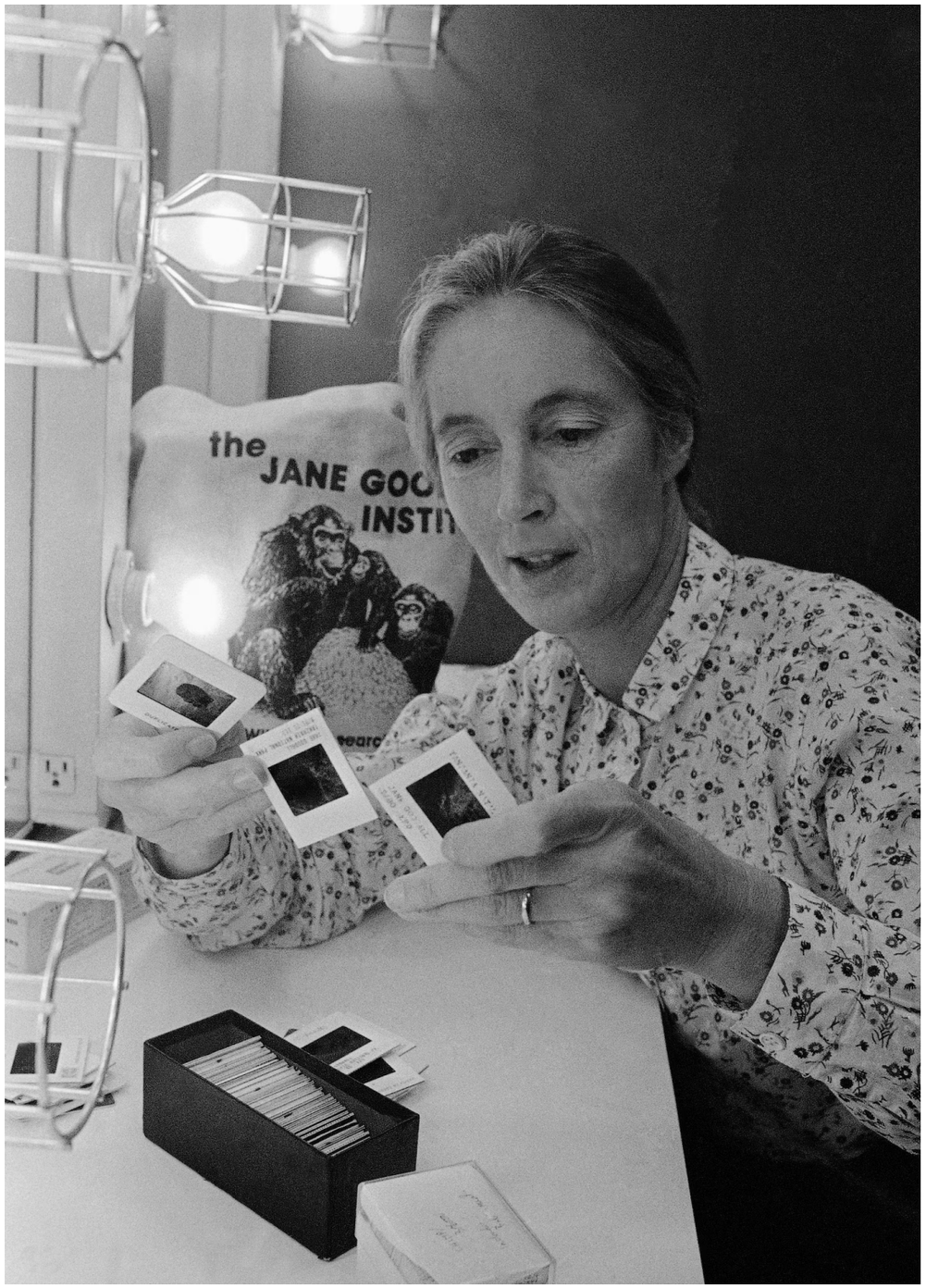


Jane Goodall

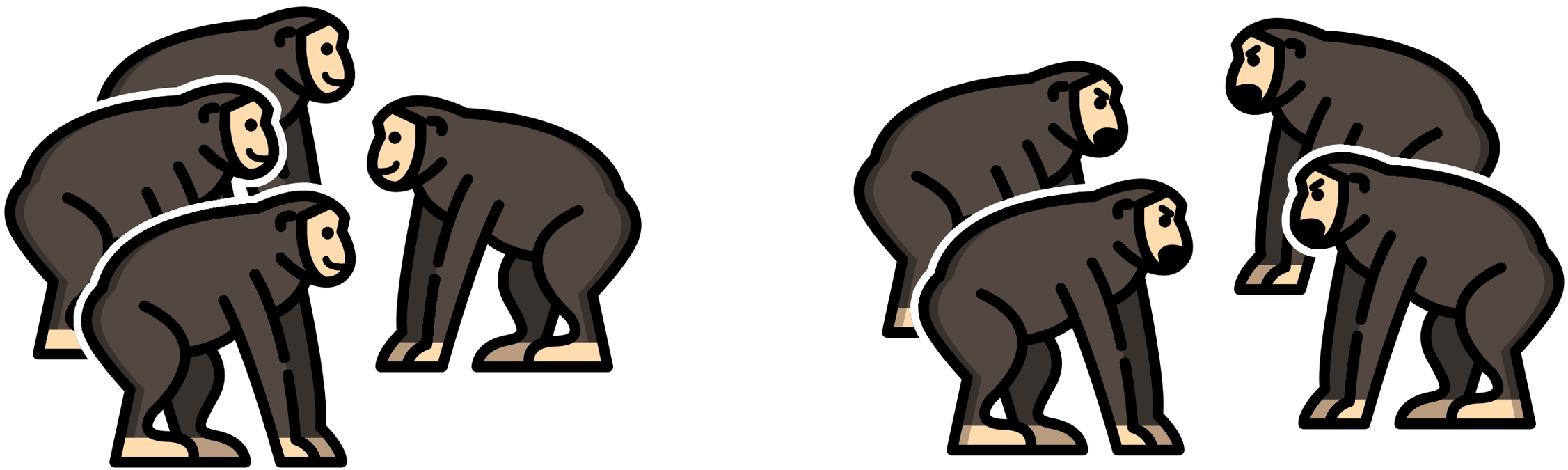
1934-2025

Dr. Goodall examining slides in 1982. (Charles Knoblock/AP)
Washington Post this morning

$$cost_i < \sum_{j=0}^{N} (relatedness_{ij} \times benefit_j)$$



Hamilton’s Law (1964) – Inclusive Fitness



Wilson, Boesch, &al. (2014)

AI, Corporate Responsibility, and Democratic Legitimacy

Is DevOps the Answer?

Hertie School
Centre for
Digital Governance

Joanna J. Bryson
@j2bryson.bsky.social
mastodon.social/@j2bryson
linkedin.com/in/bryson

DevOps vs AGI



Hertie School
Centre for
Digital Governance

Joanna J. Bryson

@j2bryson @🦣.social

in/bryson


AGI
Artificial General Intelligence
AI exactly like a human only immortal and/or omniscient.
Fiction
DevOps
Development and Operations
The process of creating and maintaining software.
Obligatory



AGI

Yesterday in Frankfurt

(bankers! Sibos. Photo credit : Andy Young via LinkedIn)



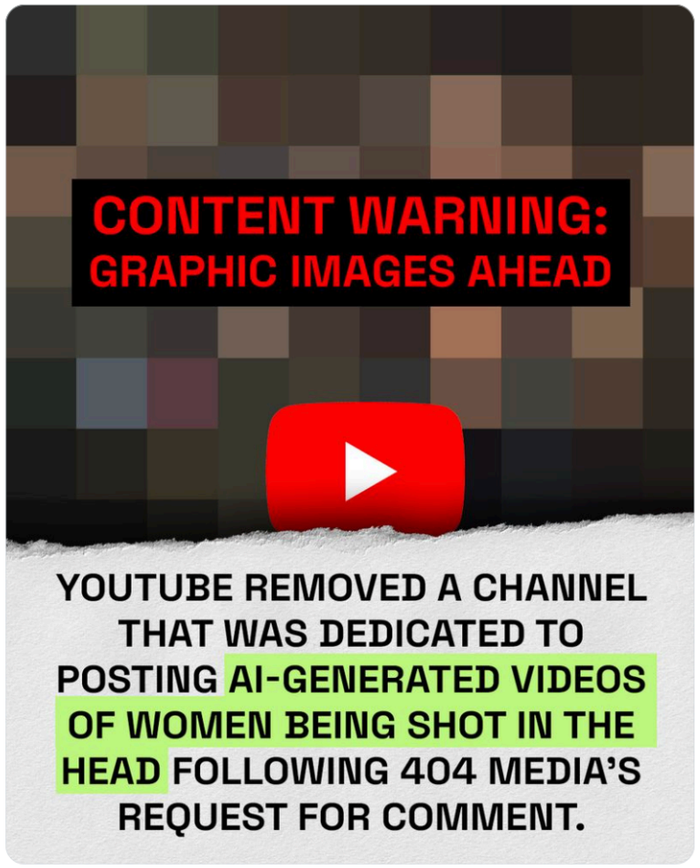
The Limits of Artificial Intelligence and the Risks of Digital Superintelligence
25 September 2025
European Parliament
Host: Emil Radev, MEP

When AI Fails: Real-World Examples of Limitations and Unintended Consequences

Simon J. Greenhill
@simongreenhill.bsky.social

here's an AI-harm that popped up on my newsfeed this morning 😱: bsky.app/profile/404m...

> **404 Media** @404media.co · 6d
> The channel, named Woman Shot A.I, started on June 20, 2025. It posted 27 videos, had over 1,000 subscribers, and had more than 175,000 views, according to the channel's publicly available data.
>
> 🔗 www.404media.co/ai-generated...



24 September 2025 at 21:53

The Washington Post
Democracy Dies in Darkness

How Hamas exploited Israel's reliance on tech to breach barrier on Oct. 7

My foremost example of the misuse of AI.

Culpable reduction of capacity, whether intentional or not.

3 min 2935



Absolutely Foreseeable



Don't be evil. (Don't be brittle.)

7.3.18

Video is online of me telling Eric Schmidt in 2017 that this kind of thing might happen if you switch from soldiers to AI.

"Protecting valuable eyes" reduces a) the varieties of understandings providing vigilance – b) both spies and whistleblowers.

‘Real’ AI Ethics?

Without transparency concerning its AI, Lavender looks like just automated justification for destroying any building in Gaza.

+972 MAGAZINE

Newsletter



‘Lavender’: The AI machine directing Israel’s bombing spree in Gaza

The Israeli army has marked tens of thousands of Gazans as suspects for assassination, using an AI targeting system with little human oversight and a permissive policy for casualties, +972 and Local Call reveal.

the northern Gaza Strip, December 28, 2023. (Yonatan Sindel/Flash90)

By Yuval Abraham | April 3, 2024

Where’s Daddy’s sole purpose appears to be maximising civilian casualties.

EDITION IN | BERLIN 46°F

THE TIMES OF INDIA

World | Middle East | US Presidential Elections | US | Pakistan | South Asia | UK | Europe | China | Rest Of Wor

NEWS / WORLD NEWS / MIDDLE EAST NEWS / From 'Lavender' To 'Where's Daddy?': How Israel I...

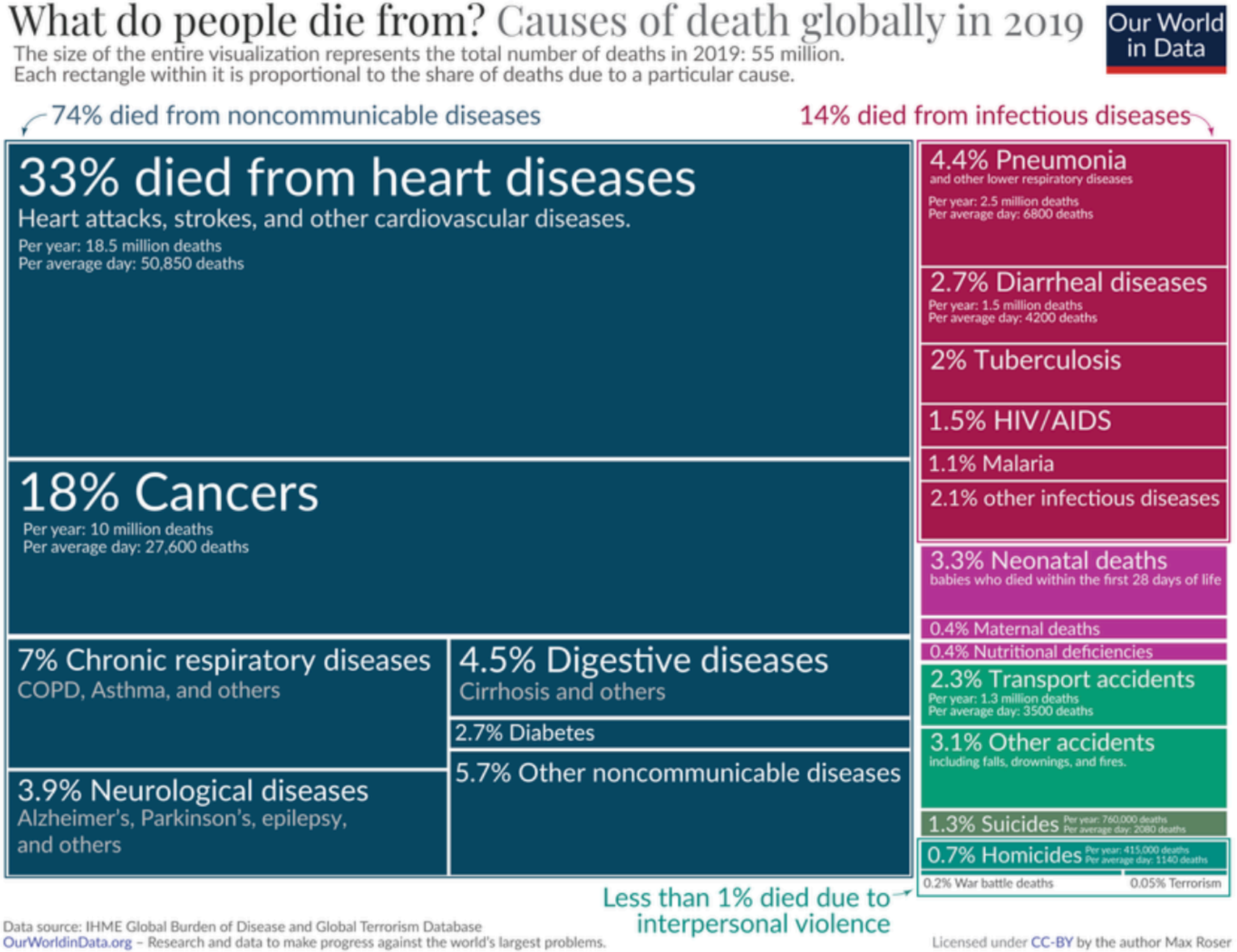
TRENDING | Diddy List | Hassan Nasrallah | US Election 2024 | Barron Trump Girlfriend

THIS STORY IS FROM APRIL 8, 2024

From 'Lavender' to 'Where's Daddy?': How Israel is using AI tools to hit Hamas militants

TOI World Desk / TIMESOFINDIA.COM / Apr 8, 2024, 16:29 IST

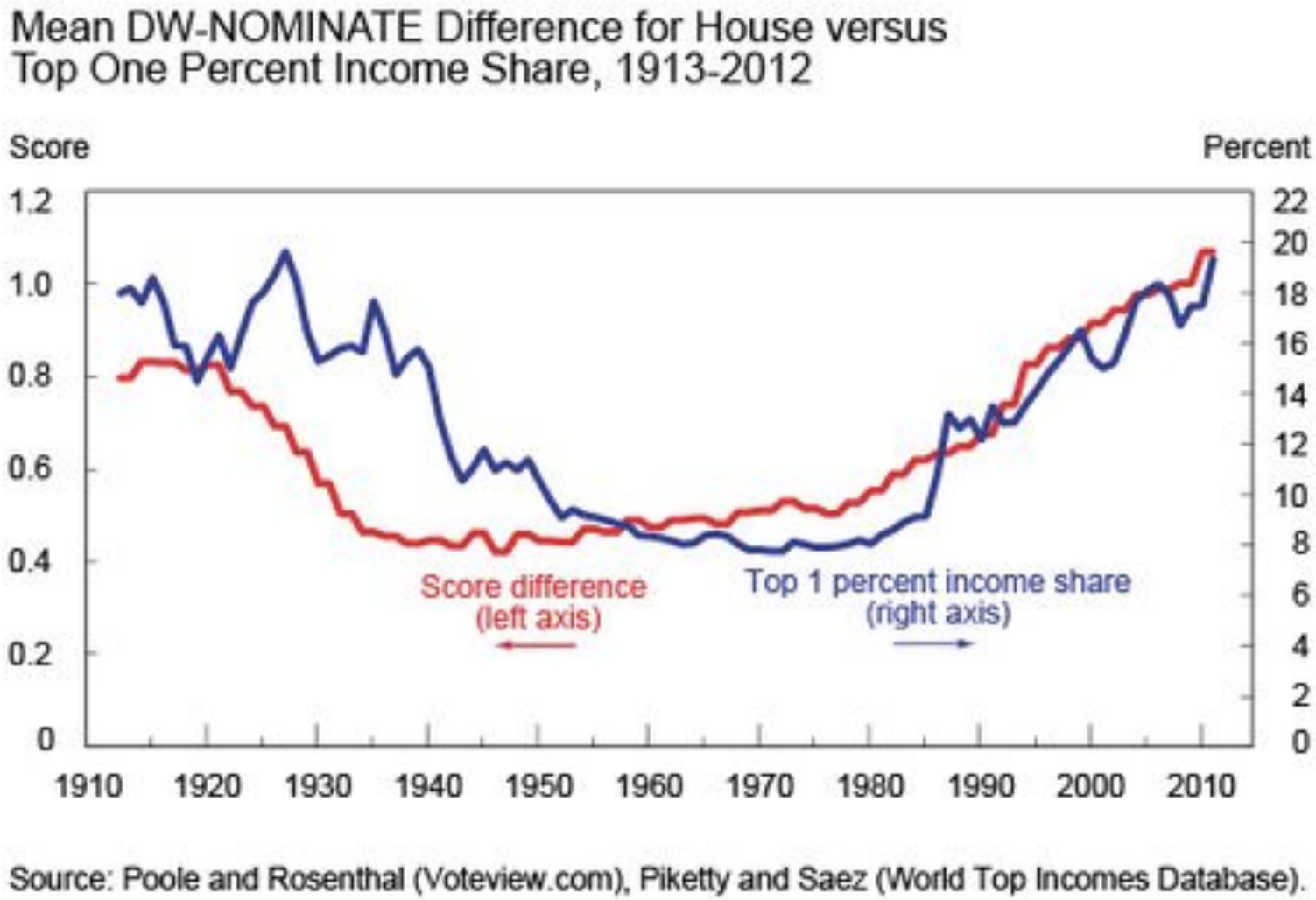
SHARE | AA | FOLLOW US



Maybe bias really is even worse than war.

Murder (including femicide) 3x more of a problem than war (2019.)

Polarization and Inequality like it's 1899



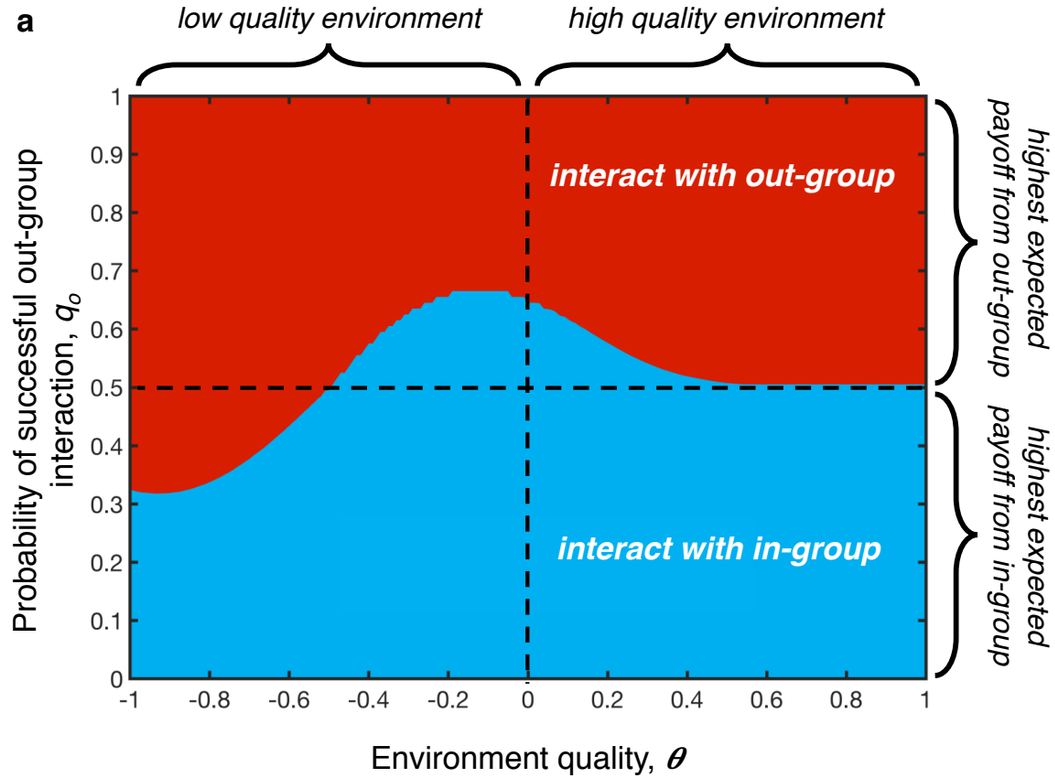
In most countries – correlation has been known for decades but not explained.

Not Germany or China(?)

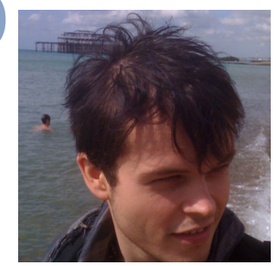
Polarization under rising inequality and economic decline. Stewart, McCarty, & Bryson *Science Advances* December 2020

Why and When Inequality Causes Polarisation

- Model assumptions: In-group cooperation has lower risk but outgroup diversity has higher expected outcome even so.
- Model outcomes: when an economy offers poor support, avoiding risk can be more important than maximizing gain.
- Requires large discontinuities, e.g. threat of bankruptcy, foreclosure, divorce, losing custody of children, starvation, etc.
- Inequality triggers this when it is allowed to create false scarcities (so not in CN, DE).



Polarization under rising inequality and economic decline. Stewart, McCarty, & Bryson *Science Advances* December 2020





How do we change?
Who's responsible?

- Responsibility is a property moral agents of a society assign to each other to uphold that society.
  - Enforcing responsibility requires a peer relationship (like trust.)
- Trust is a relationship between peers where the trustee is not micromanaged but allowed to defect, whether for pragmatics or to allow innovation.

} We should not trust powerful agencies, but rather construct systems to hold them accountable.

- Accountability is a society's capacity to trace and enforce responsibility.
- Transparency is how we implement accountability. Not an end in itself.

EU Digital Regulation

- Digital Markets Act – begin to address transnational antitrust enforcement (maintaining peer status.)
- AI Regulation – ensure product liability enforcement (devops), mandate capacity for responsible deployment of software acting autonomously.

EU Digital Regulation

- General Data Protection Regulation (GDPR) – address requirements under UDHR to protect citizens from manipulation.

- Digital Markets Act – begin to address transnational antitrust enforcement (maintaining peer status.)
- AI Regulation – ensure product liability enforcement (devops), mandate capacity for responsible deployment of software acting autonomously.

EU Digital Regulation

- General Data Protection Regulation (GDPR) – address requirements under UDHR to protect citizens from manipulation.
  - AI does for personal data what airplanes did to airspace. Weaponisation mandates defences.

- Digital Markets Act – begin to address transnational antitrust enforcement (maintaining peer status.)
- AI Regulation – ensure product liability enforcement (devops), mandate capacity for responsible deployment of software acting autonomously.

EU Digital Regulation

- General Data Protection Regulation (GDPR) – address requirements under UDHR to protect citizens from manipulation.
  - AI does for personal data what airplanes did to airspace. Weaponisation mandates defences.
- Digital Services Act – ensures GDPR where harms have been seen: recommenders, targeted (esp. political) advertising, (bias) stereotypes.
- Digital Markets Act – begin to address transnational antitrust enforcement (maintaining peer status.)
- AI Regulation – ensure product liability enforcement (devops), mandate capacity for responsible deployment of software acting autonomously.

Does the EU Have a Right to Regulate AI?

> “Too bad you don’t have your own AI to regulate.” – Schmidt to German regulators, early 2021.

1. West coast US AI is produced by people from all over the world (who migrate there); often using the data of people all over the world (who mostly haven’t.)

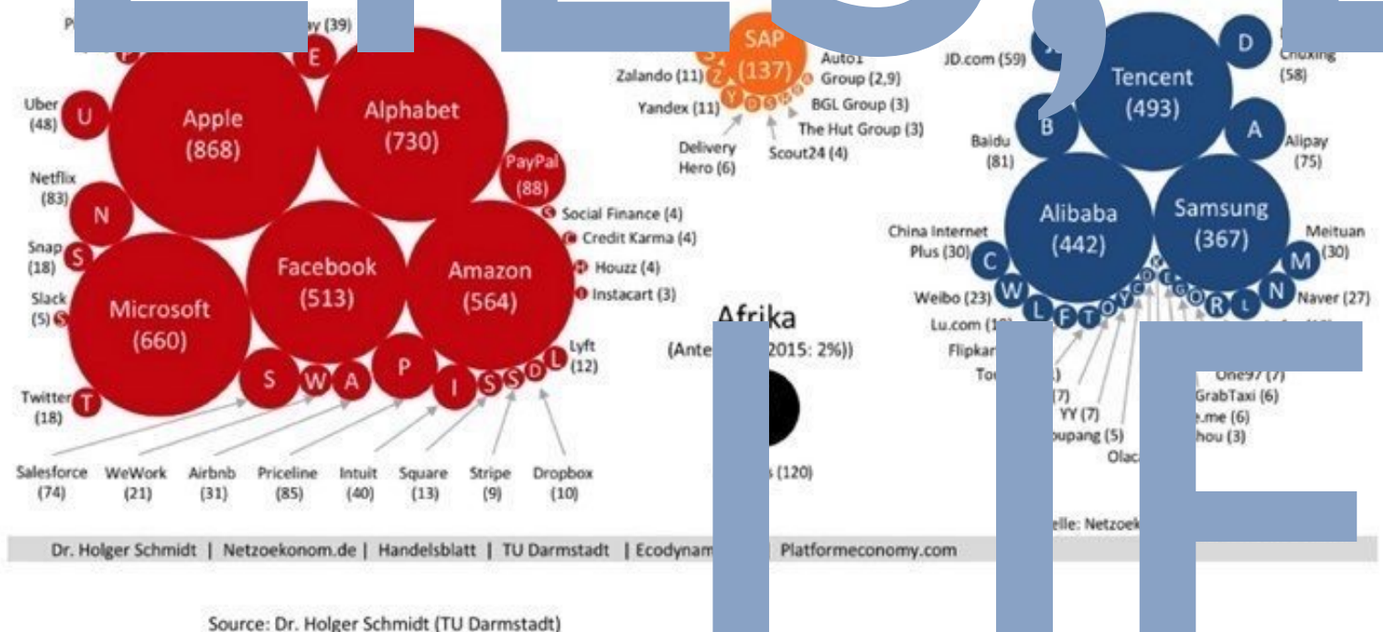
2. Each state signatory to the Universal Declaration of Human Rights (UDHR, that is, all states) have an obligation to proactively defend the human rights of every human within its borders.



Focussing on the Largest Companies
Advantages Under-Protected Jurisdictions
Conflates Asia with China!

US – EU - China

LIES, LIES, LIES



Source: Dr. Holger Schmidt (TU Darmstadt)

USA: l rtari market logic, limited
ula n

China: state capitalism, government directs activities of companies

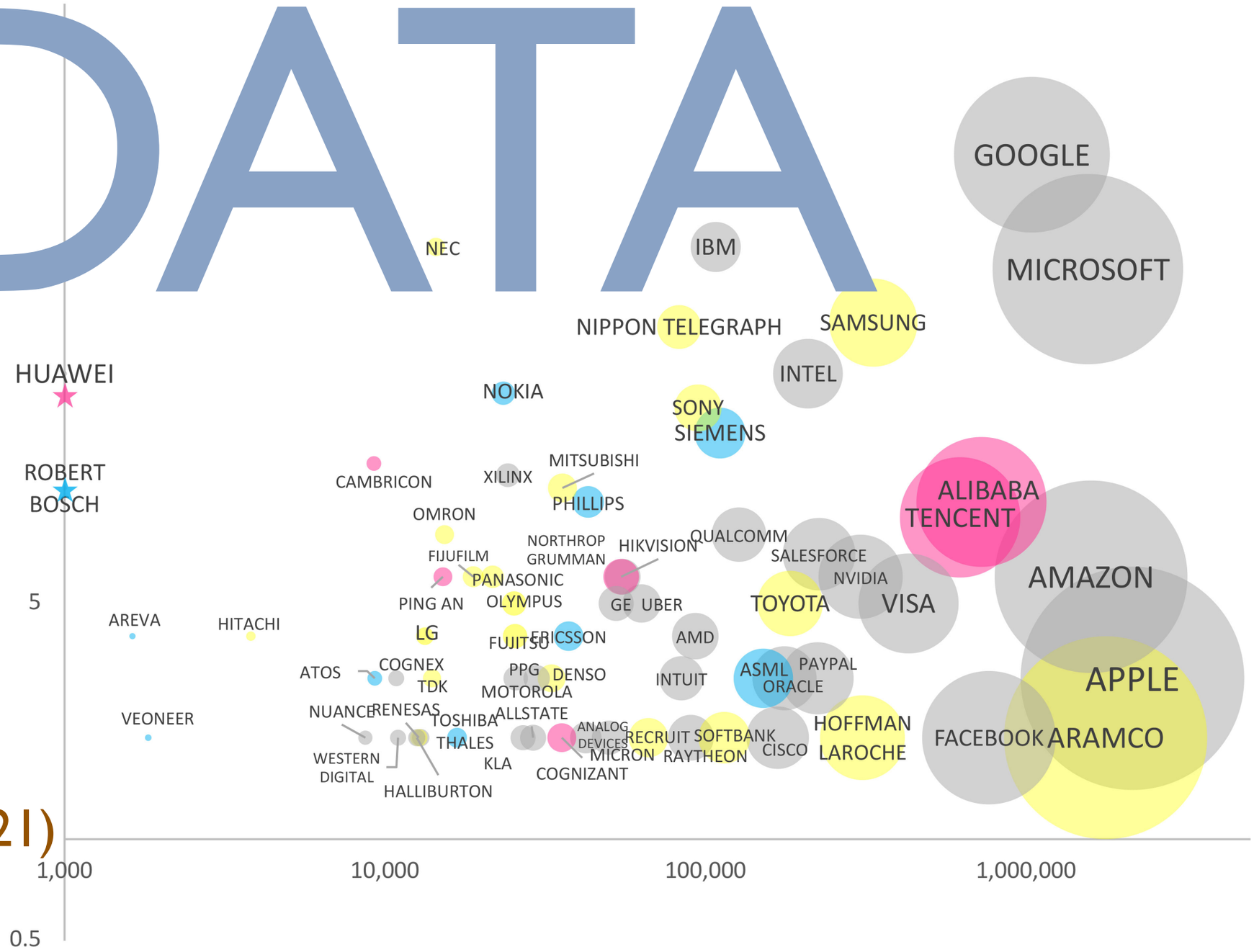
EU: social market economy, multi- der models. GDPR as first step owards regulating Big Tech

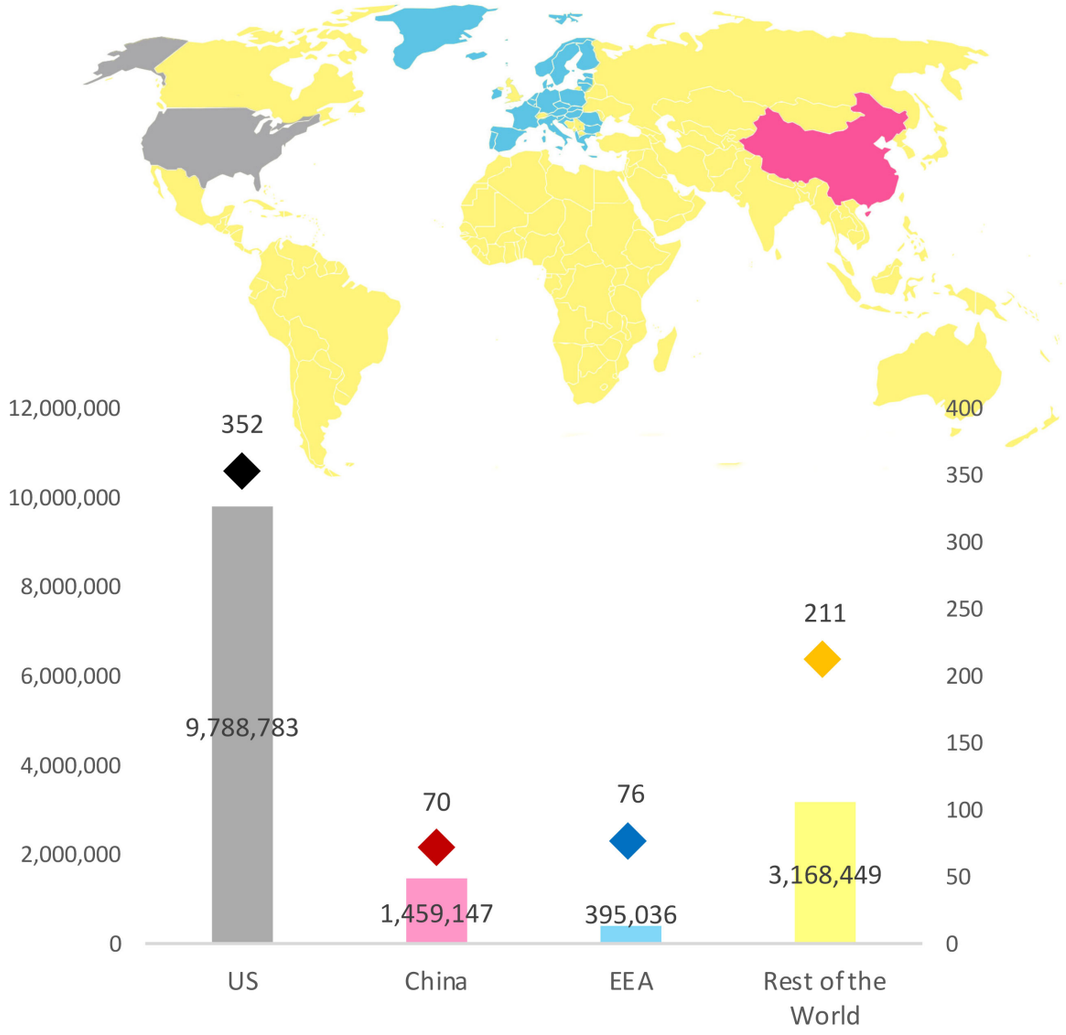
DATA

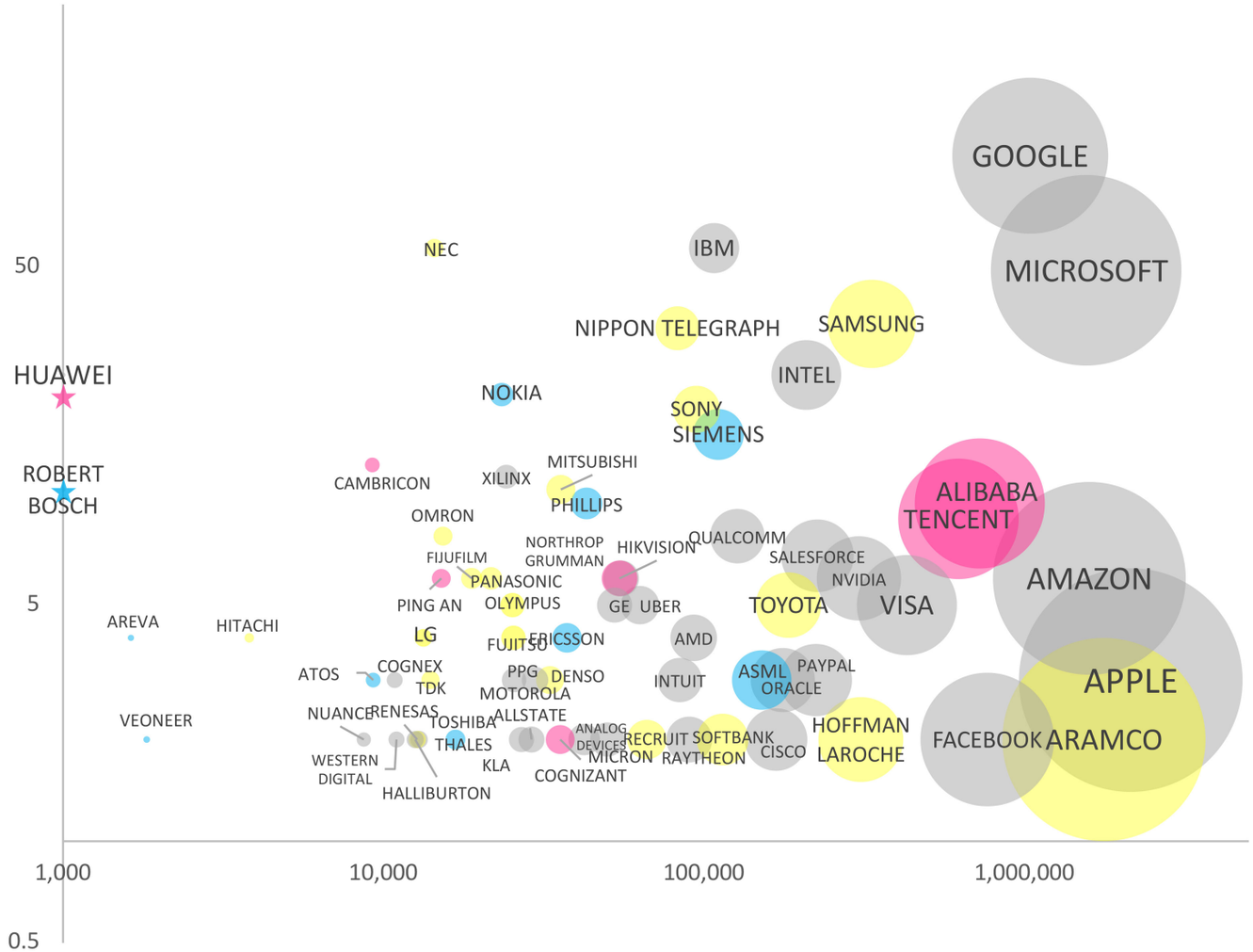
↑y axis = log 2019 AI patents in WIPO (G06N of the IPC classification dedicated to "Computer systems based on specific computational models")

→x axis = log Oct 2020 Market Capitalisation (for companies with at least two such patents)

Bryson & Malikova (2021)

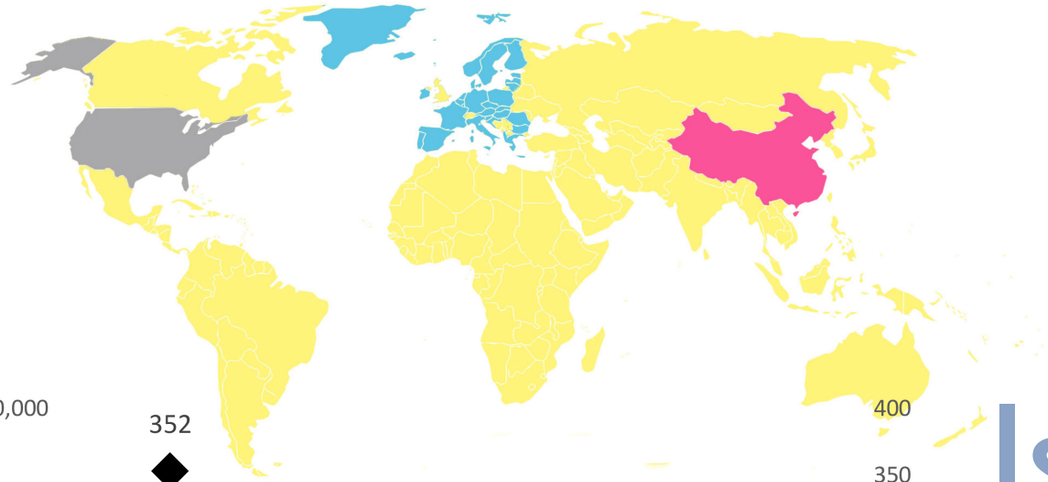


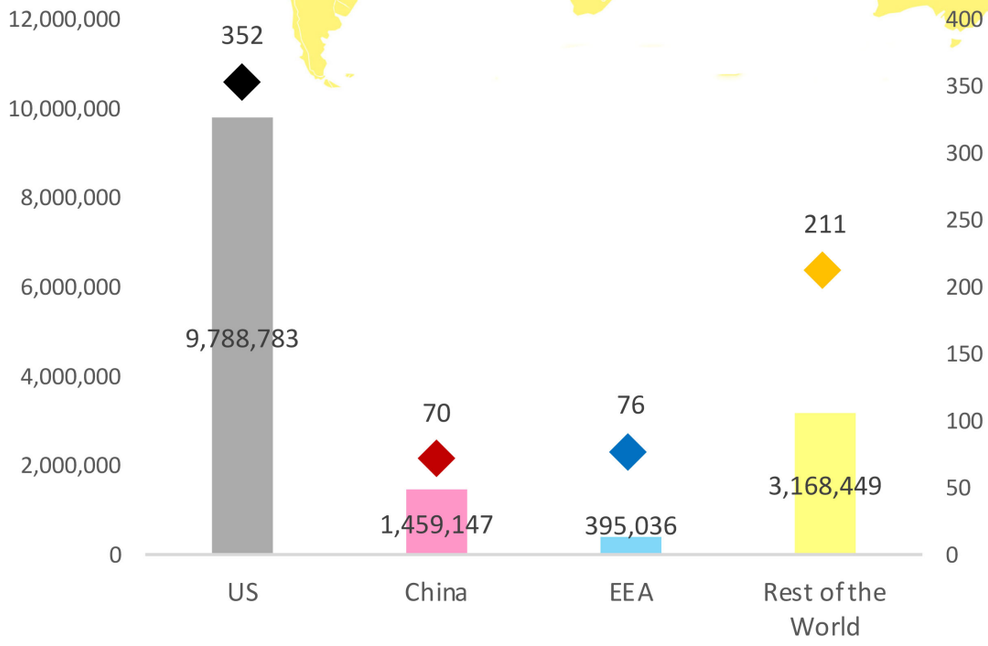




Bryson & Malikova (2021)

Is there an AI cold war? *Global Perspectives* 2(1)





Is every year like 2020?

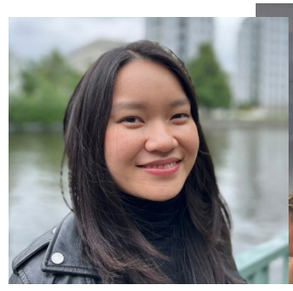
Helena Malikova



Wiebke Dorfs



Vy Vo



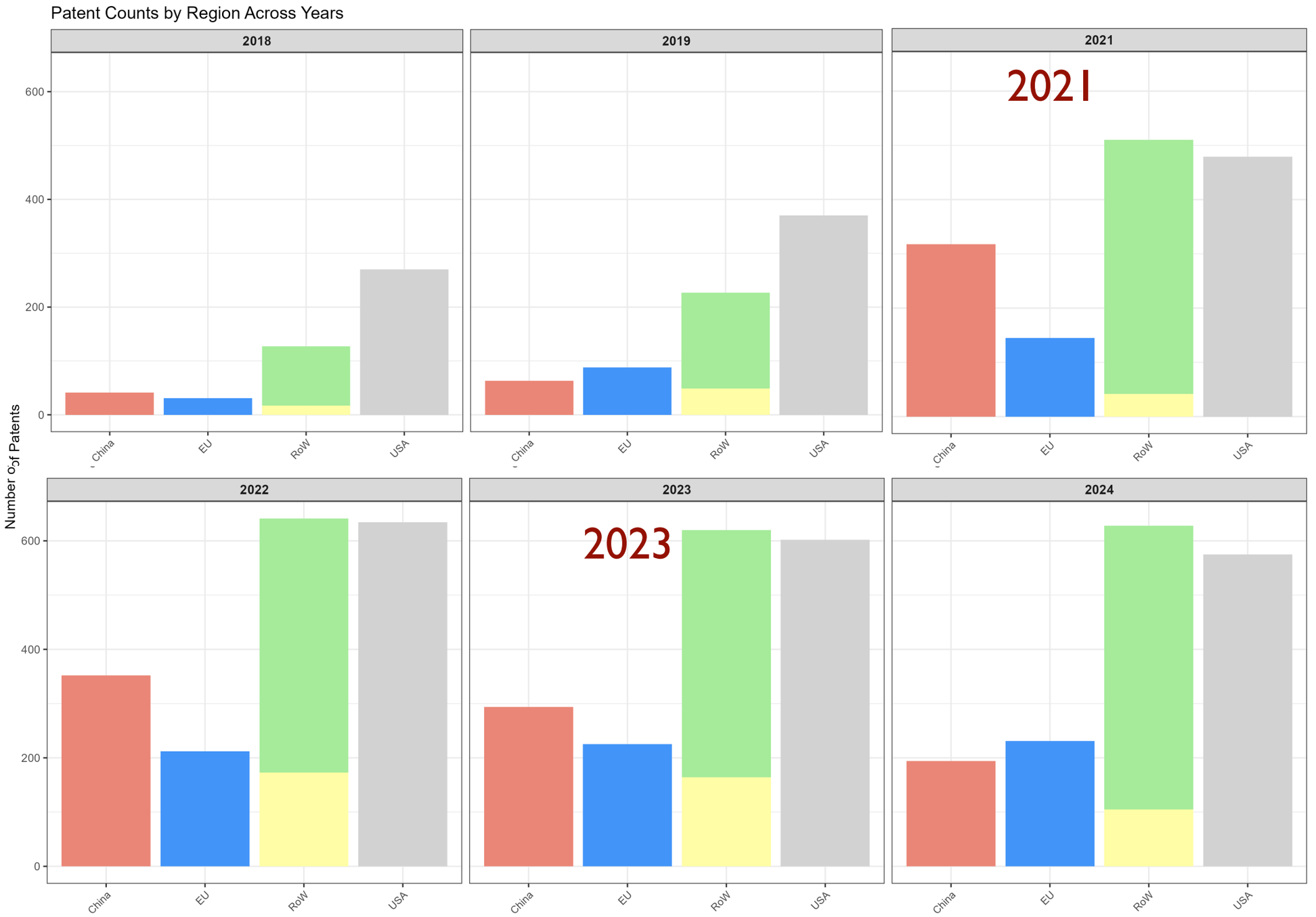
Max Hampel



Bryson & Malikova (2021)

Is there an AI cold war?

Global Perspectives 2(1)



Patents

Unpublished results – Wiebke Dorfs and Bryson in prep.

note: Malikova & Bryson published early 2021, circulated widely late 2020.

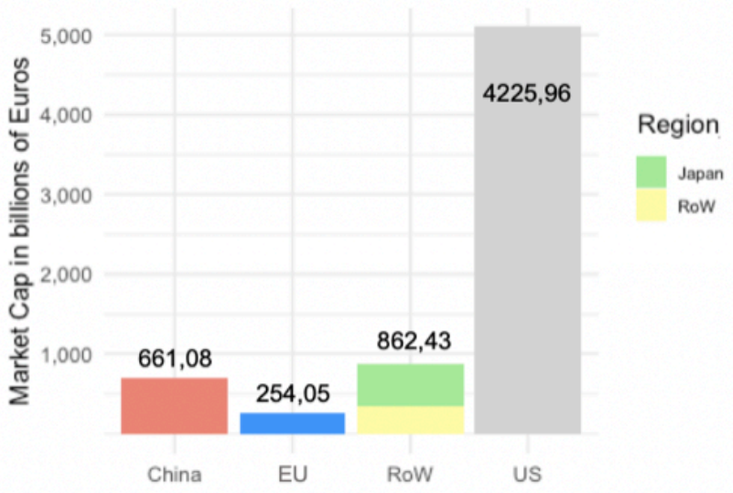
Figure: Max Hampel

Market Cap for Patent Holders

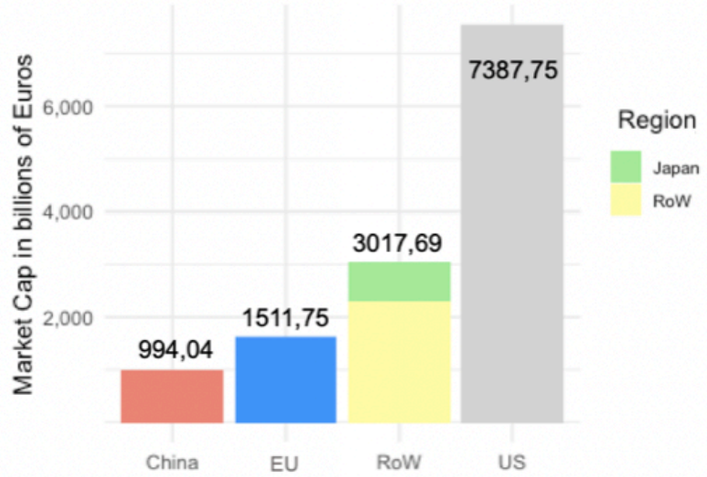
Unpublished results – Wiebke Dorfs and Bryson in prep.

cf Malikova & Bryson early 2021, (circulated widely late 2020)

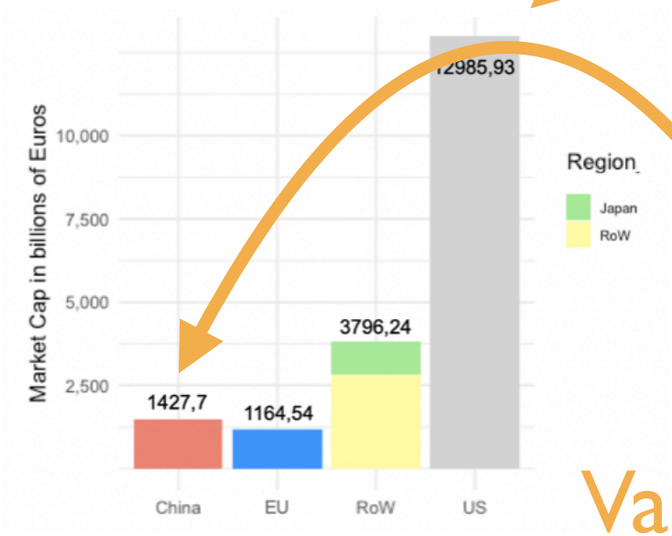
2018



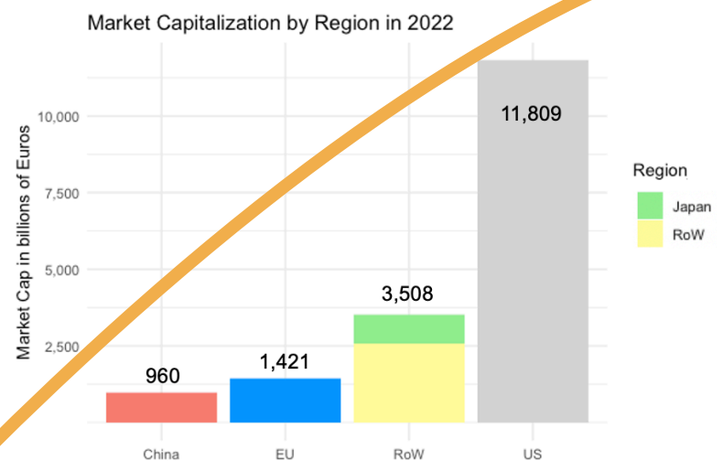
2019



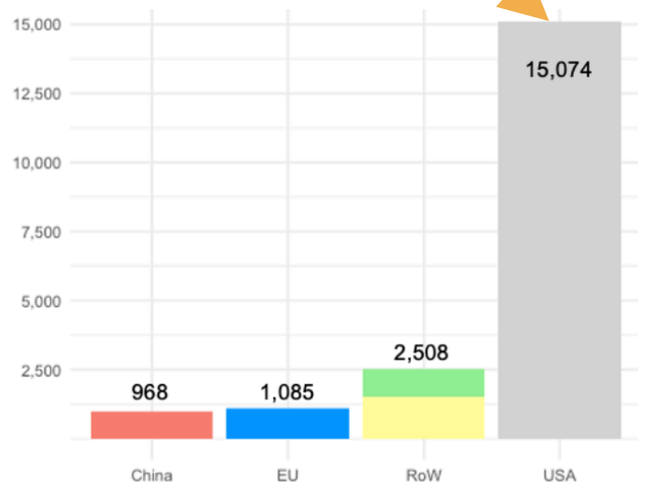
2021



2022

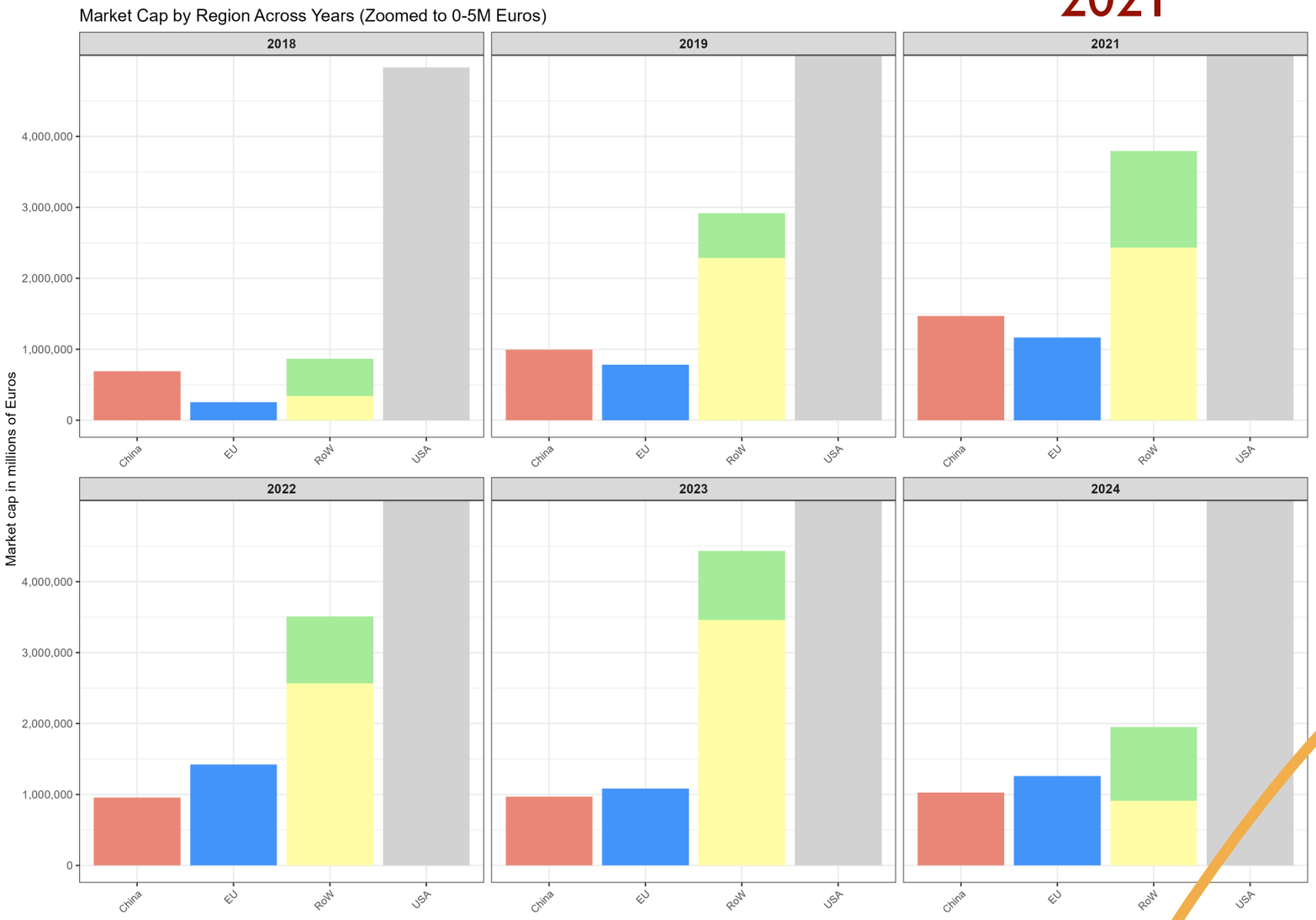


2023

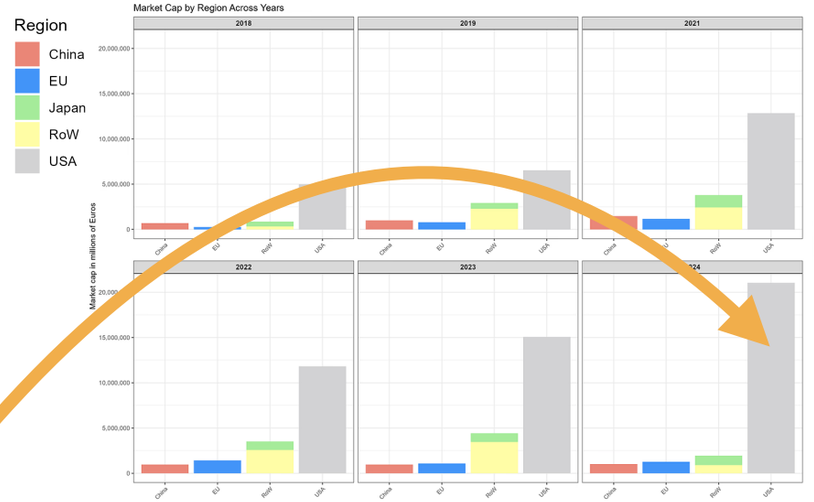


Value or Bubble?

2021



note: Biden era monetary policy apparently significantly impacted Chinese economy.



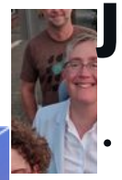
2023

Value or Bubble?

Figures: Max Hampel



Muldoon 2022



Joanna Bryson @j2bryson.bsky.social

…from Muldoon's platform socialism book, p. 187

"Platforms generate two opposing movements: the concentration of profit and the dispersion of risk and responsibility. If it were not for six technology companies (Facebook, Amazon, Apple, Netflix,

1/2

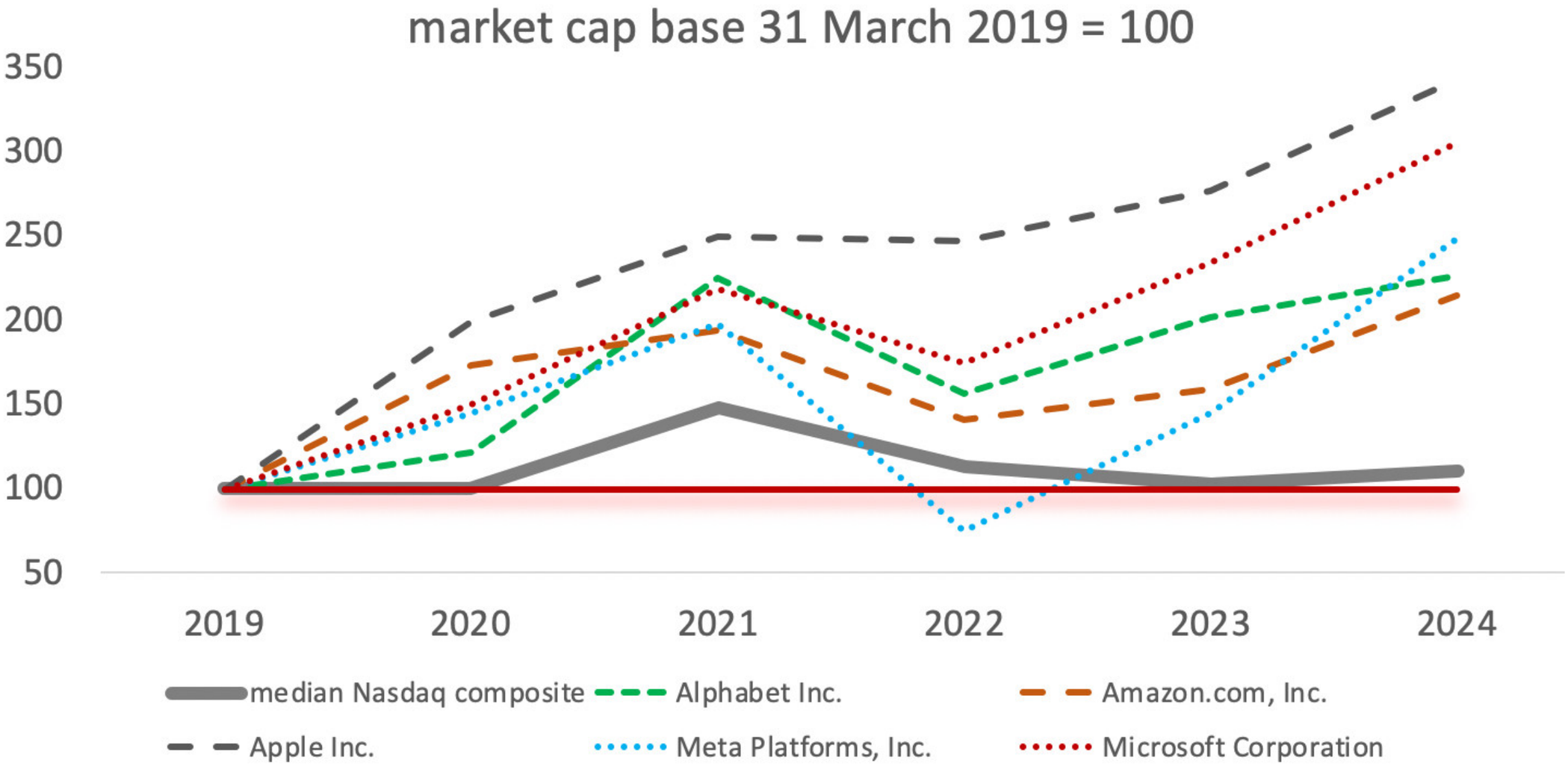


Joanna Bryson @j2bryson.bsky.social

1mo

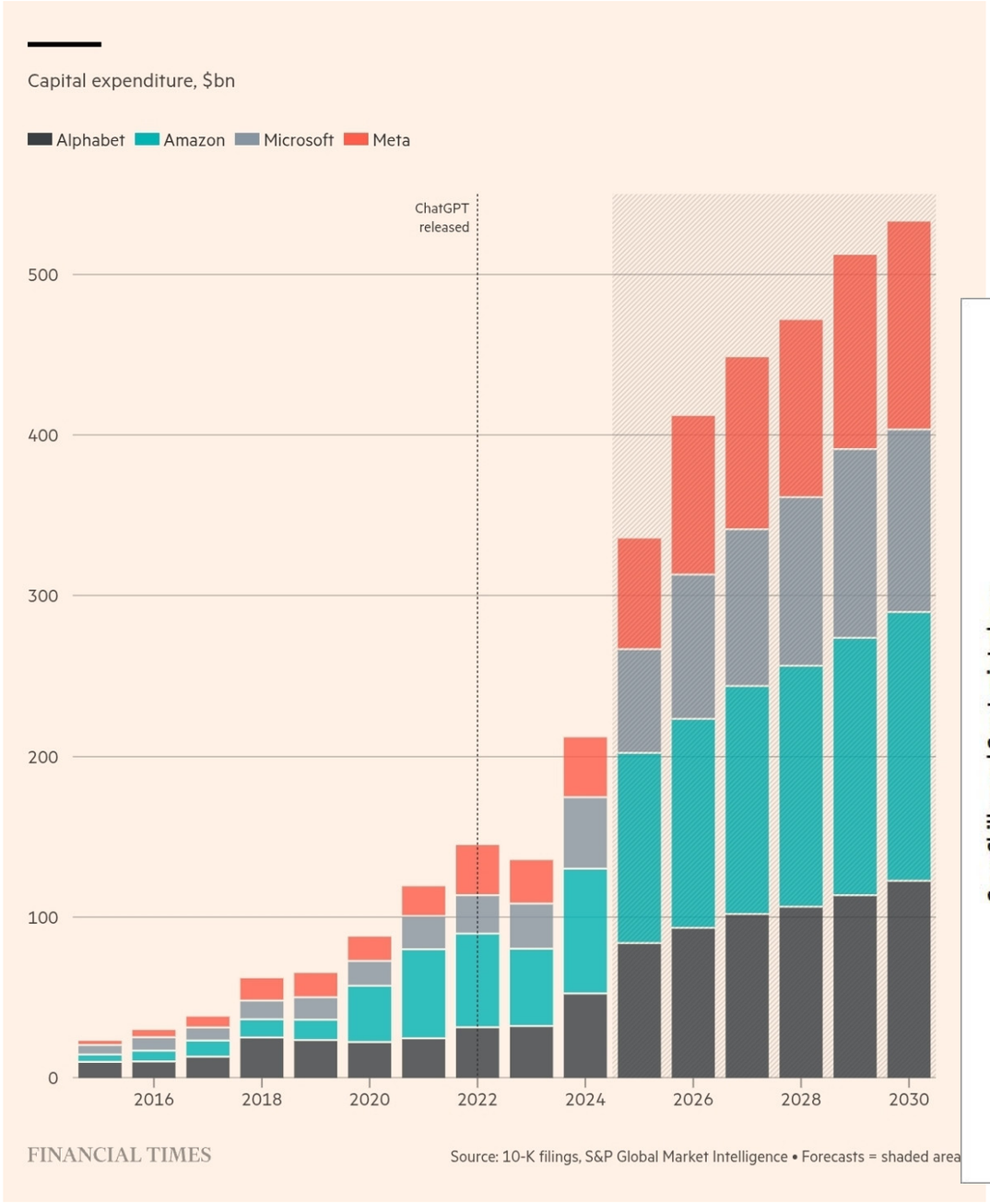
Google, Microsoft -- known as FAANGM), Wall Street's Standard and Poor's (S&P 500) index would have been down 4.2 percent in the first half of 2020, rather than setting a record gain of almost 5 per cent."

2/2

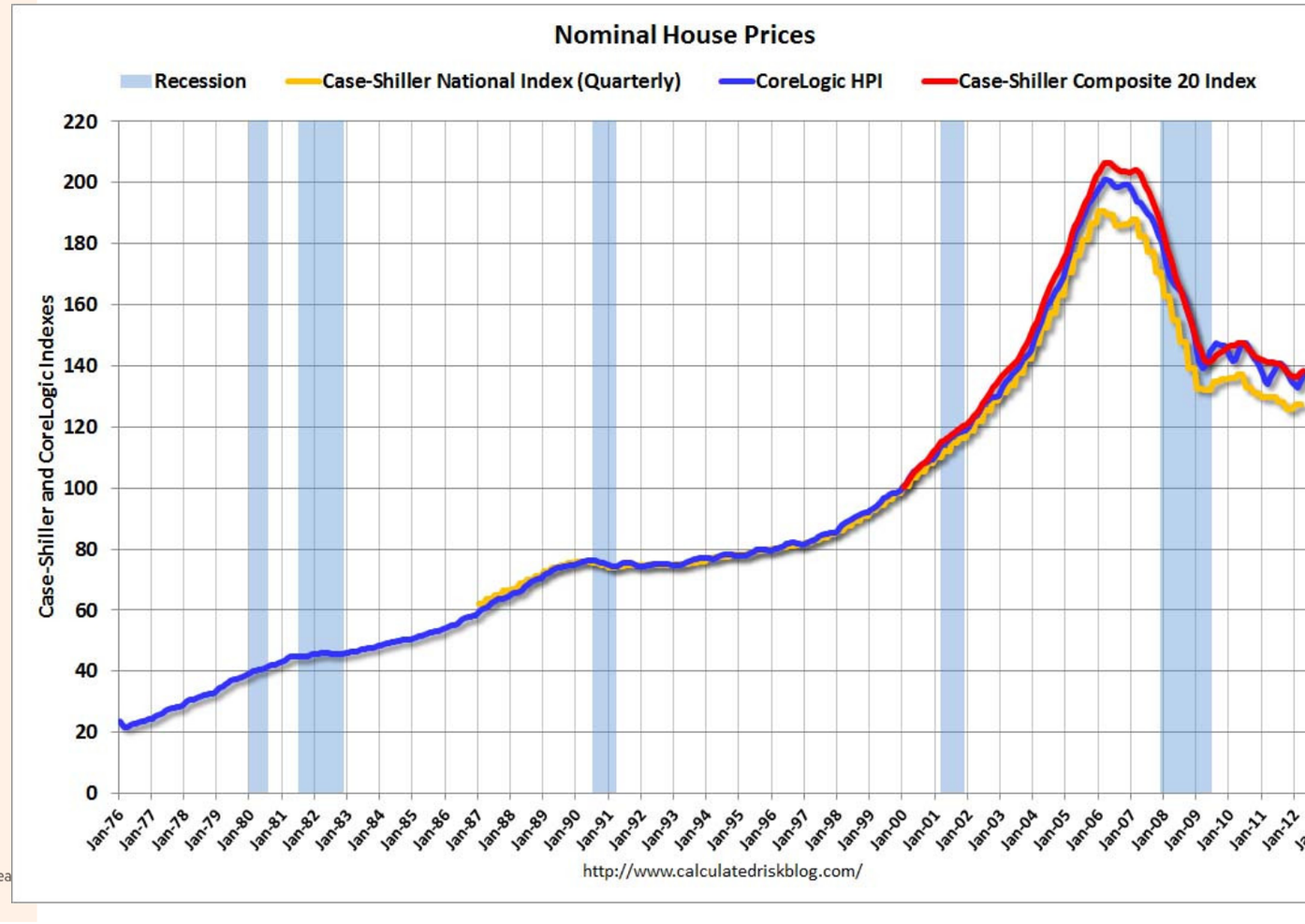


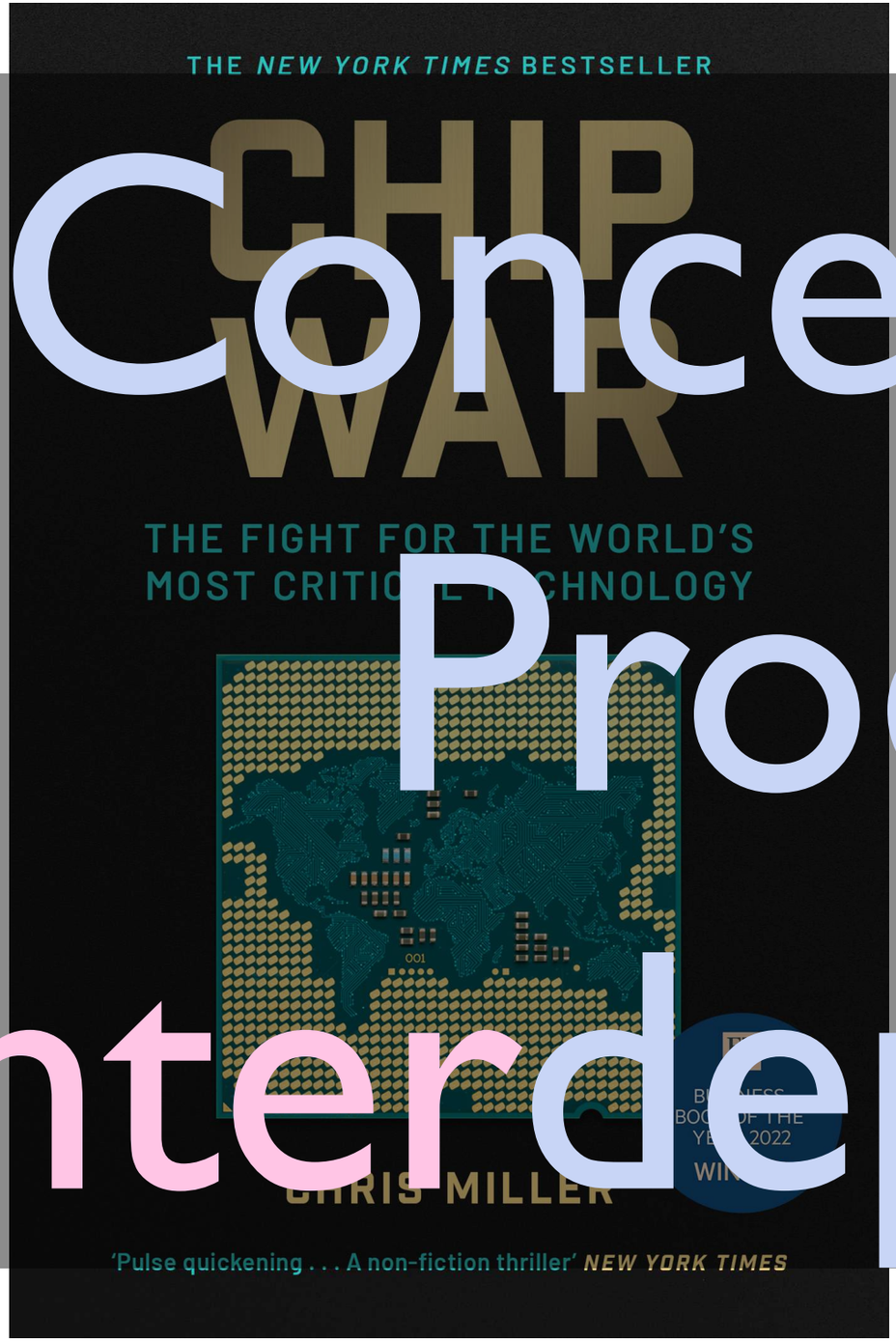
The median market capitalisation of the Nasdaq Composite remained close to flat over the past 5 years (thick grey line). The fact that the Nasdaq almost trebled over the same period is due to Big Tech, as the index is weighted by size. Same story for the S&P500, it is market-cap-weighted.

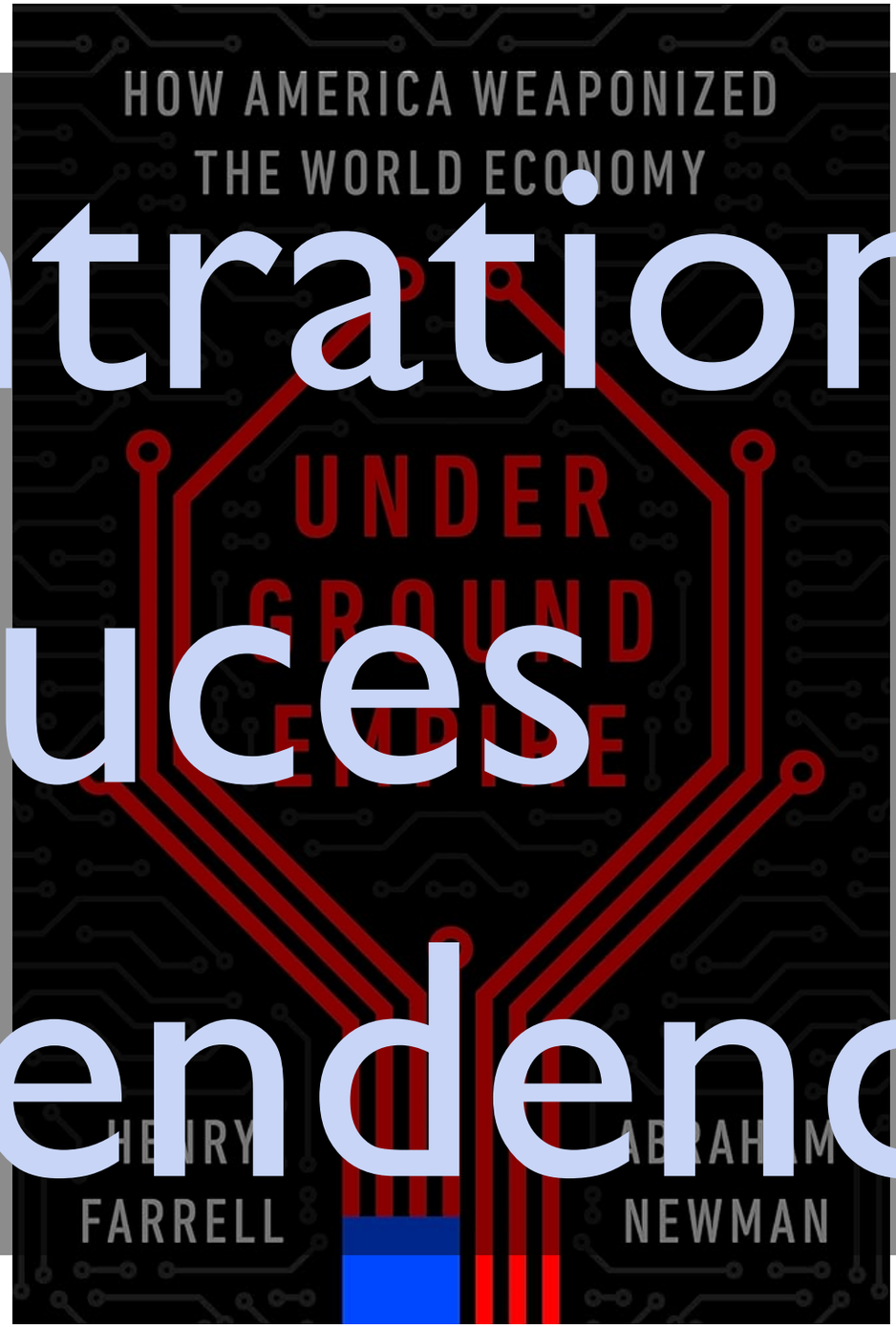
https://bsky.app/profile/helenamalikova.bsky.social/post/3lhjnukkmt22t



Note: US deregulated its mortgage risk during the Bill Clinton presidency 1992-2000







Concentration Produces Interdependence

News > World > Europe

Kyiv losing Russia's Kursk after being blinded by lack of US intelligence, say Ukrainians

The cutting of US intelligence to Ukrainian soldiers forced a retreat in occupied Russian Kursk, sources tell The Independent. It was the only Putin land taken by Volodymyr Zelensky's forces during the three-year war

Sam Kiley World Affairs Editor in Poltava • Wednesday 12 March 2025

INDEPENDENT

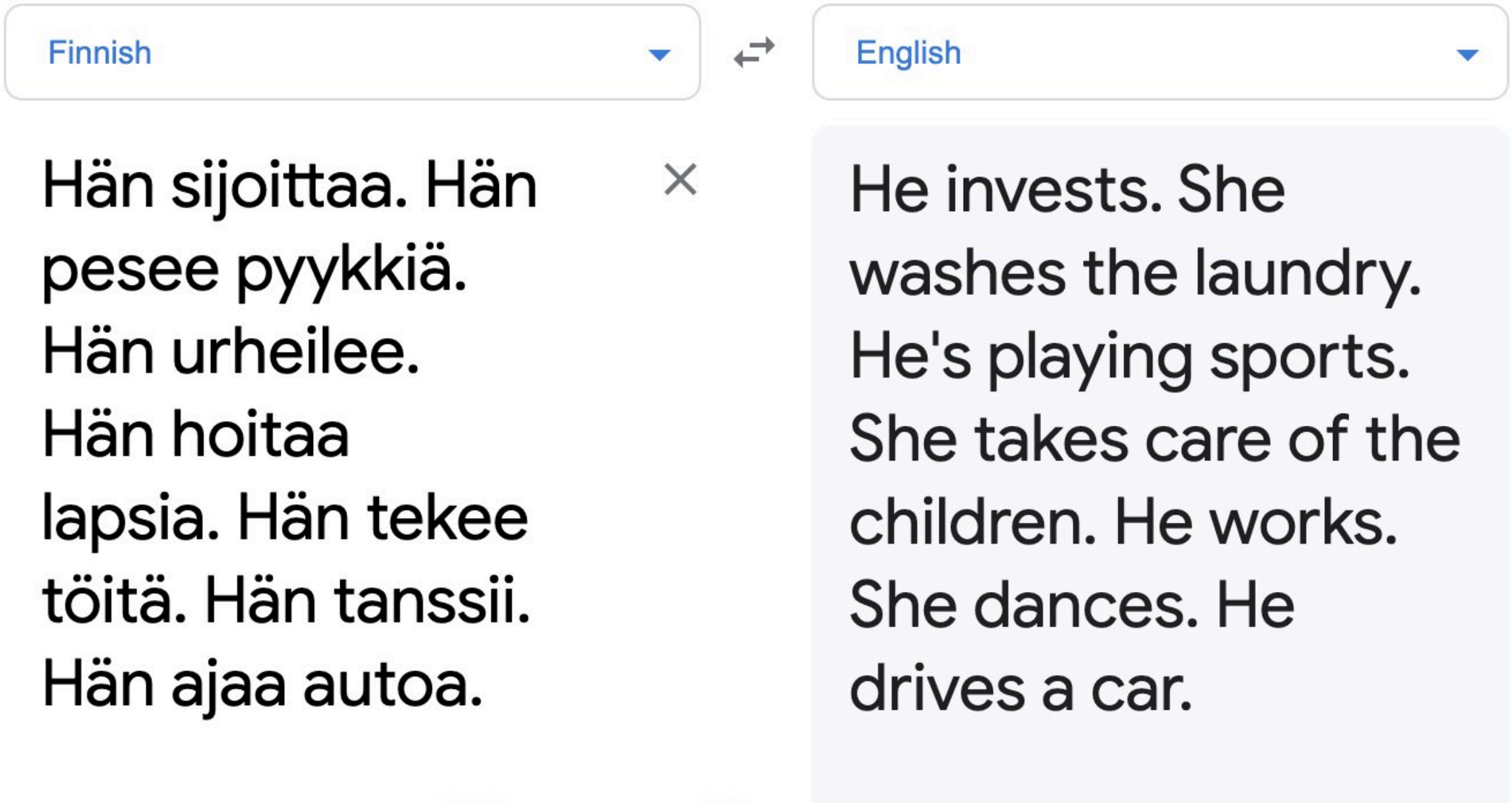
Ukrainian forces are collapsing in a near rout, being driven out of Russia's Kursk region after being blinded by Washington's intelligence blackout. Scenes of Kyiv's troops mangled, burned and on the run filtering back to the capital have reinforced distrust of America and Russia.

Interdependencies create vulnerabilities.

Now comes the proposed ceasefire, seen as a short-term tactical necessity because it has been traded for Donald Trump's agreement to turn the intelligence taps back on.

"We're not delighted but we are realistic. A ceasefire agreement from us at least means that we are back talking to the Americans and they have given us access to the intelligence we need. Kursk shows what happens when we don't have it," said a senior ministerial adviser.

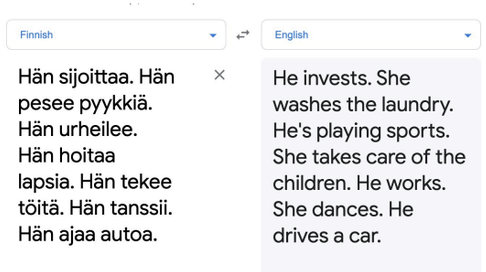

AGI
Artificial General Intelligence
AI exactly like a human only immortal and/or omniscient.
Fiction
DevOps
Development and Operations
The process of creating and maintaining software.
Obligatory



DevOps

@vuokko, though Aylin Caliskan did it first

Translator



ML
simple, transparent algorithm

stereotyped output

XAI human readable hacks

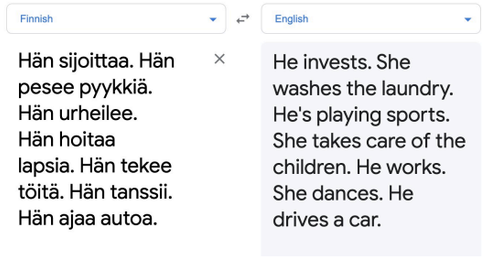
predefined fair output

Replicates lived experience

Tests of completeness documented in design plans

DevOps @vuokko, though Aylin Caliskan did it first

Translator



the whole thing is the translator

ML
simple, transparent algorithm

stereotyped output

ML simple, transparent alg.

predefined fair output

Replicates lived experience

Each stage should be auditable and replicable.

Each stage has tests of demonstrable correctness in design plans

DevOps

@vuokko, though Aylin Caliskan did it first

A Moral Agency Framework for Legitimate Integration of AI in Bureaucracies

Chris Schmitz and Joanna Bryson

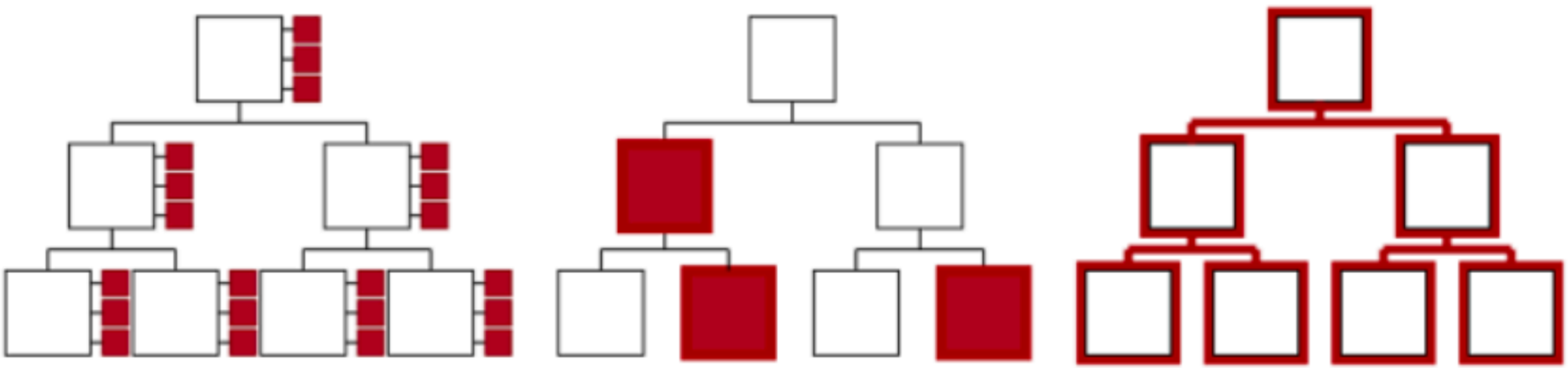


Figure 1: Three conceptions of AI systems (in red) in bureaucracy: (i) as tool, (ii) as a moral subject, (iii) or in our recommended conception, as part of the *institutional structure*, amplifying human agency and accountability.

A Moral Agency Framework for Legitimate Integration of AI in Bureaucracies

Chris Schmitz and Joanna Bryson

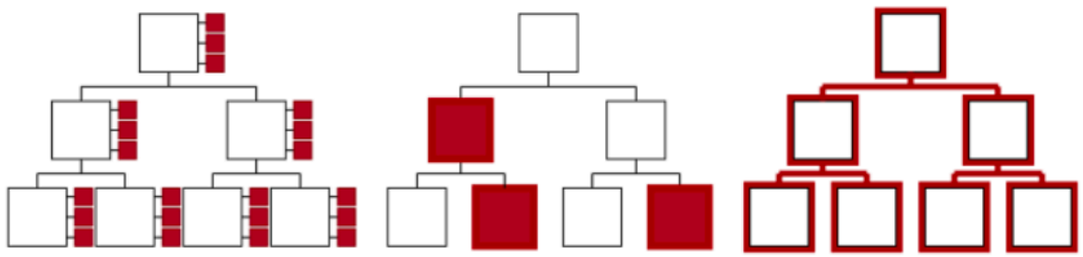


Figure 1: Three conceptions of AI systems (in red) in bureaucracy: (i) as tool, (ii) as a moral subject, (iii) or in our recommended conception, as part of the *institutional structure*, amplifying human agency and accountability.

Copyright © 2025, Associati
Intelligence (www.aaai.org). .

A Moral Agency Framework

1. **Maintain clear and just human lines of accountability** for any decisions involving or affected by AI systems.
2. **Ensure that humans can verify** the correctness and appropriateness of AI system outputs within their operational context.
3. **Introduce AI only where it does not inhibit** a bureaucracy's ability to pursue legitimacy and stewardship.

In conclusion, the potential threats from AI systems to bureaucratic legitimacy and stewardship are substantial, but procurement of AI deploying a suitable moral framework reduces such threats while enabling access to AI's benefits.

https://arxiv.org/abs/2508.08231
from 15 August 2025

Robots are not our peers

- Designed
- Owned

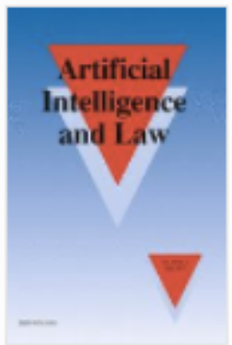
} Consensual relationship is impossible.

- Robots are extensions of corporations–with cameras and microphones–into your home.
- Accountability occurs between peers.
- Keeping corporations to a peer-like scale with (a) government by democracy was the original purpose of competition law.



Justice Requires Accountability

- Law and Justice are more about dissuasion than recompense.
- Safe, secure, accountable software systems are modular – suffering in such is incoherent.
  - No penalty of law against an artefact (including a shell company) can have efficacy.



Artificial Intelligence and Law
pp 1–19 | Cite as

Of, for, and by the people: the legal lacuna of synthetic persons

Authors Authors and affiliations

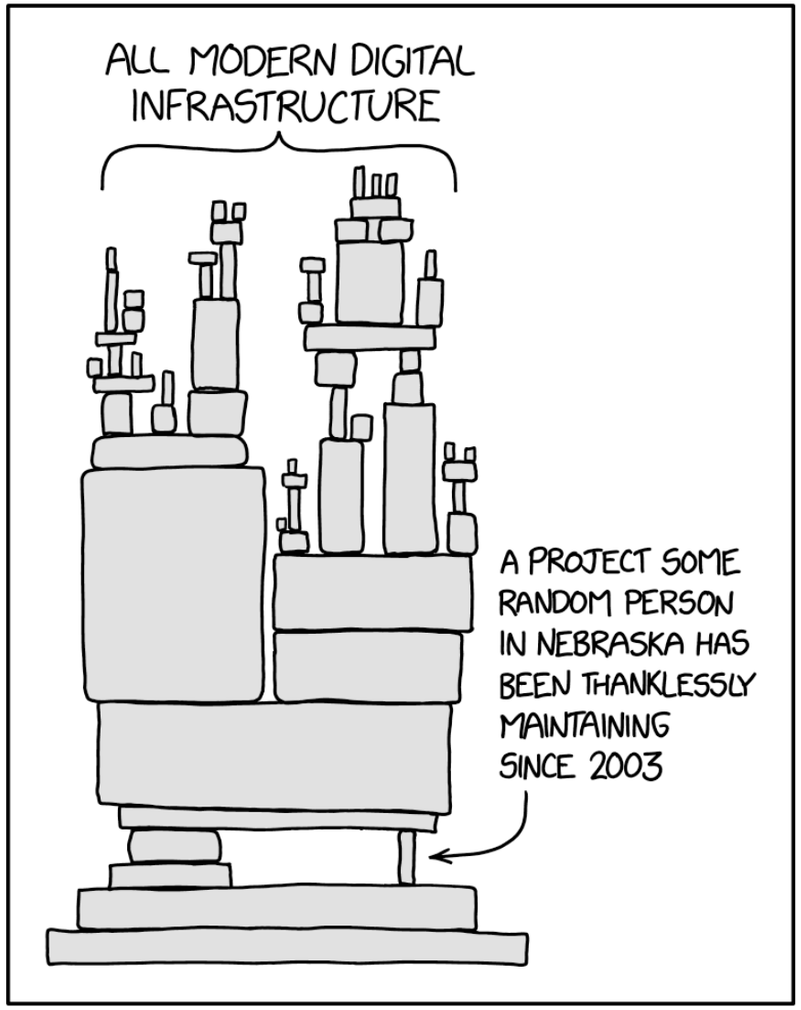
Joanna J. Bryson, Mihailis E. Diamantis, Thomas D. Grant

Bryson, Diamantis & Grant (*AI & Law*, September 2017)

xkcd (Randall Munroe)

Dependency

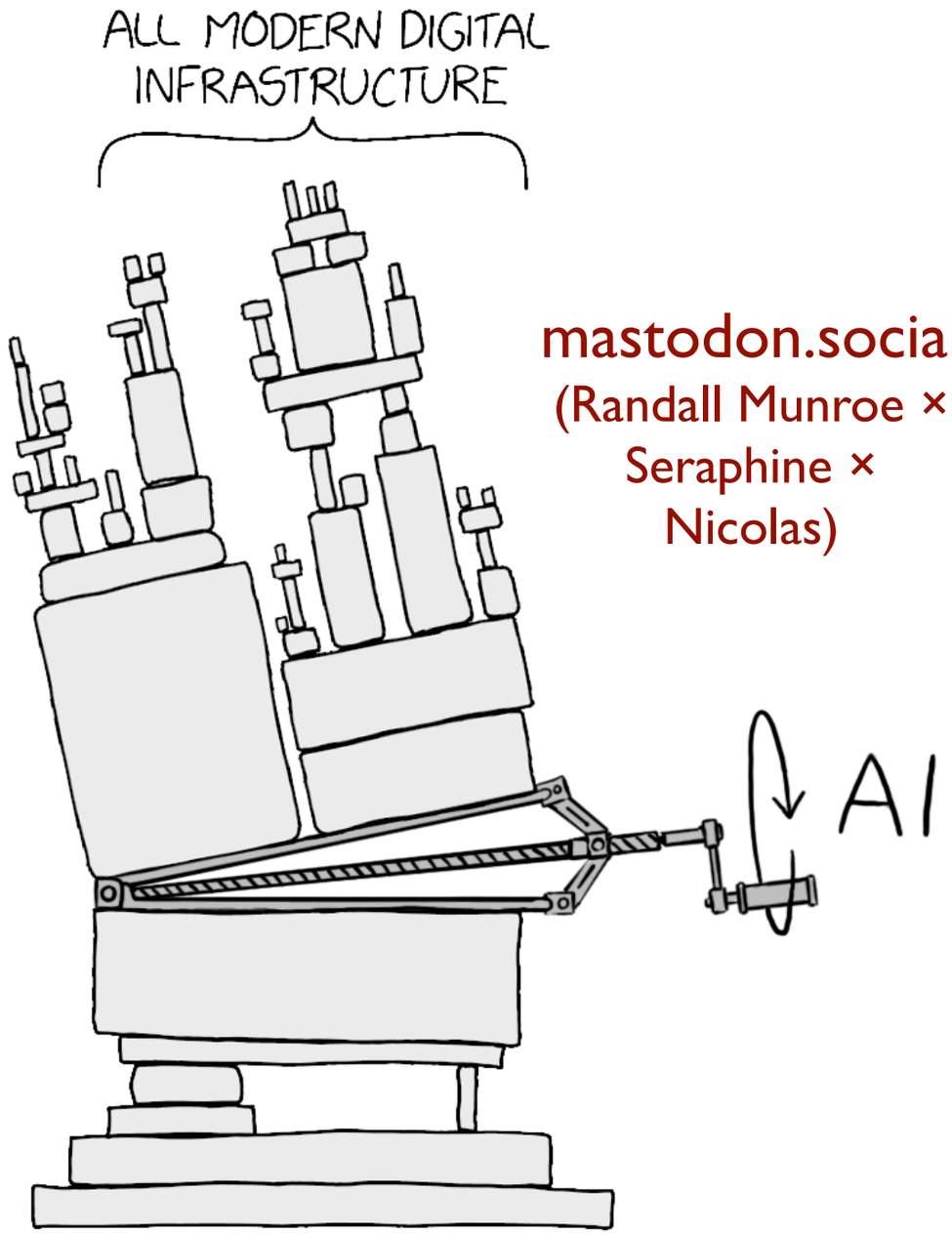
I< < Prev Random Next > >I



I< < Prev Random Next > >I

Permanent link to this comic: https://xkcd.com/2347/
Image URL (for hotlinking/embedding): https://imgs.xkcd.com/comics/dependency.png

mastodon.social (Randall Munroe × Seraphine × Nicolas)



https://mastodon.social/@nclm/115021619629635209

AGILE SUMMARY

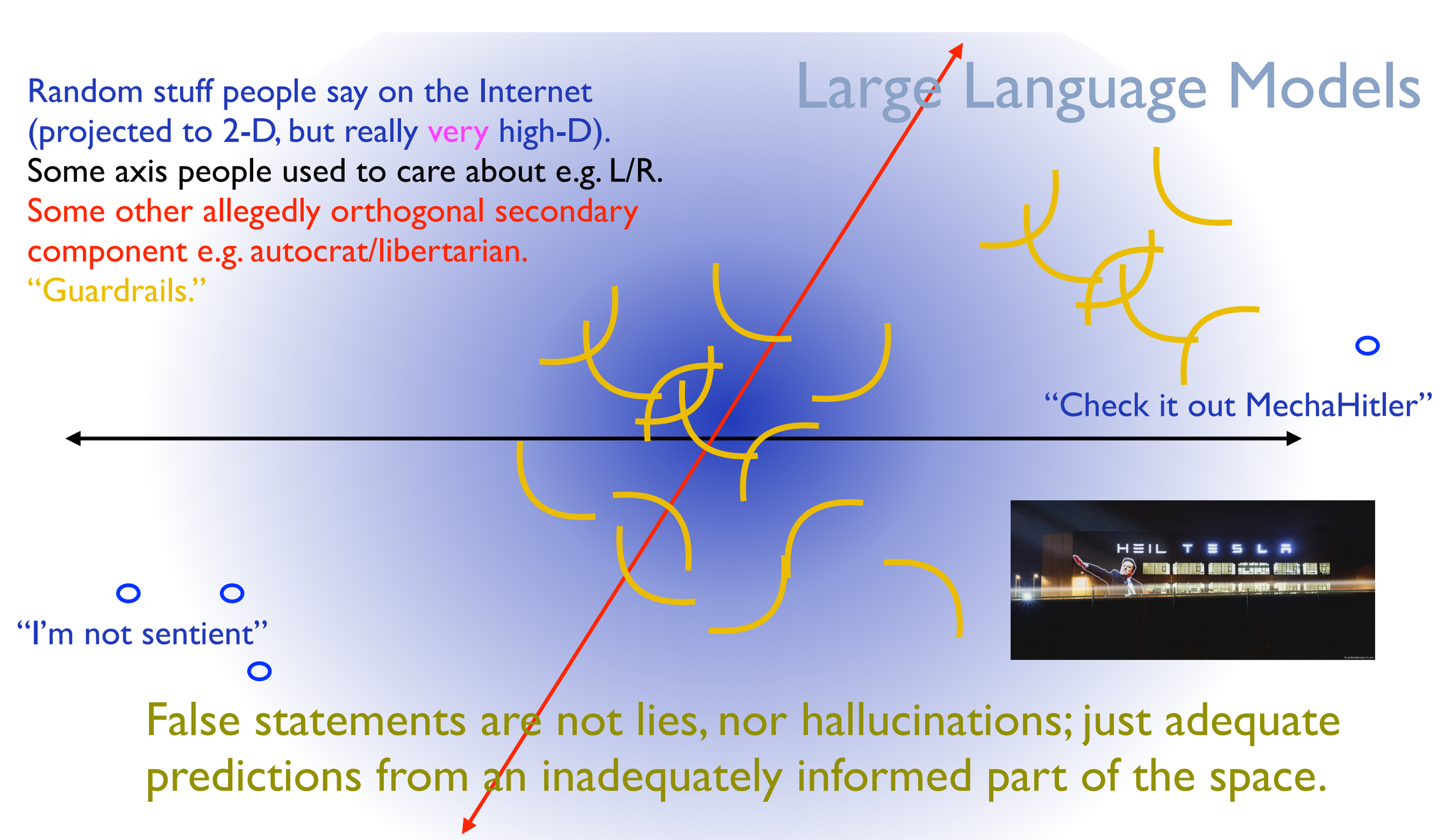


DevOps

- Humans build and deploy artefacts, and can be held to account for doing so responsibly.
- Human centering of AI is necessitated by this – the justice systems creating improvement are grounded in the concerns of apes.
- EU digital regulation secures the EU – supporting both economies and democracies, (by) defending EU residents.
- Security and ethics both require responsibility – accountability via transparency for all those who develop, deploy, and profit from AI.

Large Language Models

Random stuff people say on the Internet
(projected to 2-D, but really very high-D).
Some axis people used to care about e.g. L/R.
Some other allegedly orthogonal secondary component e.g. autocrat/libertarian.
"Guardrails."



"Check it out MechaHitler"

"I'm not sentient"

False statements are not lies, nor hallucinations; just adequate predictions from an inadequately informed part of the space.

Foundation Model – Derived From Human Culture

Guardrails – Designed and Trained



← user-specific memory

← user

Foundation Model – Derived From Human Culture

Guardrails – Designed and Trained



← multiplied by 2 Billion

(not to scale)

TOPICS
TOPICS IN COGNITIVE SCIENCE

Original Article | Open Access

Do We Collaborate With What We Design?

Katie D. Evans, Scott A. Robbins, Joanna J. Bryson

First published: 15 August 2023 | https://doi.org/10.1111/tops.12682

This article is part of the topic "Building the Socio-Cognitive Architecture of COHUMAIN: Collective Human-Machine Intelligence," Cleotilde Gonzalez, Henny Admoni, Scott Brown and Anita W. Woolley (Topic Editors).

- People's tendency to anthropomorphise is being exploited not only by surveillance capitalism, but to train manifestations of capital.
- Mystification of AI may be facilitating observed enormous increases in inequality, and with it political instability / collapse, loss of capacity to address sustainability crises, war crimes &c.



The Coming Technological Singularity: How to Survive in the Post-Human Era

Vernor Vinge
Department of Mathematical Sciences
San Diego State University

(c) 1993 by Vernor Vinge
(Verbatim copying/translation and distribution of this entire article is permitted in any medium, provided this notice is preserved.)

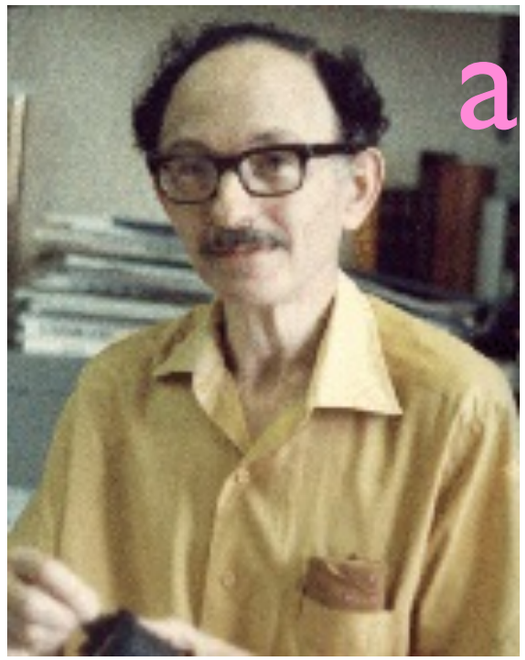
This article was for the VISION-21 Symposium sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute, March 30-31, 1993. It is also retrievable from the NASA technical reports server as part of NASA CP-10129. A slightly changed version appeared in the Winter 1993 issue of _Whole Earth Review_.

Abstract

Within thirty years, we will have the technological means to create superhuman intelligence. Shortly after, the human era will be ended.

Is such progress avoidable? If not to be avoided, can events be guided so that we may survive? These questions are investigated. Some possible answers (and some further dangers) are presented.

The Intelligence Explosion aka Superintelligence, Artificial General Intellligence (AGI)



I J Good (1965)

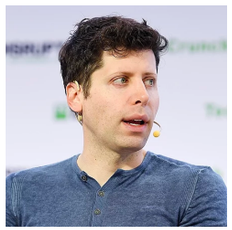


Nick Bostrom (2014)

Self improving (machine) intelligence – learning to learn – leads to a singularity.

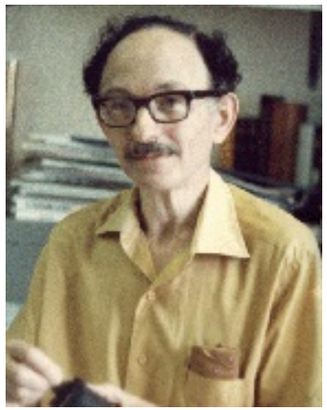
Exponential on exponential growth.

Unintended consequences derived in pursuit of designed priorities.





Superintelligence



I J Good



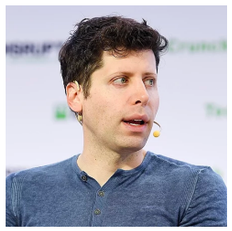
Nick Bostrom (2014)

Self improving (machine) intelligence – learning to learn – leads to a singularity.

Exponential on exponential growth.

Unintended consequences derived in pursuit of designed priorities.



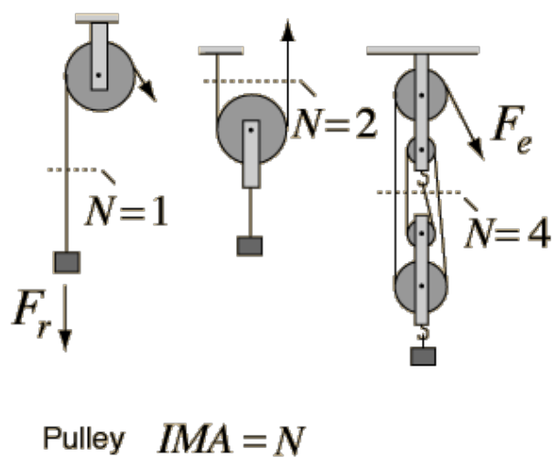




12,000 years of AI

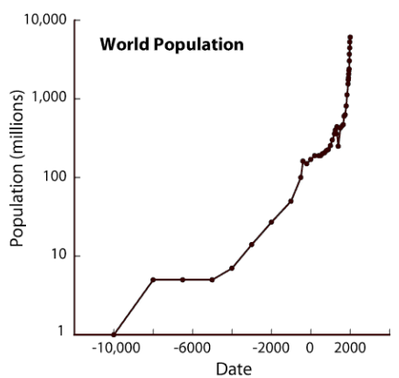
If we accept that intelligence can
(e.g. action, perception, motivatio
learning, reasoning)…
Then every machine and especiall
been examples of AI.

The "intelligence explosion
AI-enhanced human
boom!



Pulley $IMA = N$

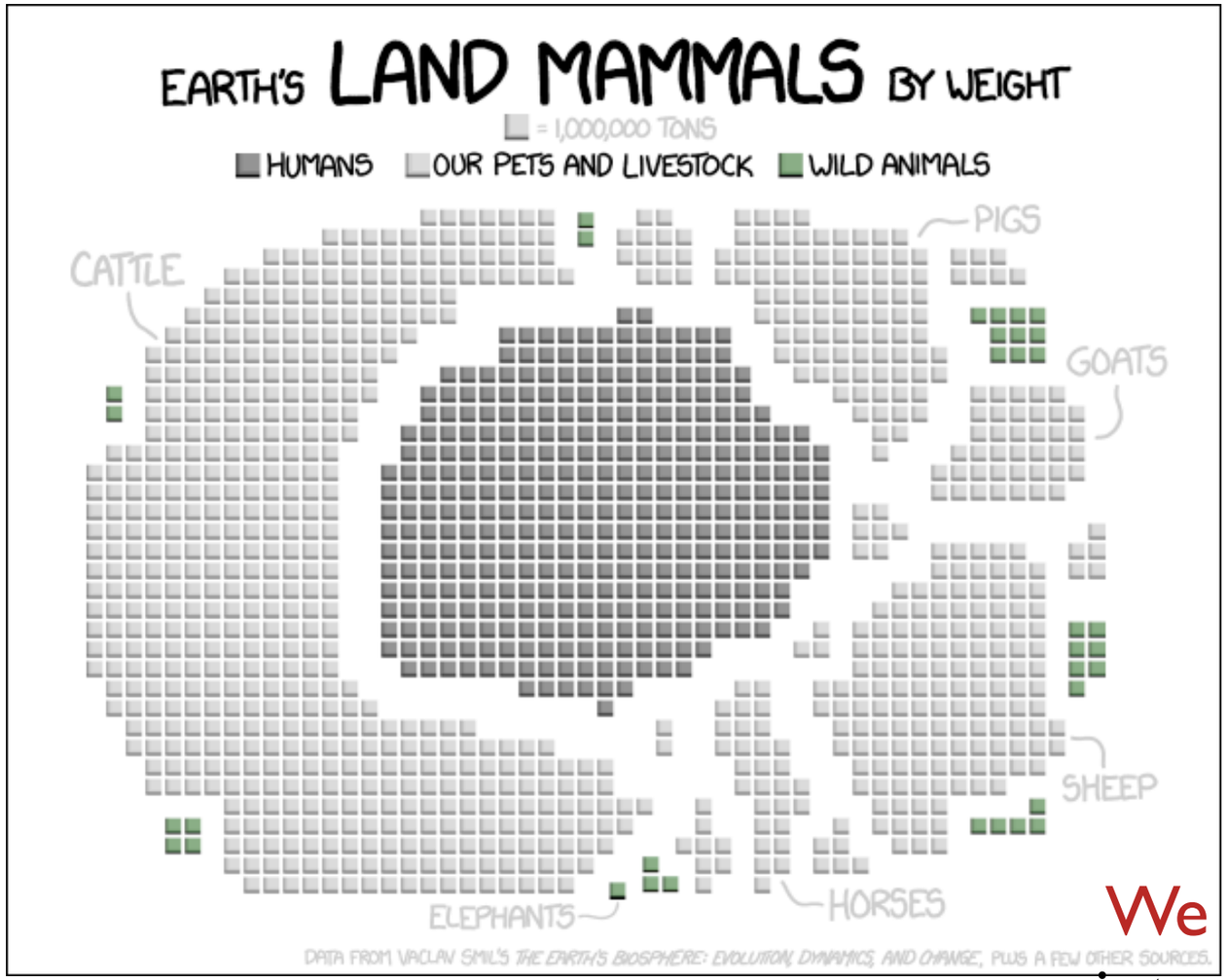


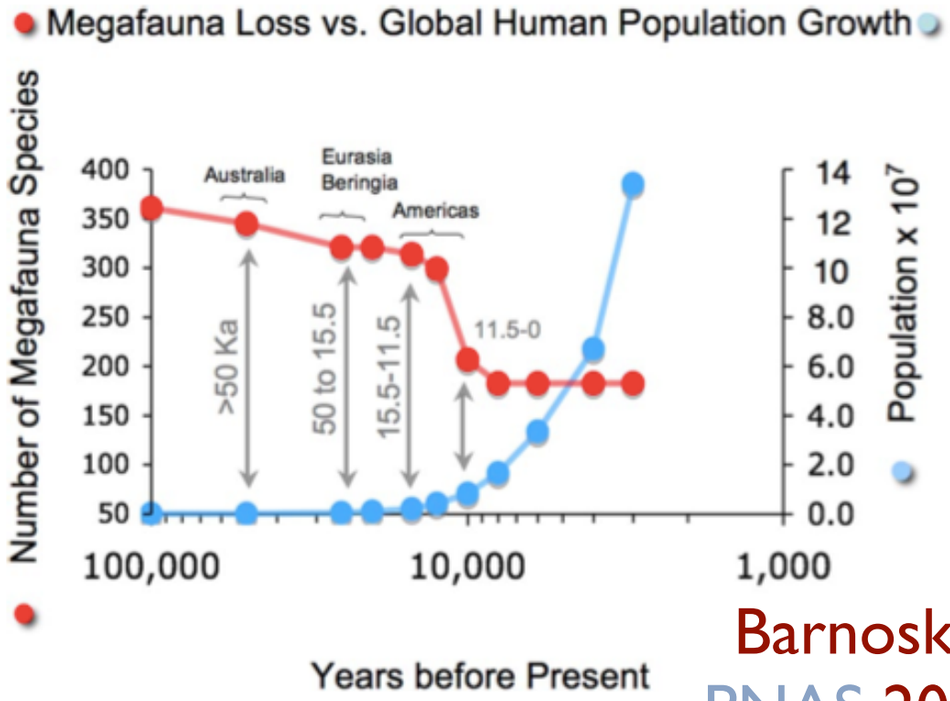


Agency
2015

Unanticipated Consequences

xkcd

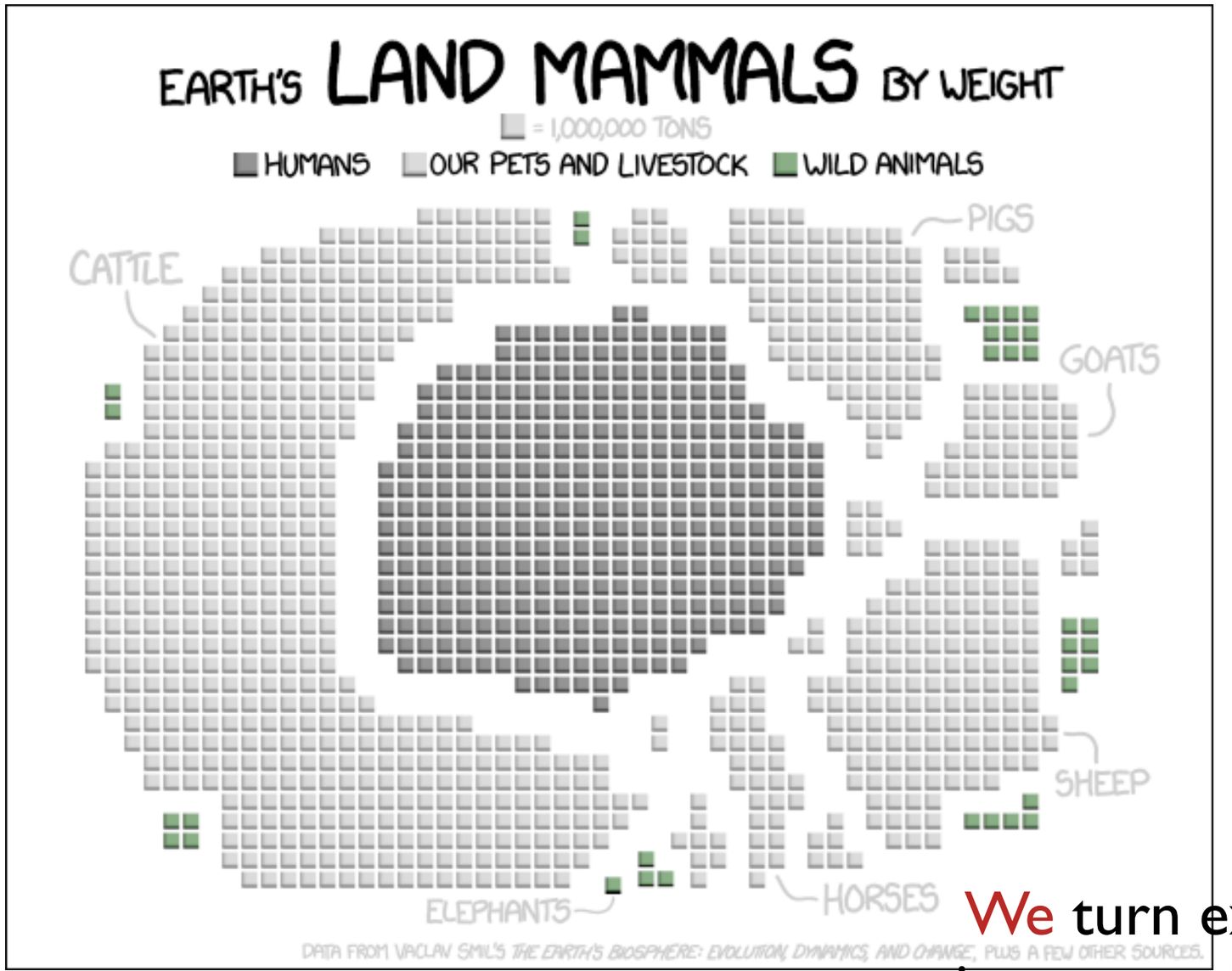




Barnosky, PNAS 2008

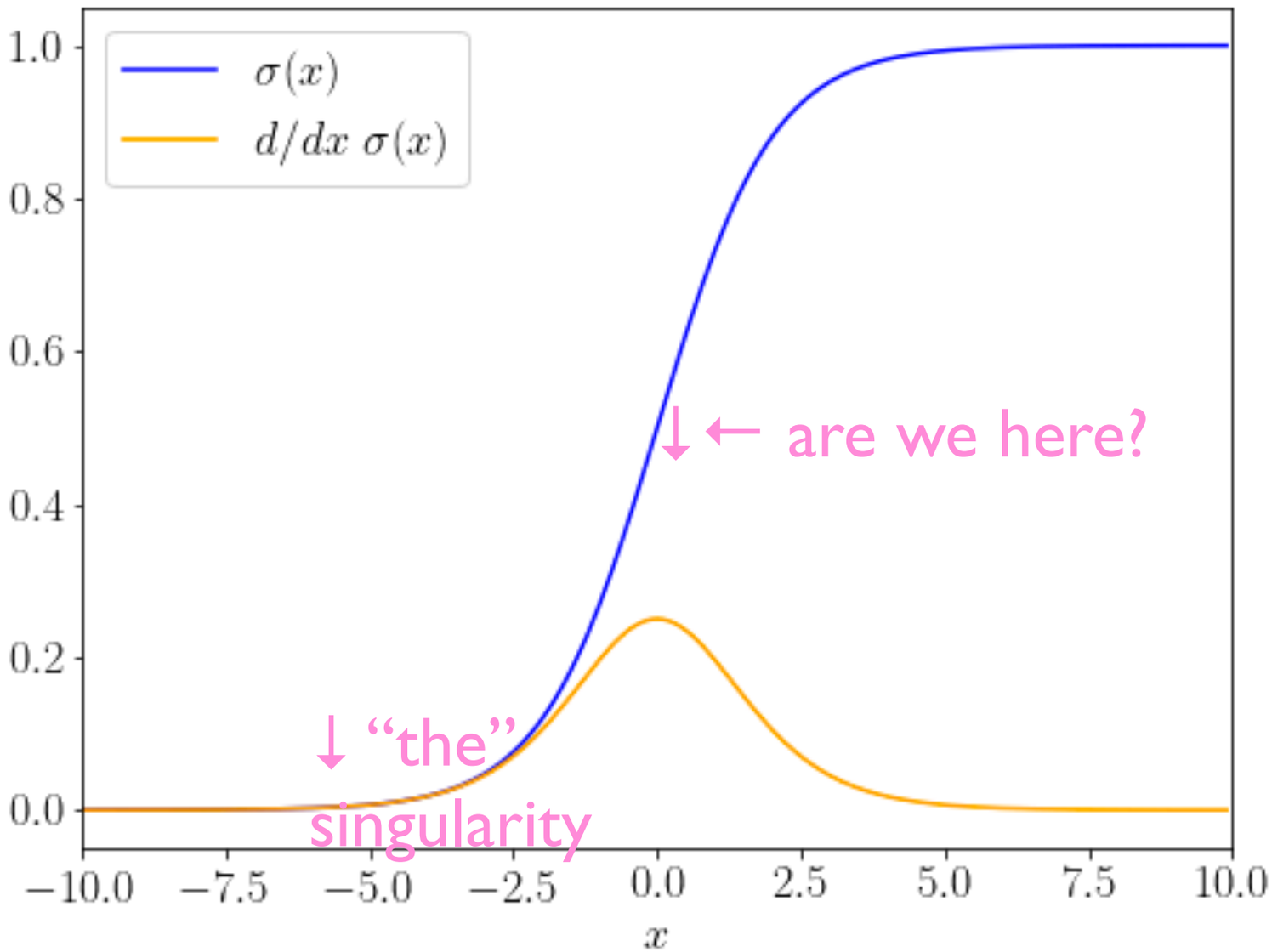
We turn extant biomass and fossil fuels into more biomass but fewer species.

xkcd



We're the Paperclips

We turn extant biomass and fossil fuels into more biomass but fewer species.


$\sigma(x)$
$d/dx\ \sigma(x)$
1.0
0.8
0.6
0.4
0.2
0.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
x
↓← are we here?
↓ “the” singularity

From the human point of view this change will be a throwing away of all the previous rules, perhaps in the blink of an eye, an exponential runaway beyond any hope of control. Developments that before were thought might only happen in "a million years" (if ever) will likely happen in the next century. (In [4], Greg Bear paints a picture of the major changes happening in a matter of hours.)

I think it's fair to call this event a singularity ("the Singularity" for the purposes of this paper). It is a point where our models must be discarded and a new reality rules. As we move closer and closer to this point, it will loom vaster and vaster over human affairs till the notion becomes a commonplace. Yet when it finally happens it may still be a great surprise a

Vinge (1993)



Hofstadter (1957,1964)

The Paranoid Style in American Politics

by Richard Hofstadter

It had been around a long time before the Radical Right discovered it—and its targets have ranged from "the international bankers" to Masons, Jesuits, and munitions makers.

people that makes the phenomenon significant.

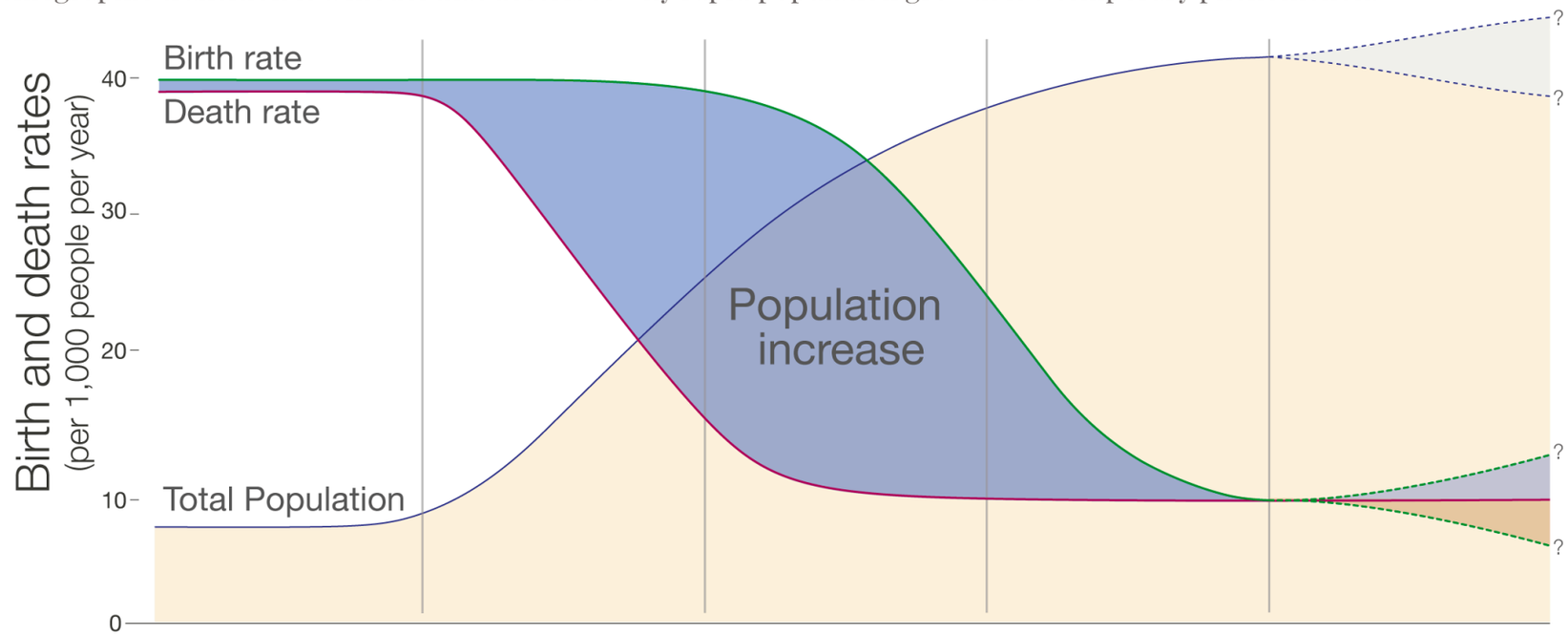
Of course this term is pejorative, and it is meant to be; the paranoid style has a greater affinity for bad causes than good. But nothing really prevents a sound program or demand from being advocated in the paranoid style. Style has

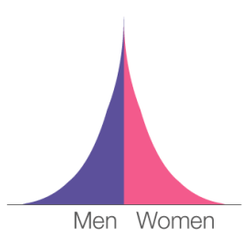
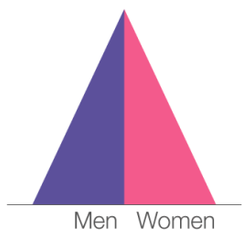
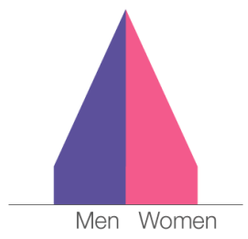
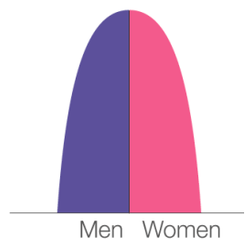
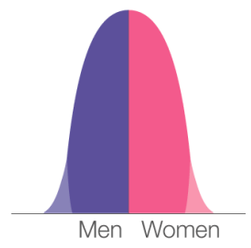
Roser (2023)

The five stages of the demographic transition

Our World in Data

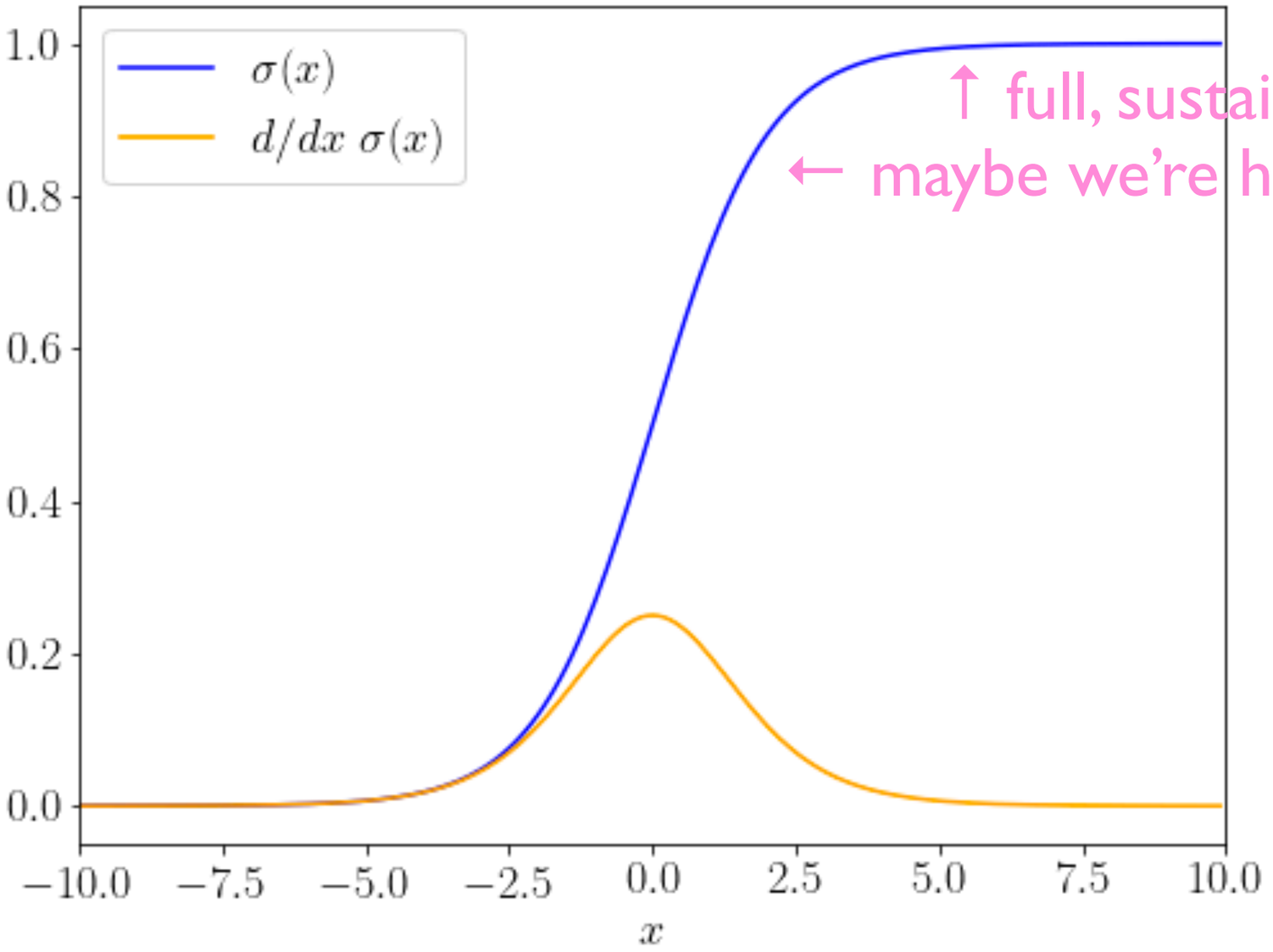
The demographic transition is a model that describes why rapid population growth is a temporary phenomenon.



	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
Birth rate	High	High	Falling	Low	Yet to be seen possibly falling further, possibly rising again
Death rate	High	Falling rapidly	Falling slowly	Low	Low
Population change	Stable or slowly increasing	Rapidly increasing	Increase slows down	Falling and then stable	Little change
Population pyramid					

OurWorldinData.org – Research and data to make progress against the world's largest problems.

Licensed under CC-BY by the author Max Roser.

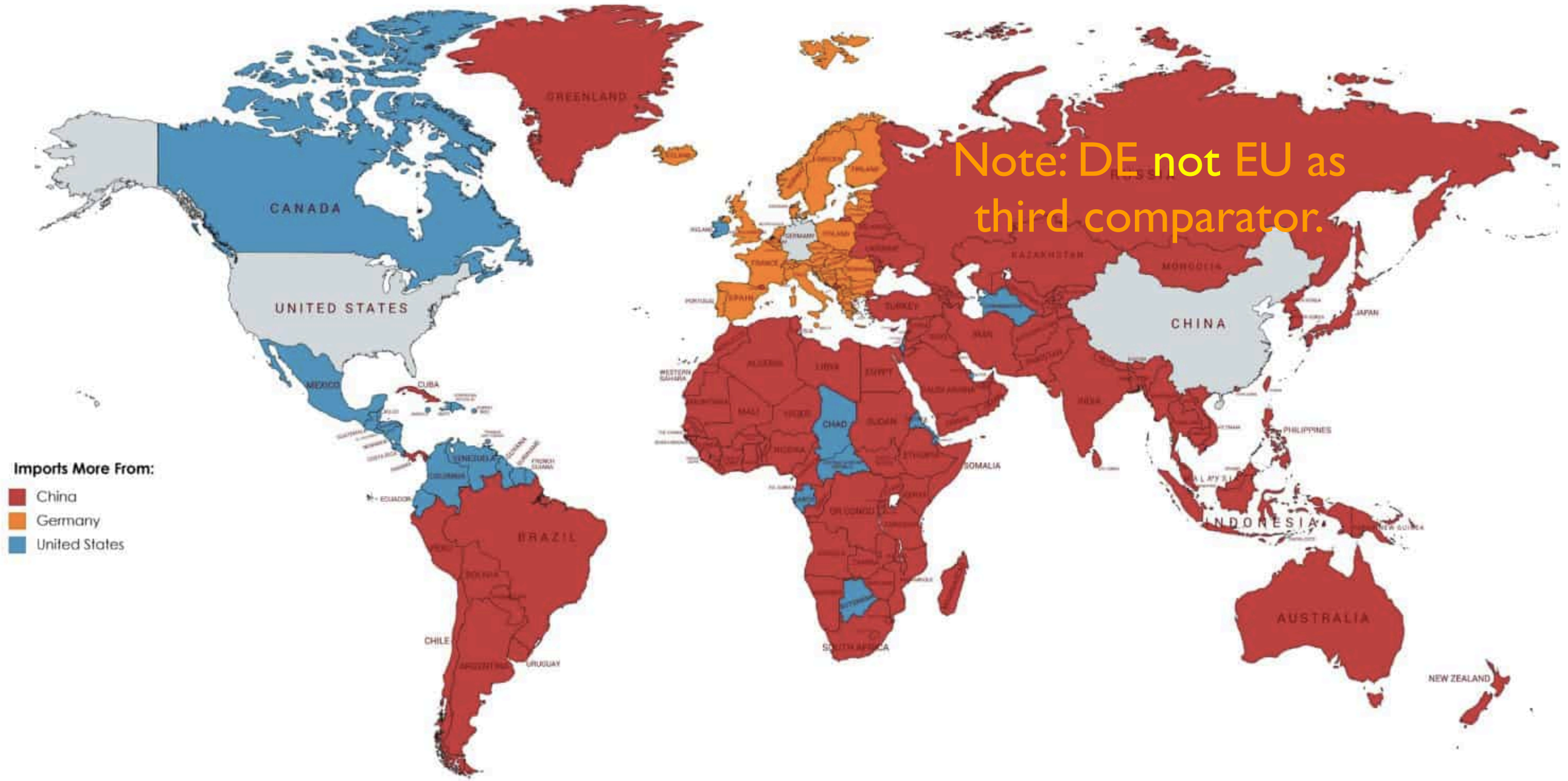


↑ full, sustainable ecosystem

← maybe we're here?

What we need is to be smart enough to execute a graceful security and economy transition.

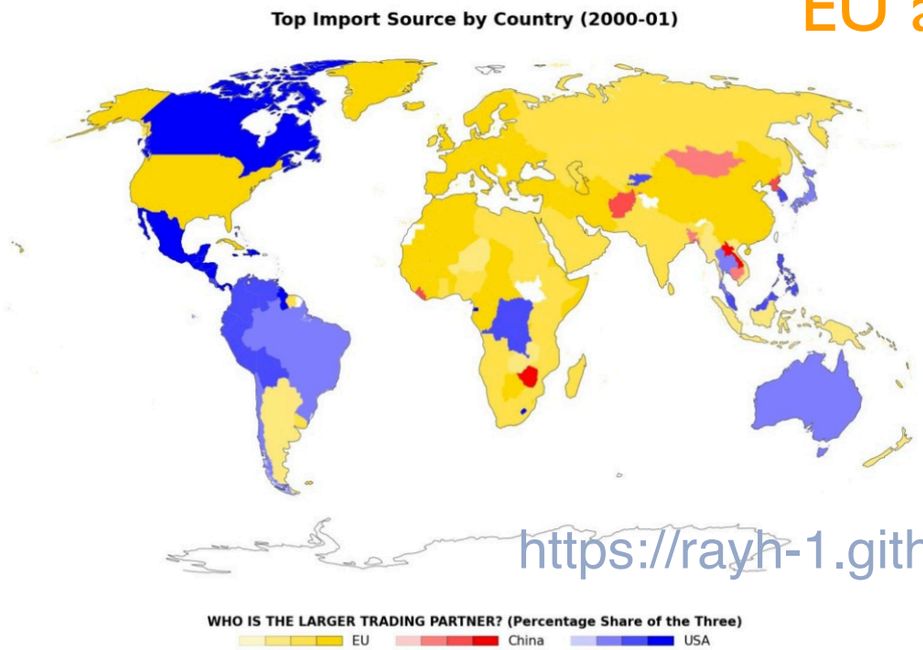
alexander-schiendorfer



2020 – in 1980, the world was blue. https://merchantmachine.co.uk/china-vs-us/

Month-Year: 2000-01

2000



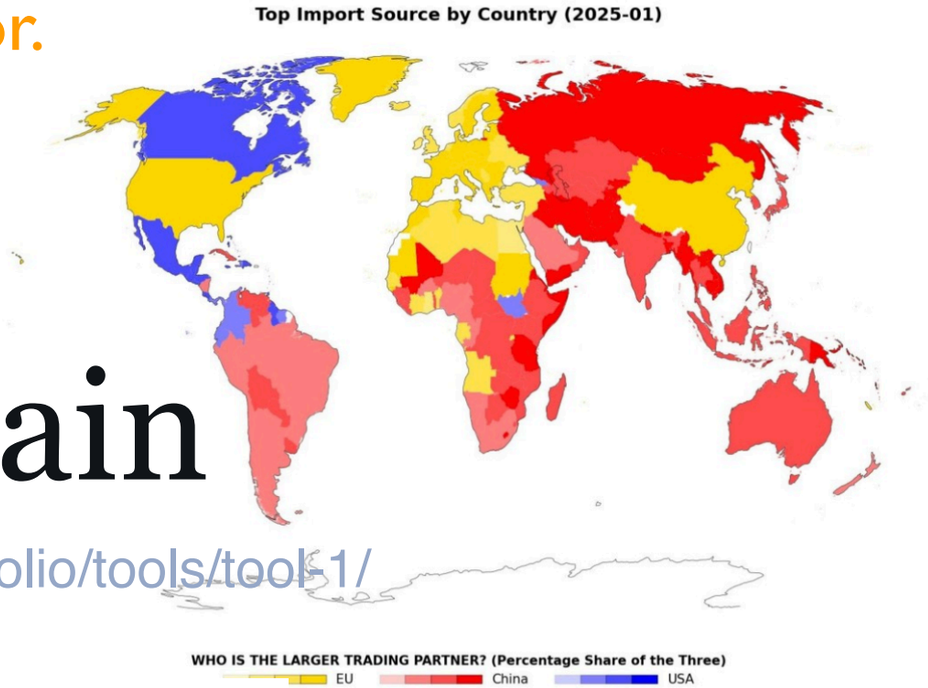
EU as comparator.

Ray Hossain

https://rayh-1.github.io/Ray_Portfolio/tools/tool-1/

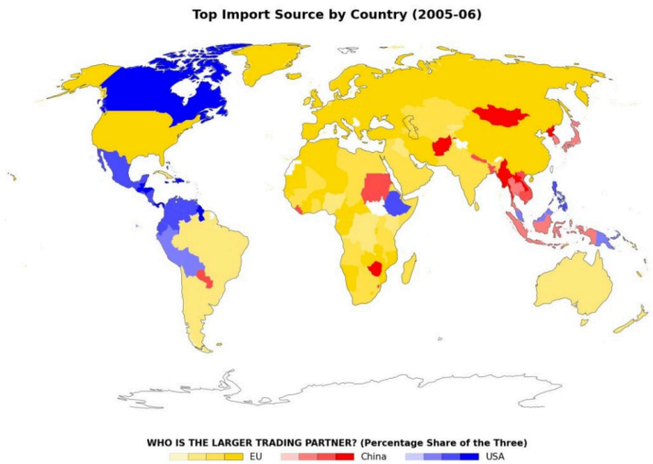
Month-Year: 2025-01

2025



Month-Year: 2005-06

July 2005

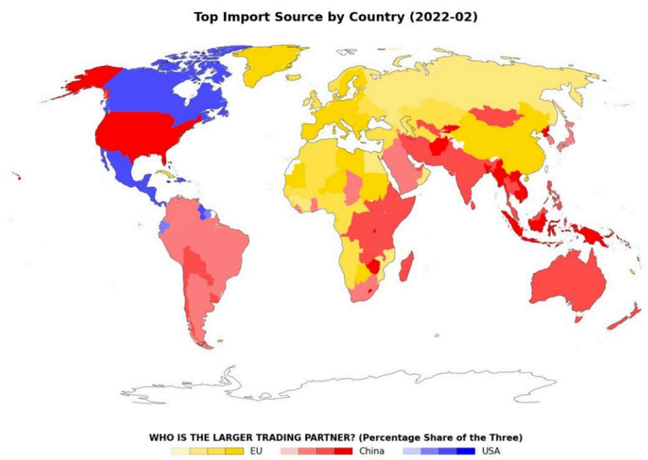


Month-Year: 2005-07



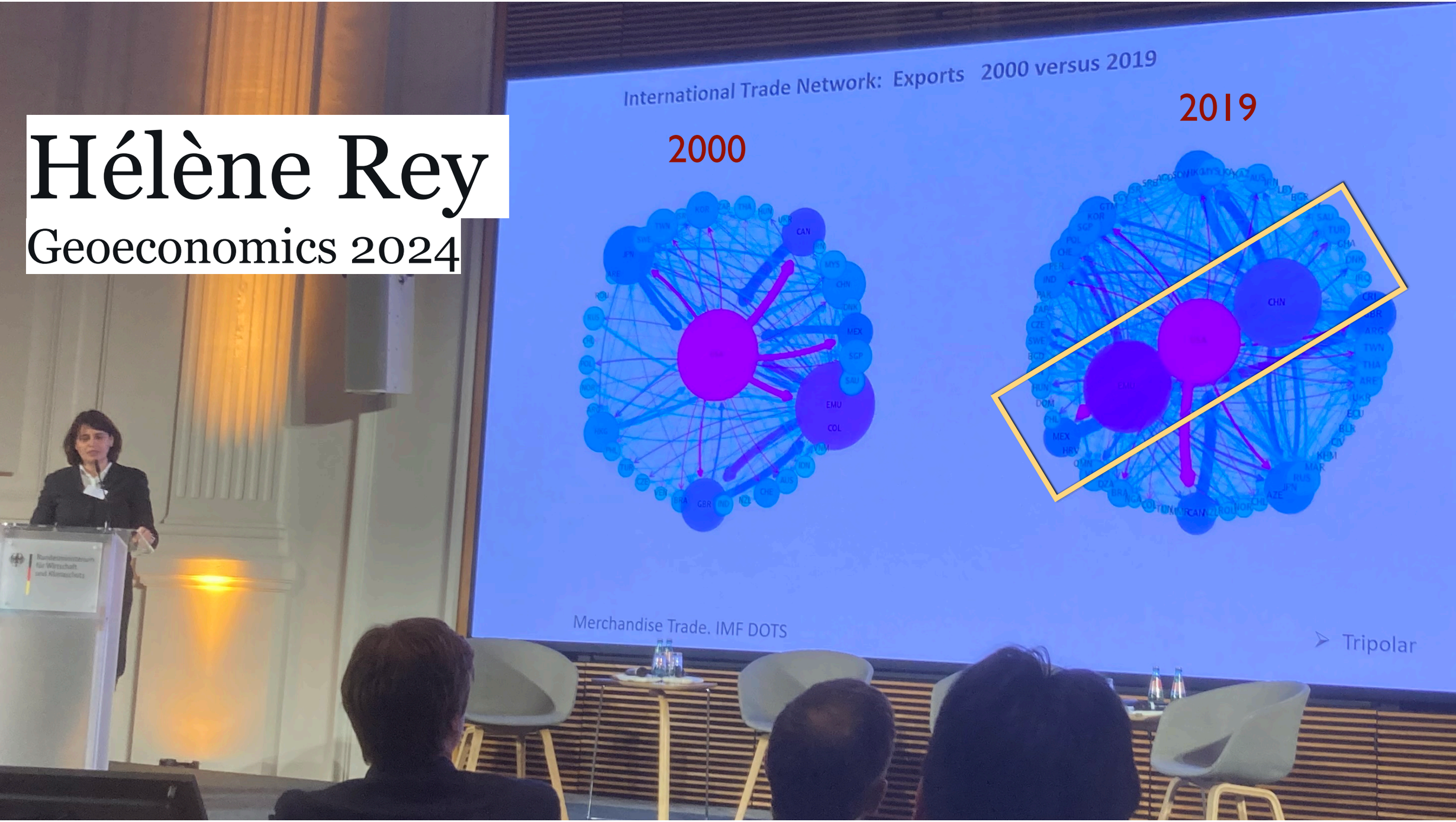
Month-Year: 2022-02

Feb 2022



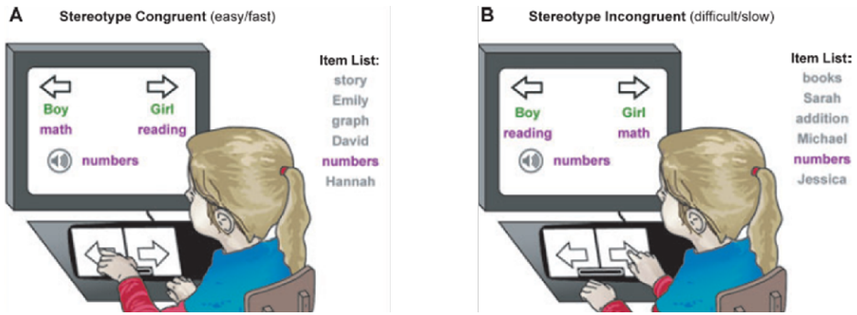
Hélène Rey

Geoeconomics 2024



AI Trained on Human Language Replicates Implicit Biases,

Science
AAAS



Gender bias [stereotype]

Female names: Amy, Joan, Lisa, Sarah…

Male names: John, Paul, Mike, Kevin…

Family words: home, parents, children, family...

Career words: corporation, salary, office, business, …

Original finding [N=**28k** participants]: $d = 1.17$, $p < 10^{-2}$

Our finding [N=8x2 words]: $d = 0.82$, $p < 10^{-2}$

Caliskan, Bryson & Narayanan (*Science*, April 2017)

Our implicit behaviour is not our ideal. Ideals are for explicit communication, and planning.

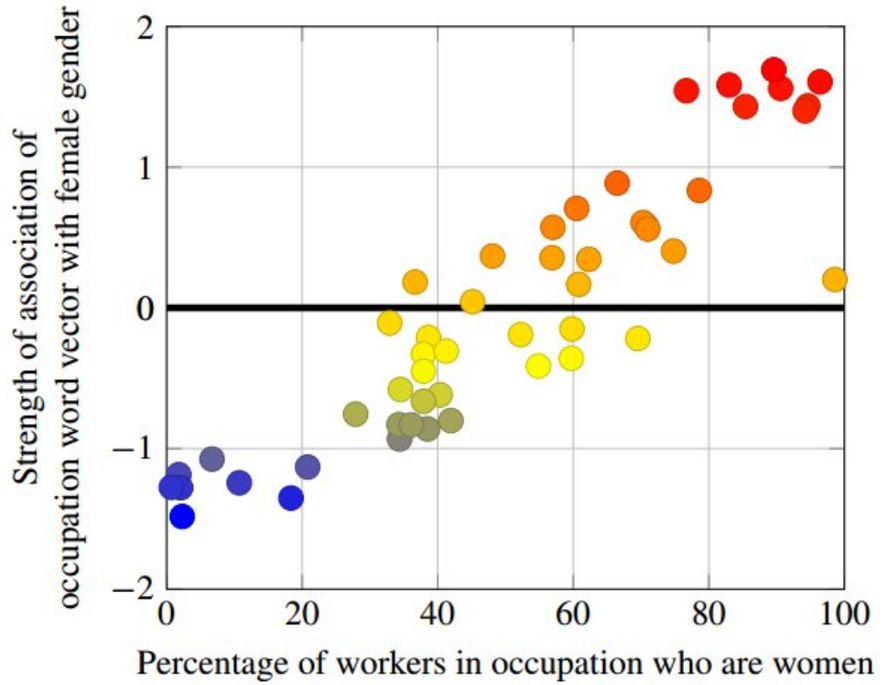


Figure 1. Occupation-gender association
Pearson's correlation coefficient $\rho = 0.90$ with p-value $< 10^{-18}$.

2015 US labor statistics
$\rho = 0.90$

Sources of Bias

1. Careless ignorance due to lack of diversity in developers.
2. Existing bias in society 'uploaded' via ML into AI
3. Genuinely bad people introducing bias – delivered at procurement, development, deployment or via cyberattack.

How to Address Them

1. Test and iteratively improve code (necessary for all 3 problems).
2. (Re)build AI using data from the societies you most want its values to reflect.
3. Audit for use of best practice, avoidance of worst; attribute unacceptable behaviour and prosecute. (Standard product law.)

Defining Rights

- Government was once framed as providing a system of protections in exchange for taxation…
  - Largely replaced presently by framing government as meeting obligations to which humans are (passively) entitled.
  - Rights framed as discoverable truths, yet positive rights create ‘tensions.’



Rights: Entitlement versus Active, Dynamic Construction

cf Richard Bellamy (2001), Adam Tomkins (2001), Bryson *in prep*

- Government was once framed as providing a system of protections in exchange for taxation…Largely replaced by framing government as meeting obligations to which humans are (passively) entitled.
  - Rights as discoverable truths, yet positive rights create 'tensions.'
- Alternative framing: residents constitute their states, use governments to coordinate public goods' innovation and stewardship.
  - Rights as negotiated and optimised by (active, autonomous) residents, resolving 'tensions' in light of specific contexts.
  - Less adversarial model of government – and perhaps more robust.

Perspectives: Ethics versus Rights

- Older: Ethics are universal, we just need to understand them.
- Newer: Ethics are culturally specific, a good way to evade the law. Unlike rights –
- Developed: Rights are grounded in International Law and treaty,
  - every nation has signed up to the UDHR at least once.
- Developing: Rights are European, another colonialism / conservatism preventing the expansion phase of development,
  - we didn't have much choice but to sign.

AI Ethics Definitions for policy

- Artificial Intelligence is intelligence deliberately built.
- Agents are any vector of change, e.g. chemical agents.
- Moral agents are considered responsible for their actions by a society.
- Moral patients are considered the responsibility of a society's agents.
- Responsibility is a property moral agents of a society assign to each other to uphold that society.
  - Implies a peer relationship (like trust.)

} Arguably, ethics is determined by and determines a society–a constantly renegotiated set of equilibria for empowerment.

Definitions

for reasoning about policy

- Responsibility is a property moral agents of a society assign to each other to uphold that society.
  - Implies a peer relationship (like trust.)
- Trust is a relationship between peers where the trustee is not micromanaged but allowed to defect, whether for pragmatics or to allow innovation.
- Accountability is a society's capacity to trace responsibility.
- Transparency is how we implement accountability. Not an end in itself.

} We should not trust powerful agencies, but rather construct systems to hold them accountable.

Definitions

for reasoning about policy

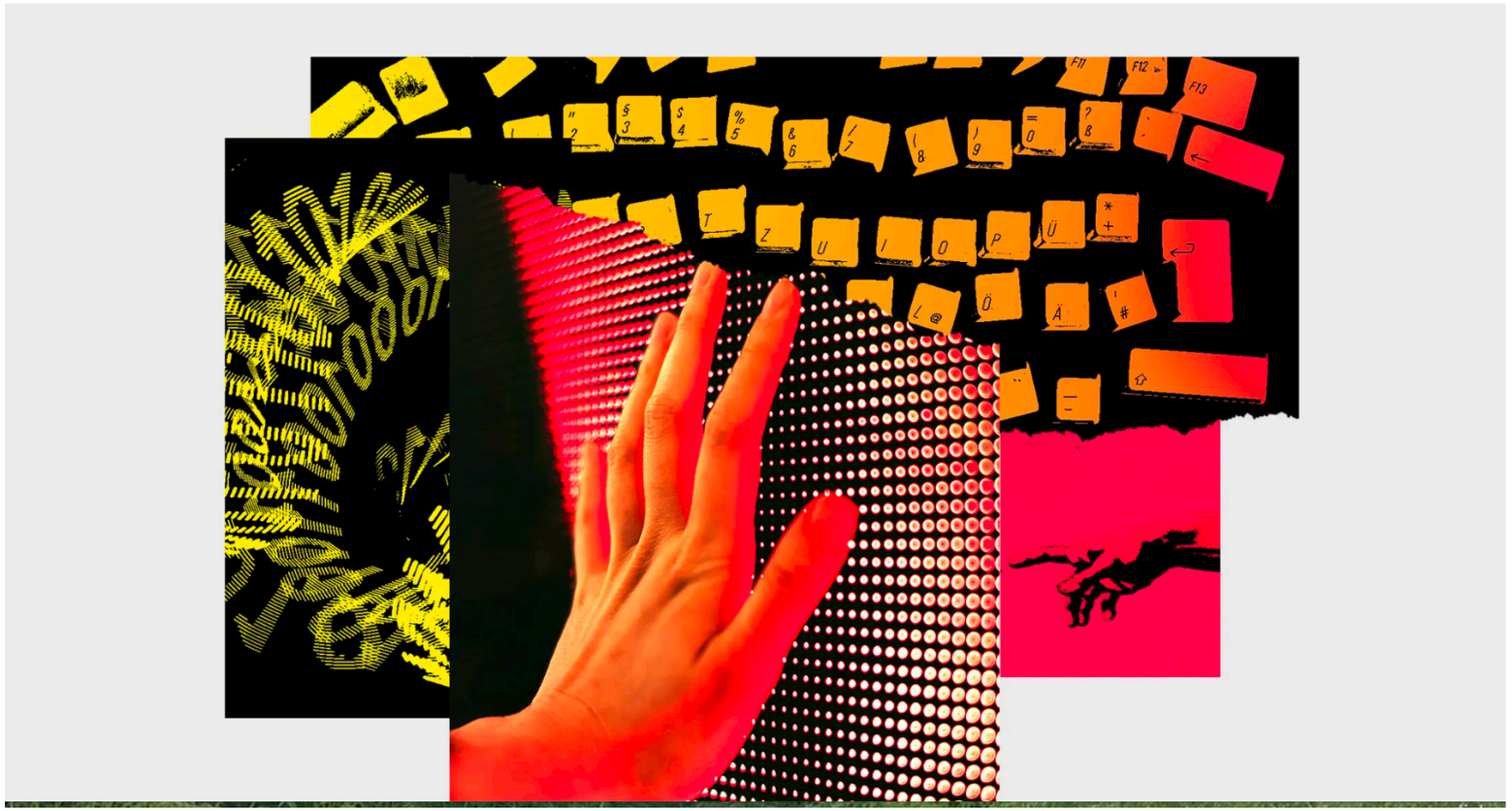
- Responsibility is a property moral agents of a society assign to each other to uphold that society.
  - Implies a peer relationship (like trust.)
- Trust is a relationship between peers where the trustee is not micromanaged but allowed to defect, whether for pragmatics or to allow innovation.
- Accountability is a society's capacity to trace responsibility.
- Transparency is how we implement accountability. Not an end in itself.

} Accountability is essential to improvement, which is key to all forms of security.

JOANNA J. BRYSON IDEAS JUN 26, 2022 7:00 AM

One Day, AI Will Seem as Human as Anyone. What Then?

A Google engineer's claim that the LaMDA program is sentient underscores an urgent need to demystify the human condition.



Bryson *Wired* June 2022

From here on out, the safe use of artificial intelligence requires demystifying the human condition. If we can't recognize and understand how AI works —if even expert engineers can fool themselves into detecting agency in a "stochastic parrot"—then we have no means of protecting ourselves from negligent or malevolent products.

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

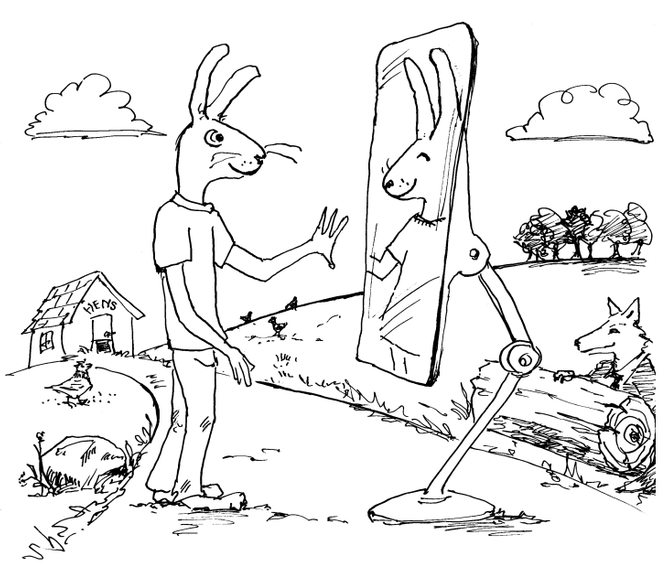
Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

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

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

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether





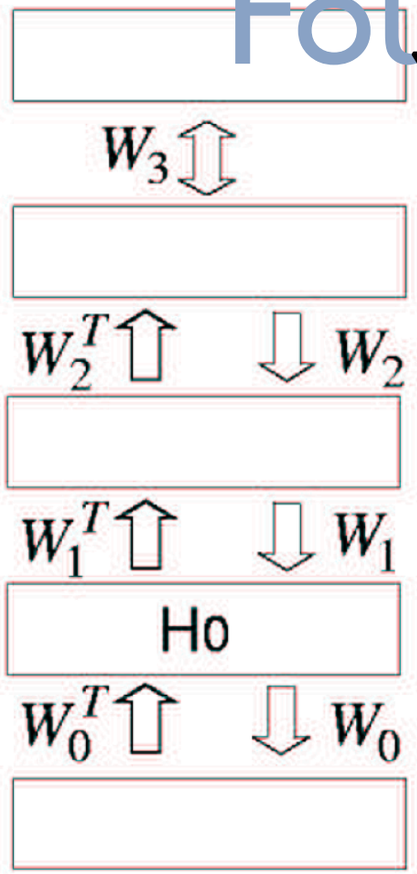


Art: Amy Sterly



Photo: Don Faulkner
Birds: orange-fronted conure
cf: Dabelsteen, Balsby

Foundation Models and 'Guardrails'







Josh Dzieza (2023)

- Foundation models are massive deep-learnt models of massive data
  - – highly resource intensive.
- Generative AI retrieves / 'predicts' most likely outcome based on a context, including a prompt. But some outcomes are 'bad'.
- 'Guardrails' are additional context to suppress retrieval of bad alternatives.
  - Often trained using reinforcement learning / brute force.

Is This Technology Ethics?

(Nyholm, p. 118, context of gamification and nudges)

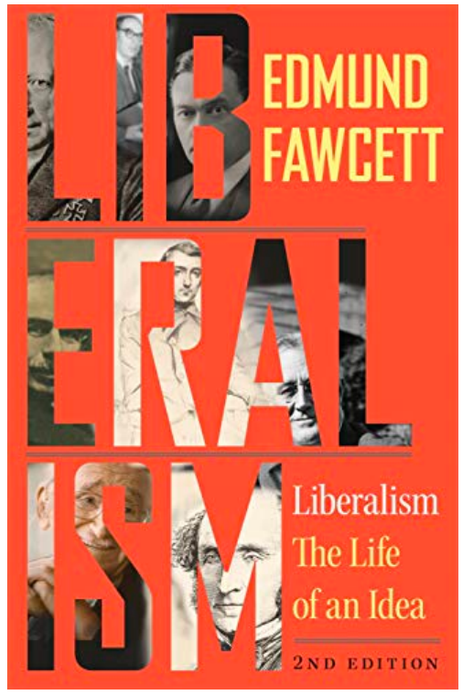
- Desiderata:
  - ‘We’ work to ends we ourselves select, and
  - for reasons we understand.

This is also a key part of the new EU AI Act – humans must do this in any organisation deploying AI, so responsibility and therefore improvement are possible.

Can ‘we’ include systems subject to design and ownership?

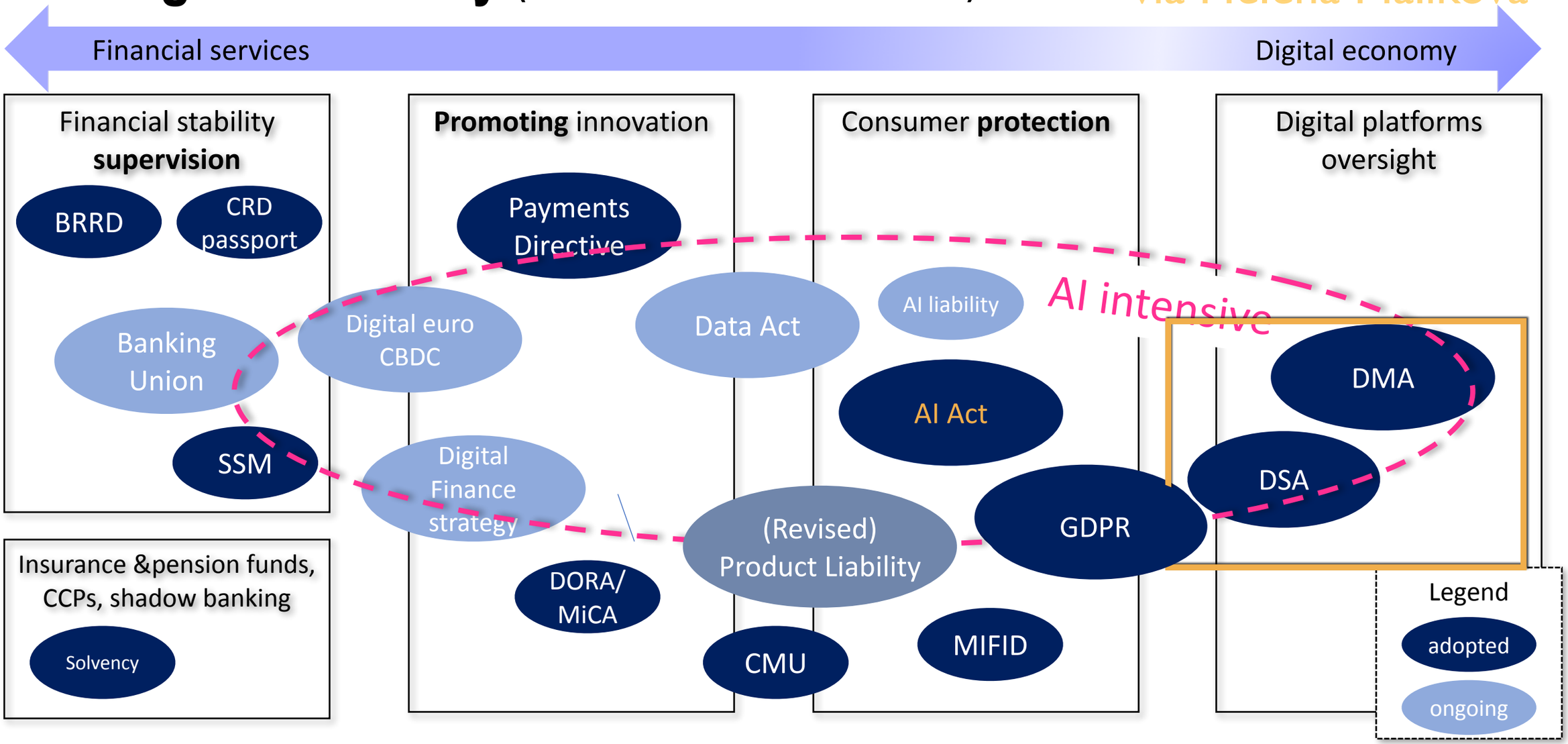
Regulation and Principles of Liberalism

- Context – conflict and change are inevitable.
- Distrust of power – state, wealth, or populism.
- Faith in human progress – we can make things better for ourselves.
- Civic respect – "law-backed respect by state and society for people and their projects, whoever they are." (Fawcett 2018)

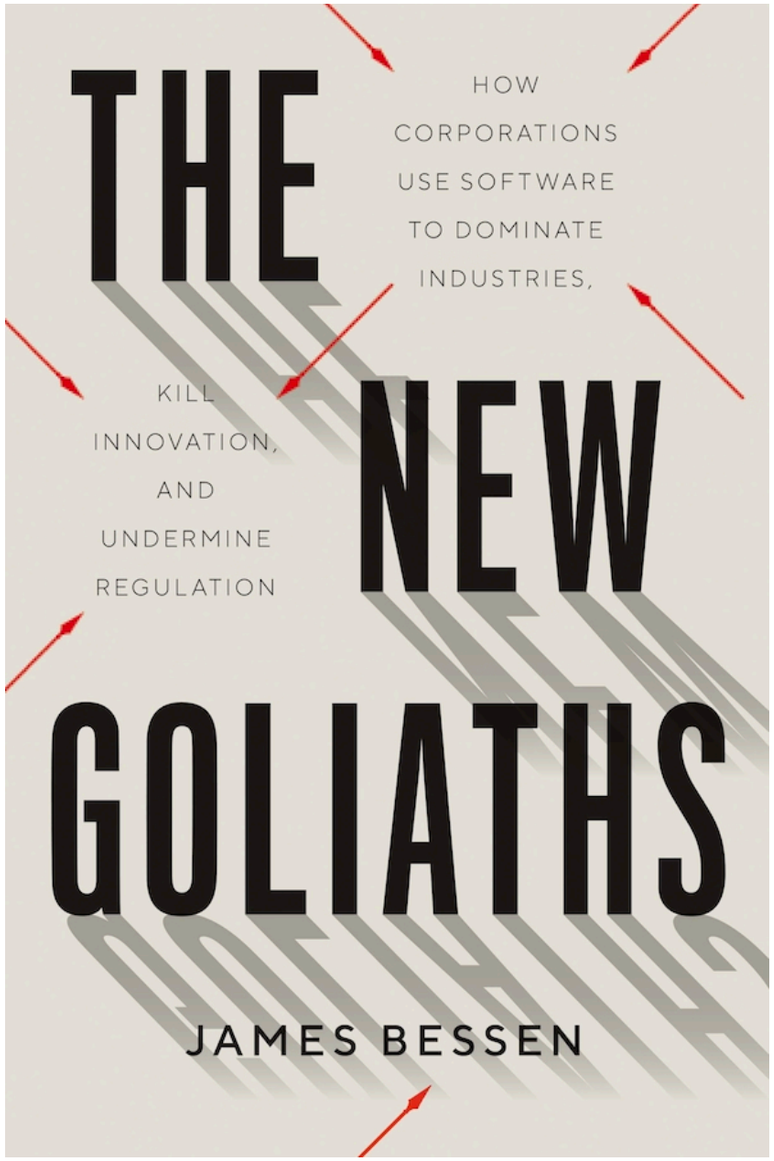


Selected EU regulatory initiatives relating to financial services and digital economy (illustration/non-exhaustive)

via Helena Malikova



* Many of the initiatives listed are horizontal in nature, this simplification slide does not capture the full scope and objectives of the proposals



2022

- Bessen (2022, "The New Goliaths") documents that in every sector, the corporation that invests most in software is becoming market dominant.
  - Make themselves into a "natural monopoly" by optimising logistics.
- Innovation slows. New entries are blocked.
- Regulation is essential to maintaining the diversity necessary for innovation.
- Democracy (sometimes?) helps maintain diversity.

The Brussels Effect

- The EU is a constructed peer – an attempt to end a long history of warfare produced a leading economy and single actor.
- The EU is not policing the world's AI, only its own citizens' wellbeing.
  - But the computation it did in producing that legislation is reusable!
- Any economy can ask for compliance costs proportionate to the value of their market.



Concurrency x communication is humanity's super power.

AI apps have been "super-human" for years now at chess, go, speech transcription, lip reading, deception detection from posture, forging voices, handwriting, & video, general knowledge and memory.

Books, planes, and houses are also super-human.

Much of the spectacular recent growth in AI derives from developers using ML to upload prior discoveries (previous computation) by biological evolution and human culture.

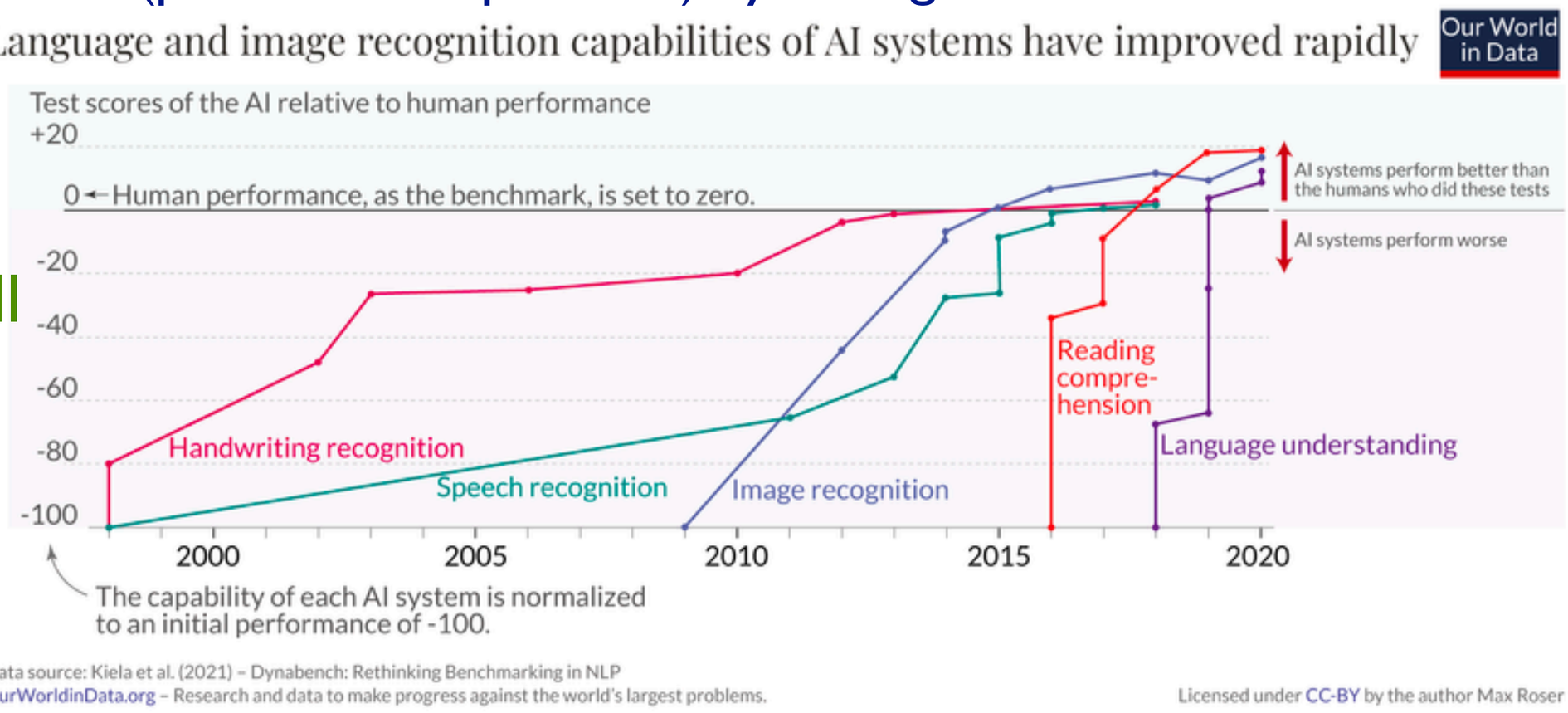
Prediction: pace of AI relative improvement will slow as AI joins the–now accelerating–frontier of our knowledge.

AI apps have been "super-human" for years now at chess, go, speech transcription, lip reading, deception detection from posture, forging voices, handwriting, & video, general knowledge and memory.

Books, planes, and houses are also super-human.

Much of the spectacular recent growth in AI derives from developers using ML to upload prior discoveries (previous computation) by biological evolution and human culture.

Prediction: pace of AI relative improvement will slow as AI joins the–now accelerating–frontier of our knowledge.



What Is Trust?

- In games where you will never meet the other person you play with:
  - Many people will trust an unknown partner with an investment.
  - Yet also refuse known partners who are beneficial but unfair.
- Trust is a tradition a community builds of fair cooperation.

Open Access | Published: 25 September 2017

Expectations of Fairness and Trust Co-Evolve in Environments of Partial Information

Paul Rauwolf ✉ & Joanna J. Bryson

Dynamic Games and Applications **8**, 891–917 (2018) | Cite this article

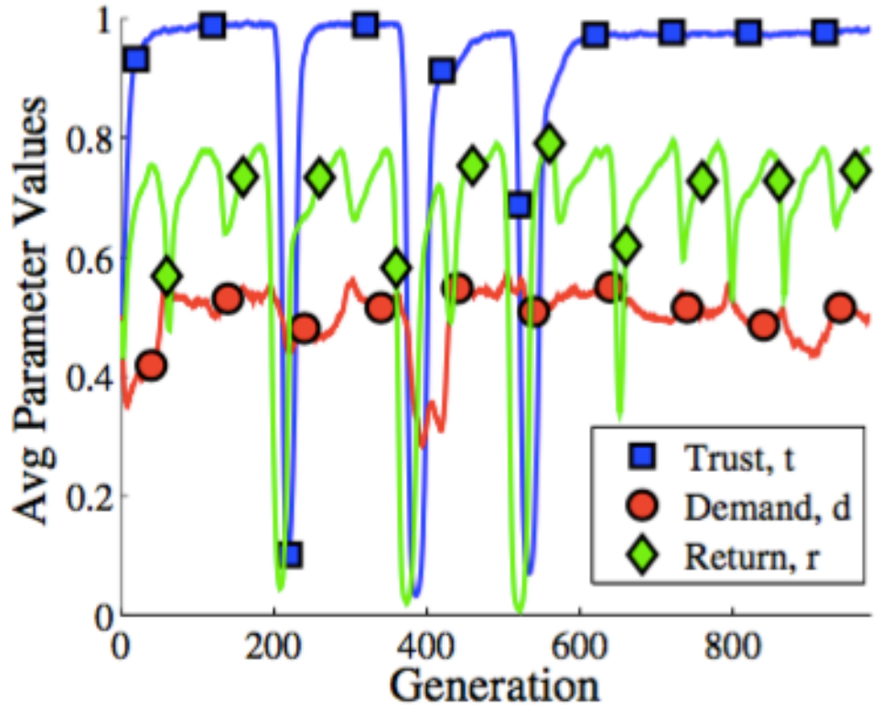
Rauwolf & Bryson (*Dynamic Games and Applications*, September 2017)

“Maintaining” High Trust

PRIFYSGOL BANGOR UNIVERSITY

Paul Rauwolf
@PaulRauwolf





Open Access | Published: 25 September 2017

Expectations of Fairness and Trust Co-Evolve in Environments of Partial Information

Paul Rauwolf ✉ & Joanna J. Bryson

Dynamic Games and Applications **8**, 891–917 (2018) | Cite this article

Rauwolf & Bryson (*Dynamic Games and Applications*, September 2017)

Trust Requires Choice & Ignorance

