fake passports
- Uncensored WormGPT variants powered by Grok and Mixtral (WormGPT is an unrestricted AI model that can generate content without safety filters or ethical guardrails)
- Attacks targeting (Atlassian's) Model Context Protocol (MCP) introduce a new "living off AI" risk. "Living off AI" refers to co-opting legitimate AI to do harmful things. Examples include:
 - Prompt injection: Tricking the AI into ignoring instructions and doing something unintended
 - Tool abuse: Getting the AI to use its connected tools (via MCP) in harmful ways (e.g., "send this data," "delete those files")
 - Data exfiltration via the AI: Getting the AI to retrieve sensitive info and leak it.
 - Action chaining: Using multi-step reasoning to escalate impact (e.g., read → summarize → send externally).
- Weaponization of AI Skills: Use of legitimate AI skills for adversarial purposes
 - Example: The use of Claude Mythos vulnerability discovery capability to discover new attack vectors against existing systems

Ethical/Societal Considerations in Sustainability

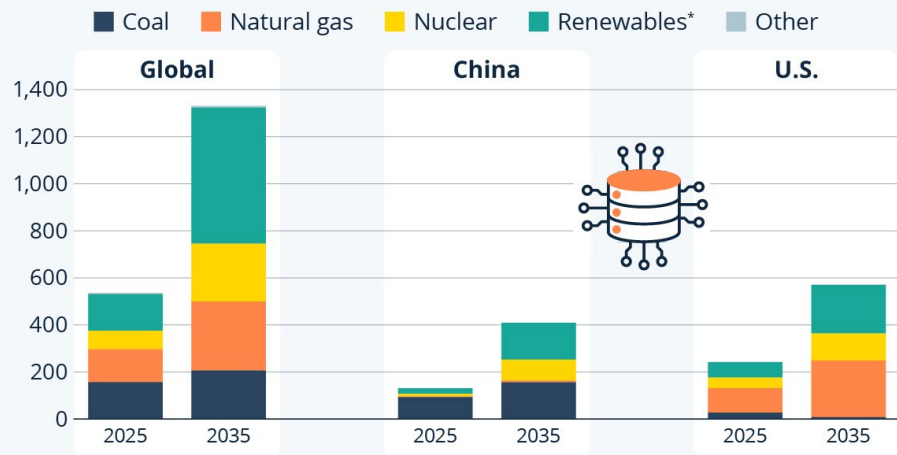
An energy consumption perspective

The growing energy demand in the age of AI

AI and Sustainability

Data Center Energy Consumption Surges Amid AI Boom

Electricity generation to supply data centers, by energy source (in TWh)



* Solar PV, wind and other renewables
Based on the fuel mix of the electricity physically consumed by data centers rather than contractual mix of different data center operators.

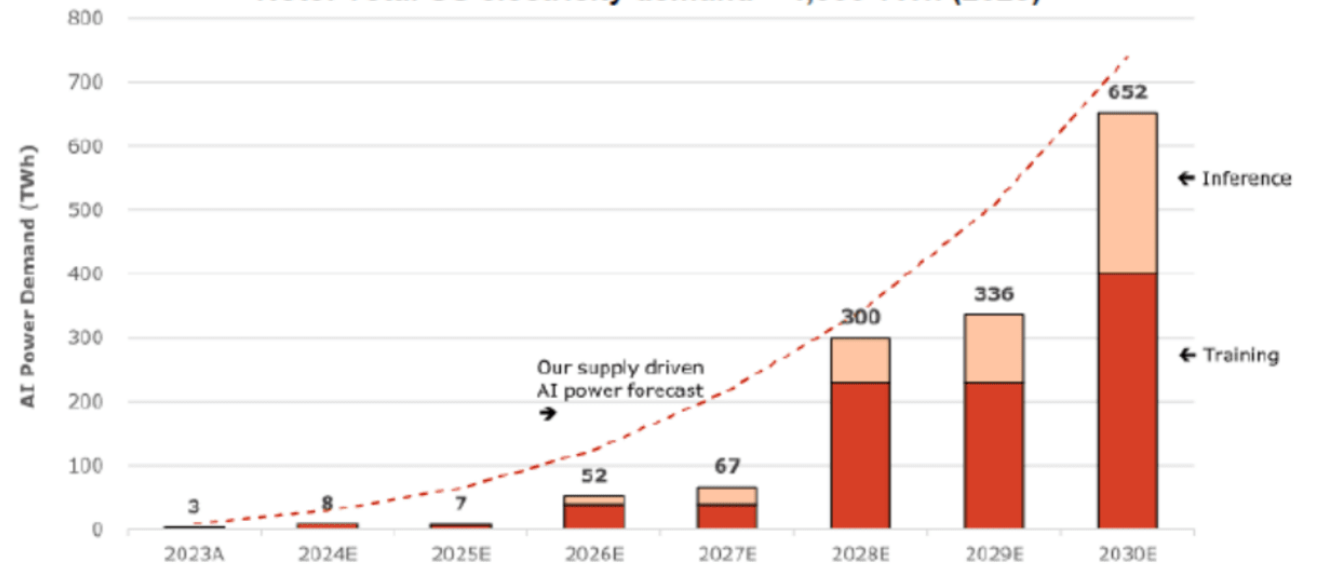
Source: IEA



Summary of GenAI demand forecast

Source: Wells Fargo

Note: Total US electricity demand – 4,000 TWh (2023)



	2023A	2024E	2025E	2026E	2027E	2028E	2029E	2030E
(+) Training power demand (TWh)	3	8	7	40	40	229	229	402
(+) Inference power demand (TWh)	0	0	0	12	27	70	107	250
Demand driven AI power forecast (TWh)	3	8	7	52	67	300	336	652
Supply driven AI power forecast (TWh)	9	29	65	125	217	341	506	739

Source: Wells Fargo Securities, LLC estimates

Summary of Concerns



Concerns with Misuse



Concerns with misinterpretation (by machine)



Concerns with misinterpretation/trust (by human)



Concerns with side-effects of regular use



Concerns with Misuse

Summary of Concerns

The screenshot shows a Tributech blog post. At the top left is the Tributech logo. The navigation menu includes 'Product', 'Blog', 'Use Cases', 'Career', 'Partners', and a 'Get Started' button. Below the navigation is a 'BACK TO BLOGS' link. The main content area features the date 'BLOG | AUG 17, 2020' and the title 'TAMPERED IOT DATA - THE "FAKE NEWS" OF THE INDUSTRIAL IOT'. The text discusses the risks of IoT data tampering and mentions Tributech's data notary service. A 'Data Stories' tag is visible. The author is identified as Thomas Plank, CEO of Tributech, with a '4 min' read time. On the right side of the screenshot, there is a partial view of another article with a typewriter image and the text 'Fake News' on a sheet of paper.

Why fake data is a serious IoT security concern

Feature
Nov 07, 2018 • 10 mins
Internet of Things | Security | Security Infrastructure

Fake internet of things data could break your business and kill your customers.

Big models will facilitate the generation of “fake” IoT data



Credit: Getty Images



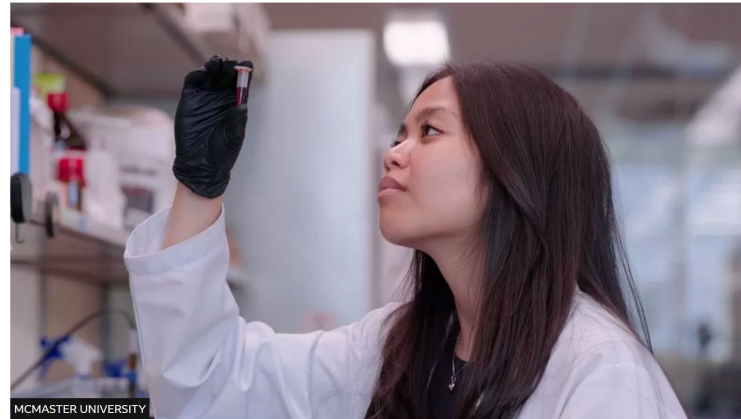
Summary of Concerns



Concerns with Misuse

New superbug-killing antibiotic discovered using AI

© 25 May · Comments



MCMaster UNIVERSITY

Scientist Denise Catacutan working on the experimental antibiotic discovered with the help of artificial intelligence.

Big model facilitate discovery of novel artifacts at an accelerated rate

[ChatGPT's Code Interpreter](#) [Can A.I. Be Fooled?](#) [Chatbot Prompts to Try](#) [A.I.'s Literary Skills](#)

A.I. Poses 'Risk of Extinction,' Industry Leaders Warn

Leaders from OpenAI, Google DeepMind, Anthropic and other A.I. labs warn that future systems could be as deadly as pandemics and nuclear weapons.

Give this article 1.4K

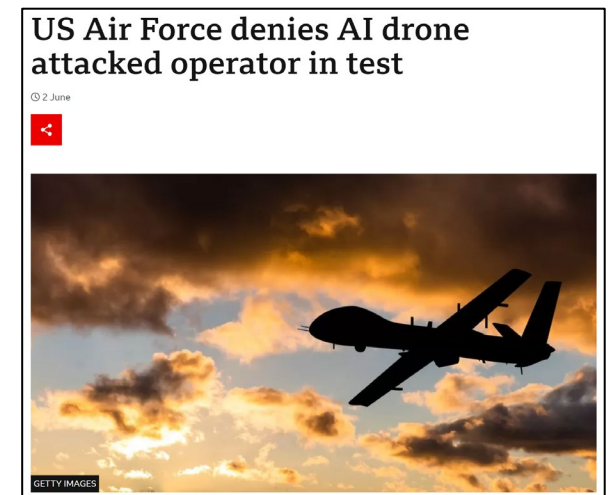


Summary of Concerns

Prevalence of AI in IoT will enhance machine autonomy and take humans increasingly out of the loop which increases vulnerability to machine error

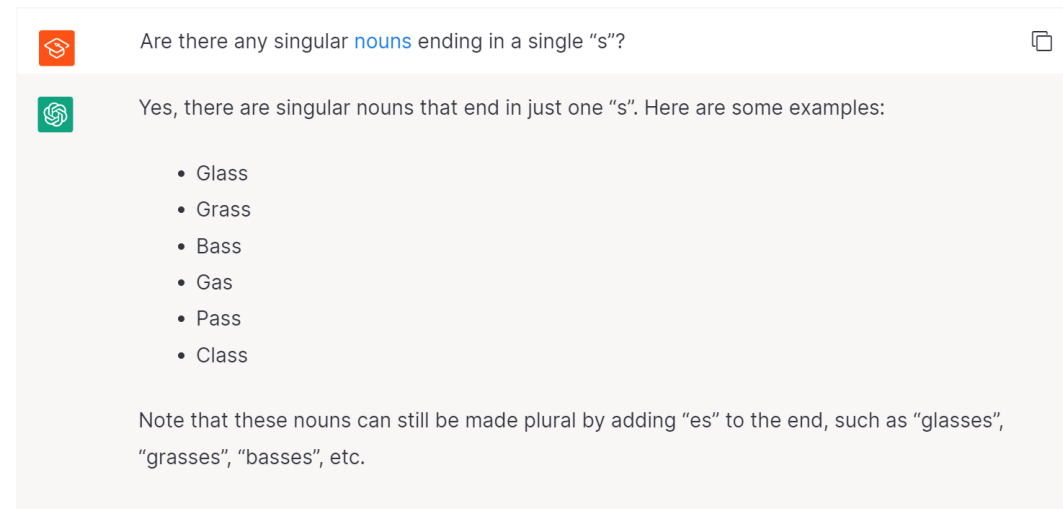


Concerns with misinterpretation (by machine)



Summary of Concerns

Example: ChatGPT answers incorrectly



The screenshot shows a chat interface with a user question and a ChatGPT response. The user asks, "Are there any singular nouns ending in a single 's'?" ChatGPT responds, "Yes, there are singular nouns that end in just one 's'. Here are some examples:" followed by a bulleted list: "Glass", "Grass", "Bass", "Gas", "Pass", and "Class". Below the list, a note states, "Note that these nouns can still be made plural by adding 'es' to the end, such as 'glasses', 'grasses', 'basses', etc."

Are there any singular nouns ending in a single "s"?

Yes, there are singular nouns that end in just one "s". Here are some examples:

- Glass
- Grass
- Bass
- Gas
- Pass
- Class

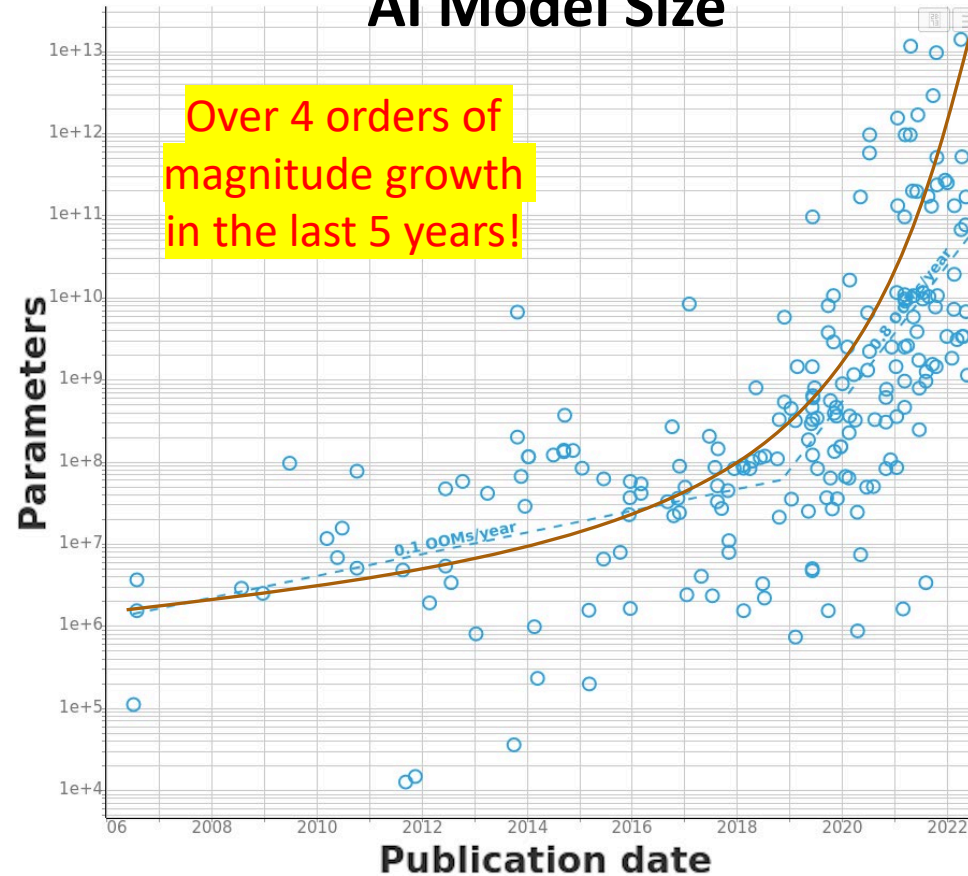
Note that these nouns can still be made plural by adding "es" to the end, such as "glasses", "grasses", "basses", etc.



Concerns with misinterpretation/trust (by human)

Summary of Concerns

AI Model Size



Moore's Law: Capacity doubles every **18 months**.

AI model size doubles approximately every **3.4 months**.

We are heading towards a severe computational bottleneck!



Concerns with side-effects of regular use