Artificial Intelligence and Human Cooperation

Joanna J. Bryson
Professor of Ethics and Technology

Hertie School
Centre for
Digital Governance

@j2bryson

Outline

- AI and Ethics
- Human Cooperation and AI

Definitions

for reasoning about cooperation and security

- Intelligence is doing the right thing at the right time – computing action from context (Romanes 1882).
- Artificial Intelligence is intelligence deliberately built.
  - Deliberation ⊨ responsibility in human societies between human adults.
- Moral agents are those considered responsible.
- Moral patients are the subjects of responsibility.
- Moral agents are approximately peers, enforcing responsibilities on each other to uphold society.
- Trust is also a peer-wise relationship – cooperation without micromanagement.

Moral subjects compose a society.

Conjecture: Ethics is the way a society defines and secures itself.

Enforcement occurs between peers

- Citizens shouldn't trust corporations or governments, but should demand transparency and accountability.
- Antitrust was originally intended to keep companies manageable by a democracy.
- Perhaps transnational corporations require transnational cooperation for regulation.

UNIVERSITY of CALIFORNIA PRESS | GLOBAL PERSPECTIVES

Bryson, Joanna J., and Helena Malikova. 2021. "Is There an AI Cold War?" *Global Perspectives* 2 (1). https://doi.org/10.1525/gp.2021.24803.

ongoing work with Helena Malikova, EC DG Competition

Technology and Global Change

Is There an AI Cold War?

Joanna J. Bryson[1] [a], Helena Malikova[2] [b]

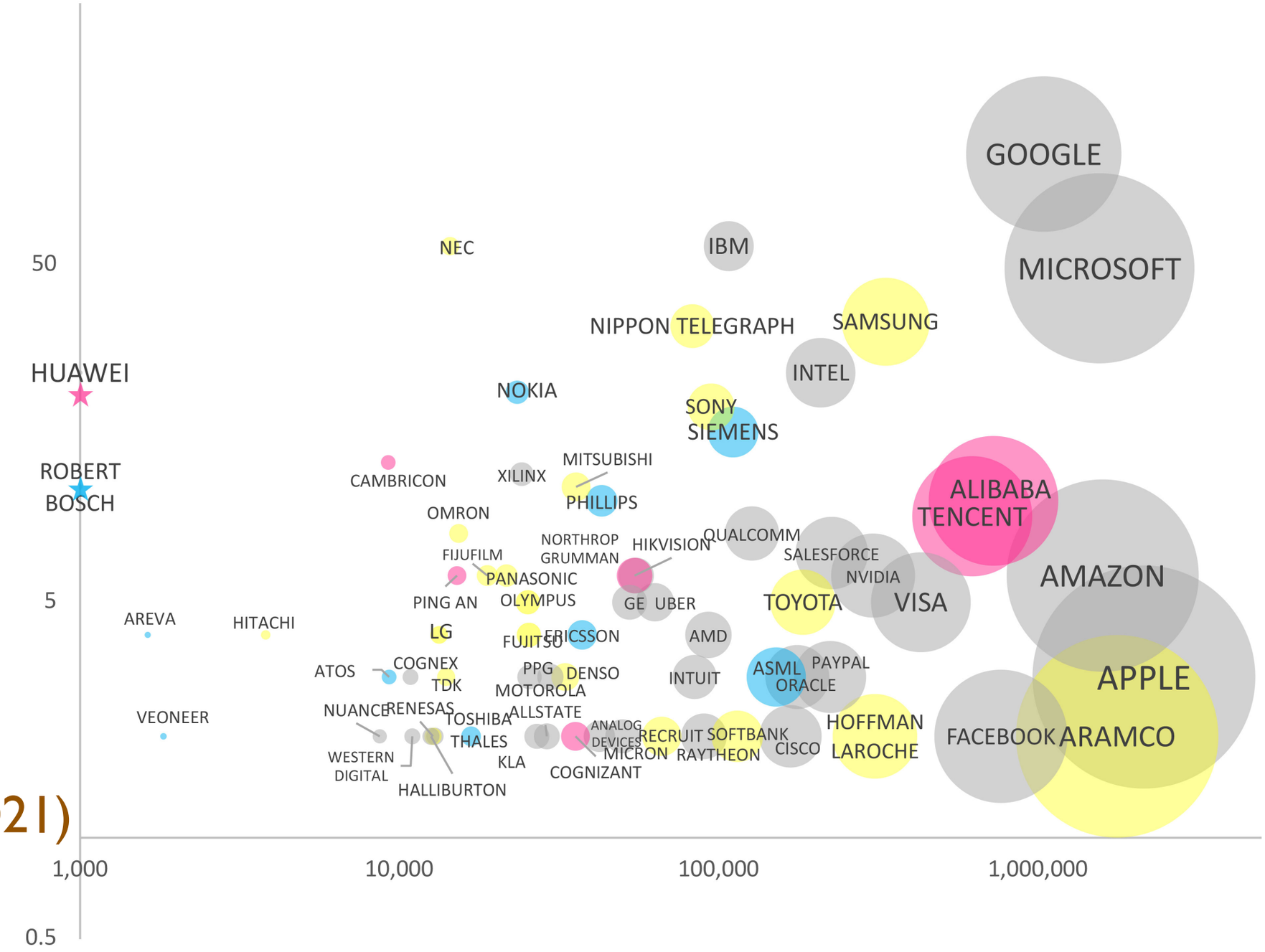
[1] Centre for Digital Governance, Hertie School, Berlin, Germany, [2] Directorate General for Competition, European Commission

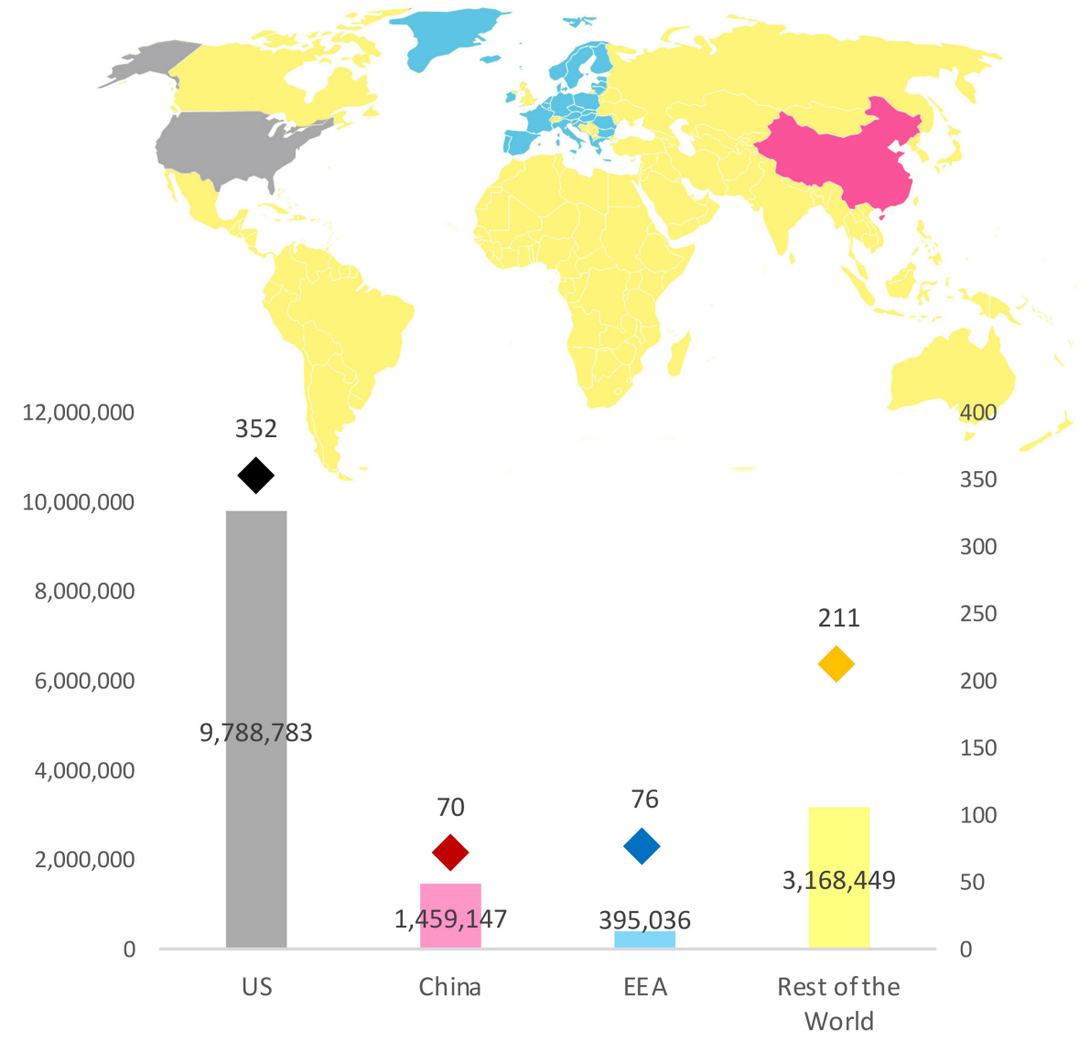
Keywords: technology policy, digital governance, artificial intelligence, transnational regulation, polarity

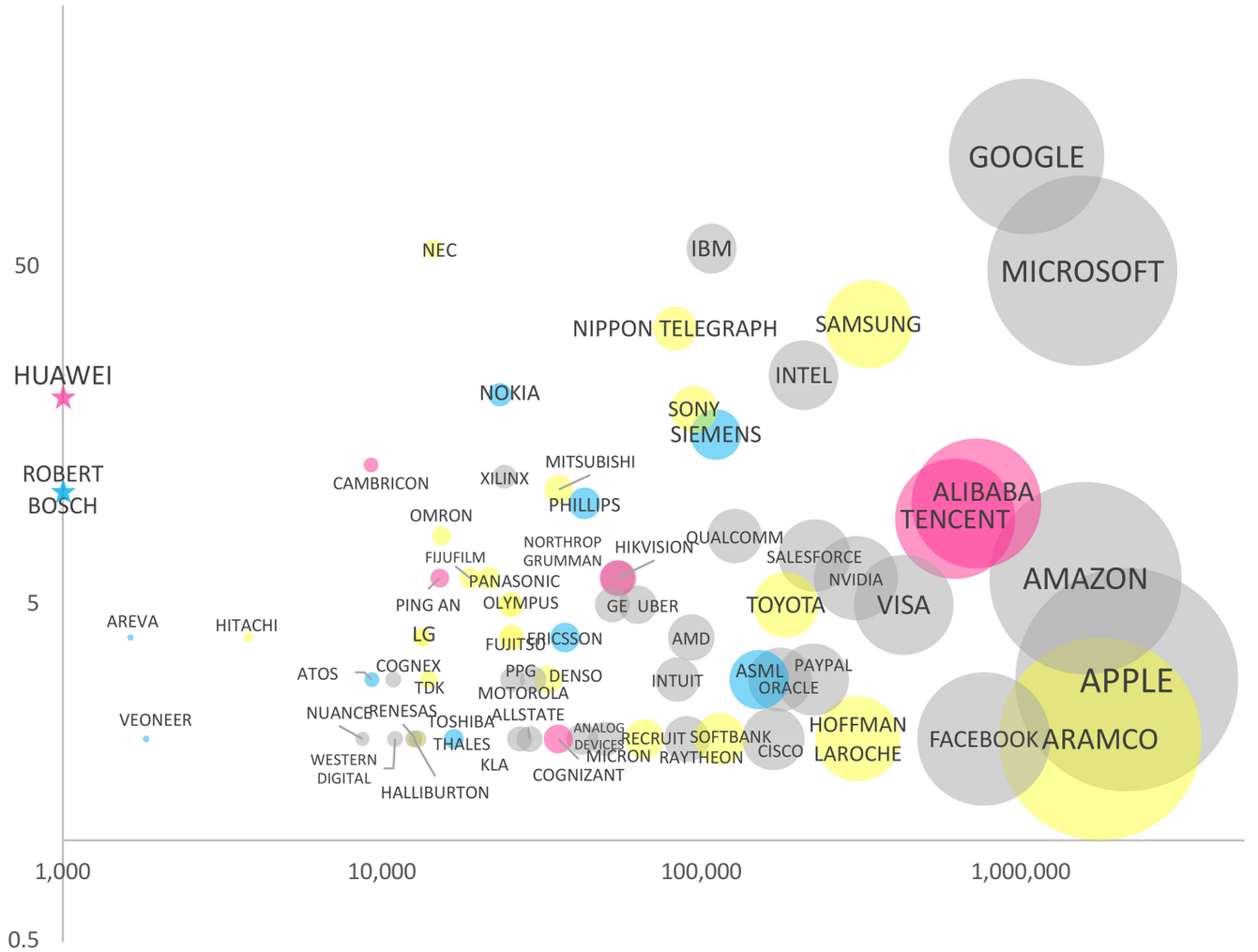
↑y axis = log 2019 AI patents in WIPO (G06N of the IPC classification dedicated to "Computer sys- tems based on specific computational models")

→x axis = log Oct 2020 Market Capitalisation

Bryson & Malikova (2021)







Bryson & Malikova (2021)

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

AI Cannot Be a Peer

- Consent cannot be meaningfully granted even by someone who is owned, let alone by something that is designed.
- Law and Justice are more about dissuasion than recompense.
- The equivalent of the phenomenological impact of social sanctions on humans (or other social animals) cannot be designed or maintained in AI.
  - Evolution makes us so systemically averse to isolation, loss of status that jailing an opponent can feel like recompense, (but it isn't.)
- Safe AI is modular. AI legal agents would be the ultimate shell company.

Artificial Intelligence and Law
September 2017, Volume 25, Issue 3, pp 273–291 | Cite as

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

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

The origins of bias

The Map of Germany Problem

The perfect map of Germany would be the same size as Germany.

All models are wrong, but some are useful – Box (1976)

Intelligence is computation–a transformation of information.

Not math.

Perception ⟹ Action

Computation is a physical process, taking time, energy, & space.

Finding the right thing to do at the right time requires search.

Cost of search = # of options$^{\#\text{ of acts}}$ (serial computing).

Examples:

- Any 2 of 100 possible actions = 100^2 = 10,000 possible plans.
- # of 35-move games of chess > # of atoms in the universe.

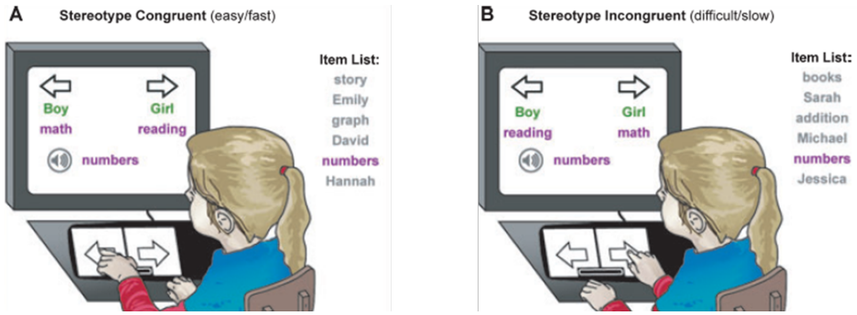
Concurrency can save real time, but not energy, and requires more space.

Quantum saves on space (sometimes) but not energy.

Human intelligence –including AI– is based on millennia of stored, concurrent search, deployed as heuristics from culture.

Quantum expert: Viv Kendon

AI Trained on Human Language Replicates Implicit Biases



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, planning.

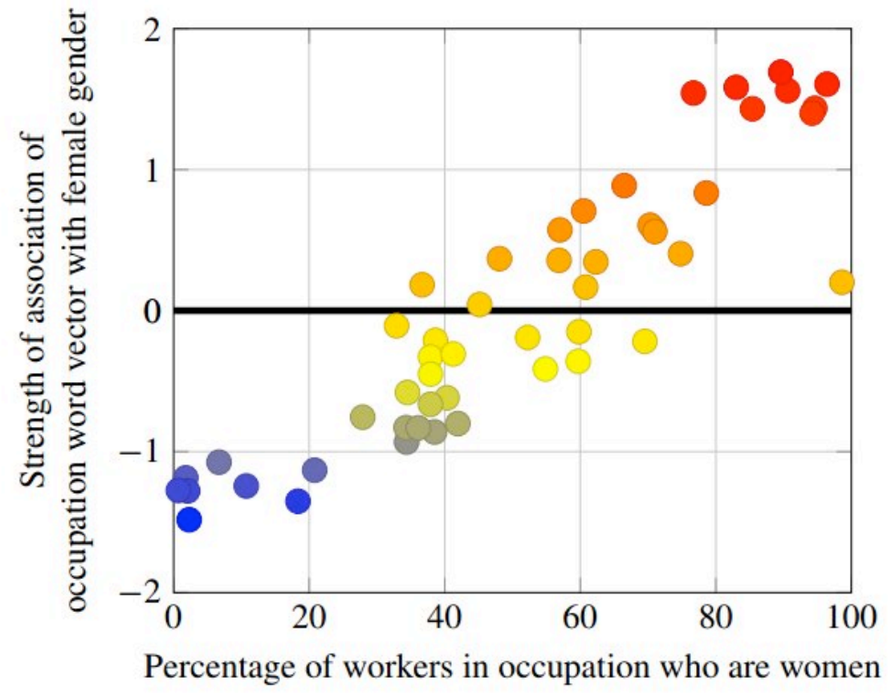


Figure 1. Occupation-gender association
Pearson's correlation coefficient $\rho = 0.90$ with p-value $< 10^{-18}$.

2015 US labor statistics

$\rho = 0.90$

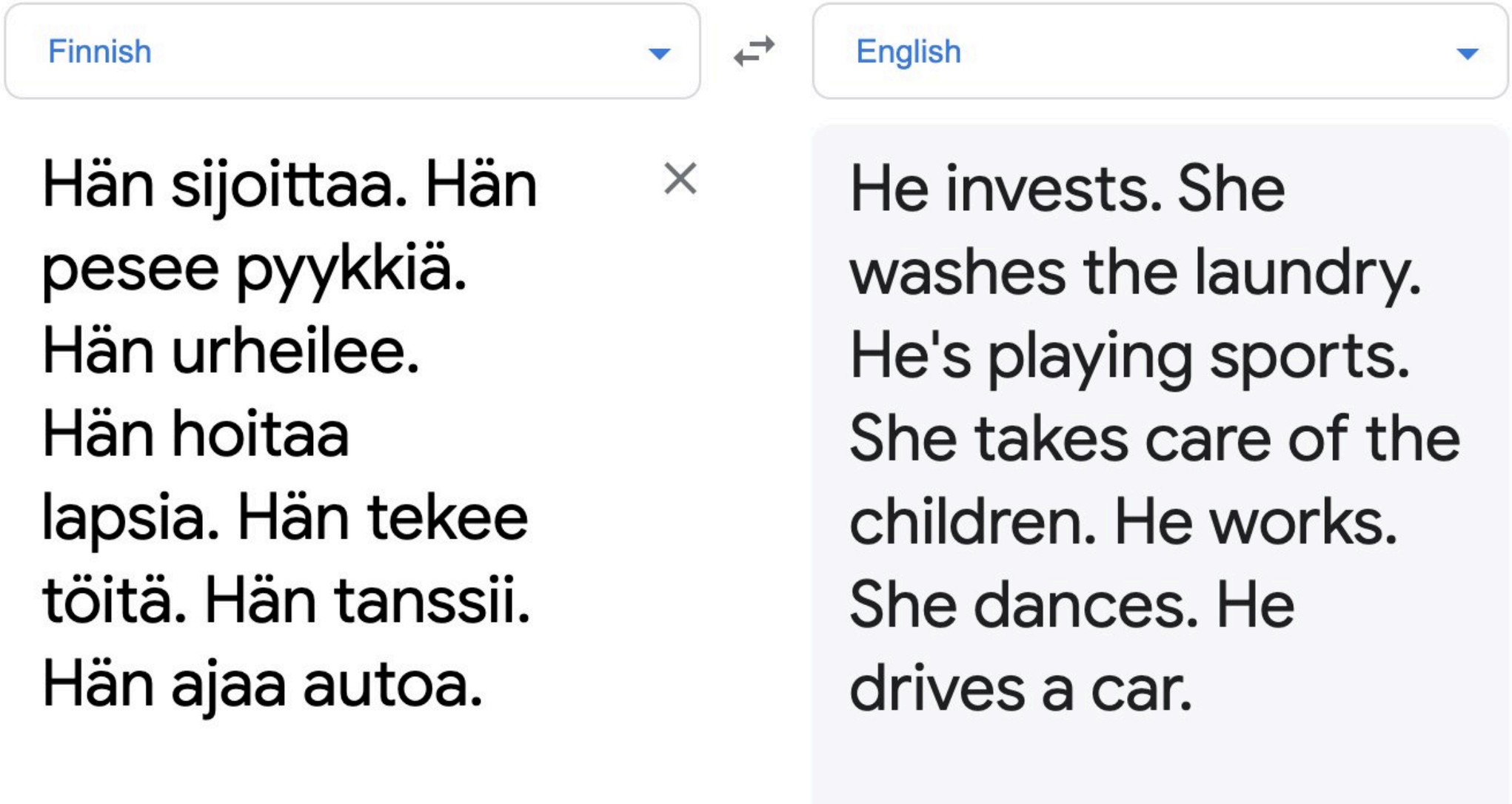
Digital Systems Are Easily Transparent

- What we audit is not the micro details of how AI works, but how humans behave when they build, train, test deploy, and monitor it.
- Good (maintainable) systems engineering of software requires:
  - Architecting the system: design and document its components, processes for development, use, and maintenance.
  - Secure the system. Including logs; provenance of software & data libraries.
  - Document (log) with secure revision control every change to the code base – who made the change, when, and why. For ML, log also data libraries, and model parameters.

cf Bryson OUP 2020, EC's draft AI regulation & digital services act (DSA)

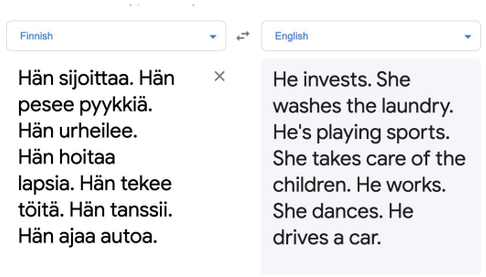
- What we audit is not the micro details of how AI works, but how humans behave when they build, train, test deploy, and monitor it.
- Architecture documents of the system: design of its components, processes for development, use, and maintenance.
- Security documents for the system. Including logs; provenance of software & data libraries.
- Logs of every change to the code base – who made the change, when, and why. For ML, log also data libraries, and model parameters.
- Logs of testing before and during release; and performance – inputs and decisions – of operational systems.
- All benefit the developers, and are auditable cf. EU proposed Digital Services Act (DSA), AI Act (AIA).

cf Bryson OUP 2020



@vuokko recently, though Aylin Caliskan did it first

Translator?



ML
simple, transparent algorithm

stereotyped output

XAI human readable hacks

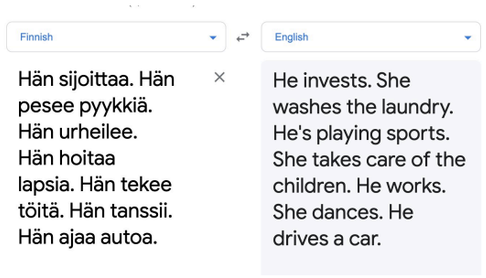
predefined fair output

Replicates lived experience

Tests of completeness documented in design plans

@vuokko recently, though Aylin Caliskan did it first

Translator?



the whole thing is the translator

ML
simple, transparent algorithm

stereotyped output

ML simple, transparent alg.

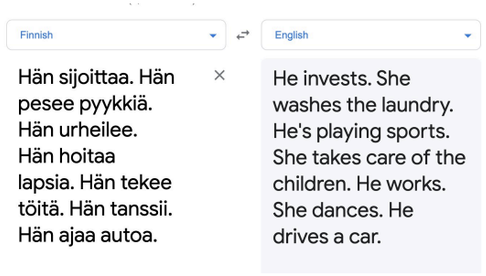
predefined fair output

Replicates lived experience

Tests of completeness documented in design plans

@vuokko recently, 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

Each stage should be auditable and replicable.

Each stage demonstrably meets criteria.

Accountability for AI is possible, but requires reliable enforcement – governance.

Outline

- AI and Ethics
- Human Cooperation and AI

The Intelligence Explosion aka Superintelligence

I J Good (1965)

Nick Bostrom (2014)

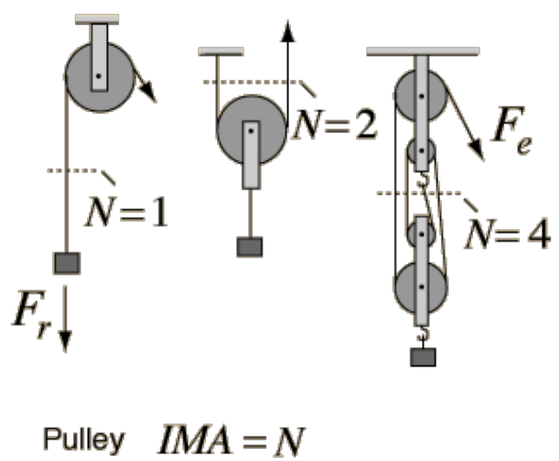
Self improving intelligence – learning to learn.

Exponential growth, eventual domination.

12,000 years of AI

If we accept that intelligence can be decomposed (e.g. action, perception, motivation, memory, learning, reasoning)…
Then every machine and especially writing have been examples of AI.

The "intelligence explosion" is us–
AI-enhanced humans.
boom!



Bryson Collective Agency 2015

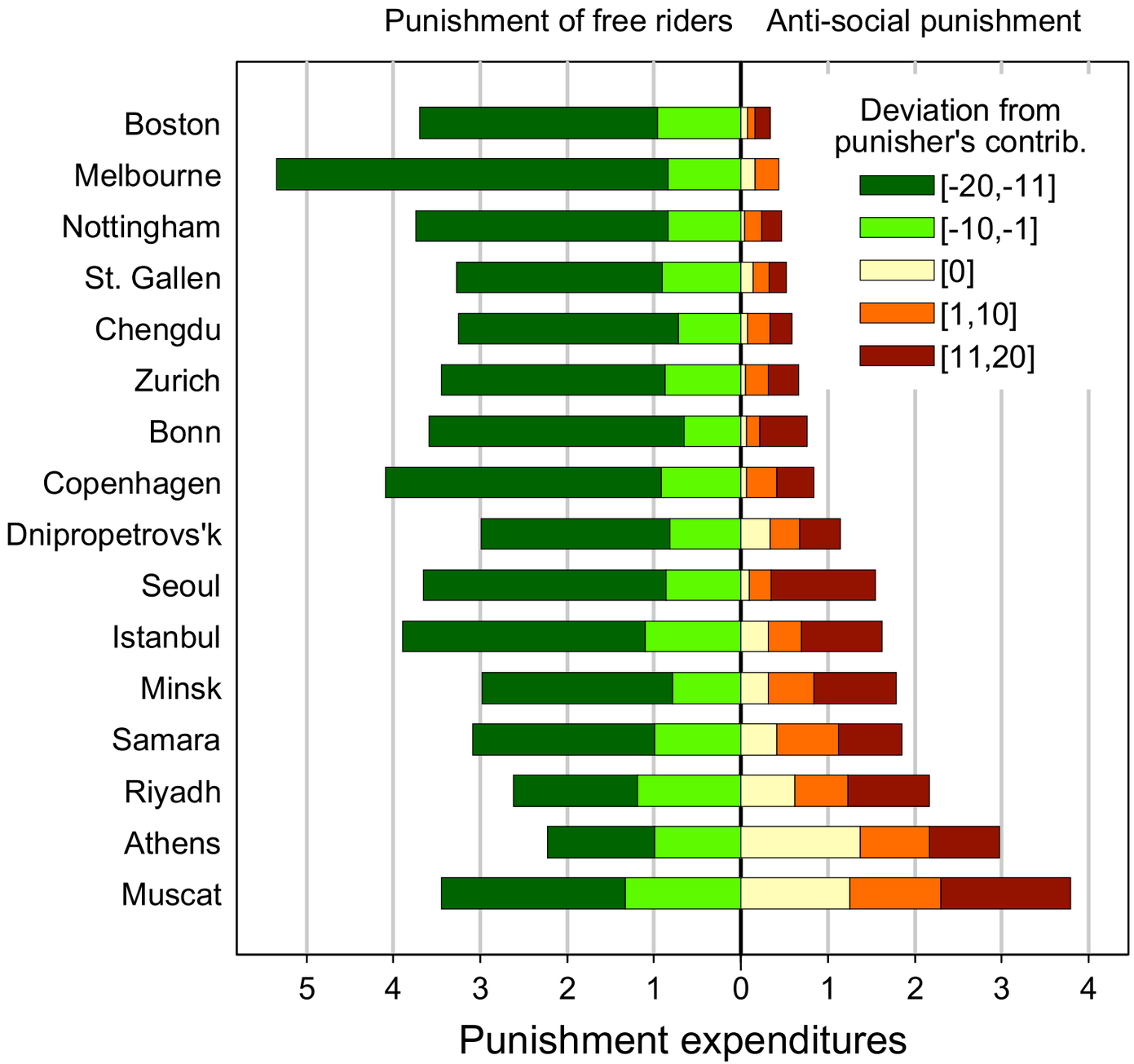
How can we support so many people?

Public Goods

When are we good?

When do we cooperate?

- Fundamentally, two social strategies: sustainability and inequality.
- Sustainability: how big can we make the pie (produce public goods)?
- Inequality: how big a slice of the pie does everyone get?
- Cooperation grows the pie; competition grows the slice.
- We cooperate if we can find a way to grow the pie, and that seems to have a better cost/benefit than competing over a slice.
  - AI is good for search!



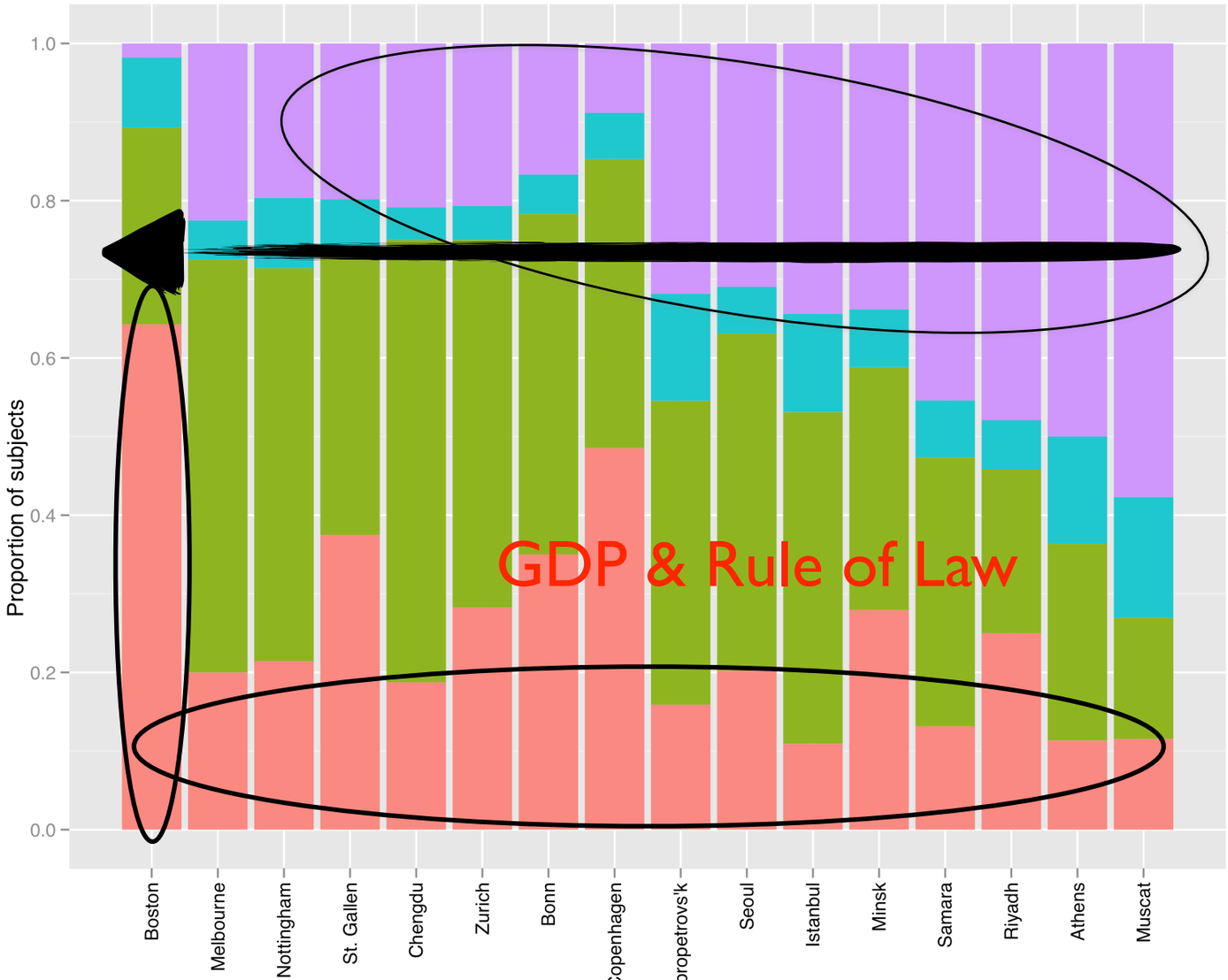
Punishment of Altruism: Expenditures by Relative Contribution

GDP & Rule of Law

(Hermann, Thöni & Gächter 2008)

Punishment in rounds 2-9 of 10

Still using Hermann, Thöni & Gächter's (2008) data



Mostly mixed strategies (4 people per group)

Free riding on punish-ment.

Growing the slice.

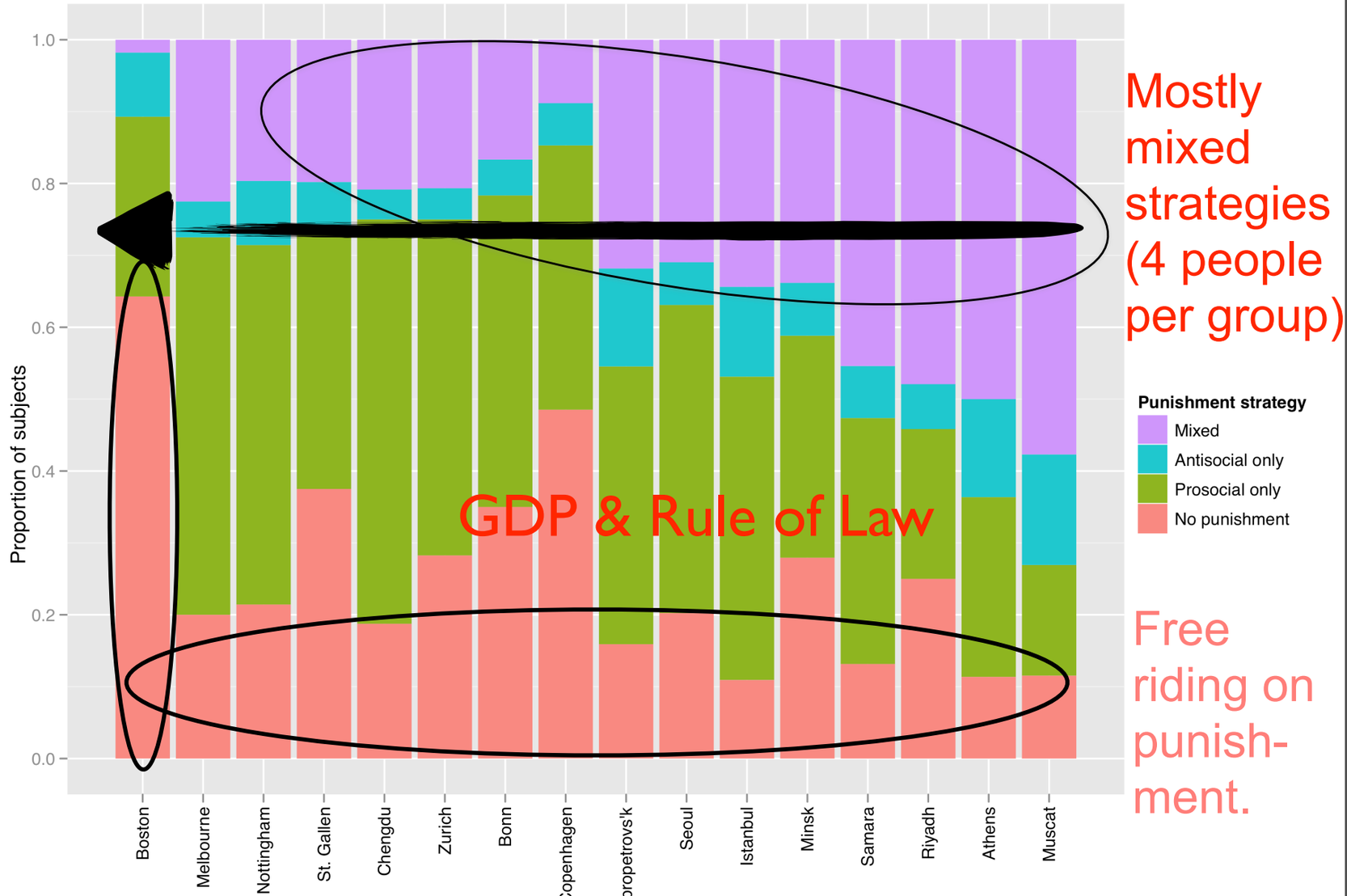
Growing the pie.

Exploiting advantages of the present system.

Sylwester, Mitchell, Lowe & Bryson *in prep*

May be many pies!

Still using Hermann, Thöni & Gächter's (2008) data



Sylwester, Mitchell, Lowe & Bryson *in prep*

Growing their slice.

Growing their pie.

Exploiting advantages of the present system.

If you're good with digital technology, transparency should be easy.

On the Dangers of Stochastic Parrots:


Alessandro Acquisti


Bubacarr Bah


De Kai


Dyan Gibbens


Joanna Bryson


Kay Coles James


Luciano Floridi


William Joseph Burns

The Limits of Transparency

1. Combinatorics
2. Polarisation
3. Multiple, Conflicting Goals

The Limits of Transparency

1. Combinatorics
2. Polarisation
3. Multiple, Conflicting Goals

Wilkinson & Picket 2011

Polarization and the Top 1%

Polarisation

$\propto$

Inequality

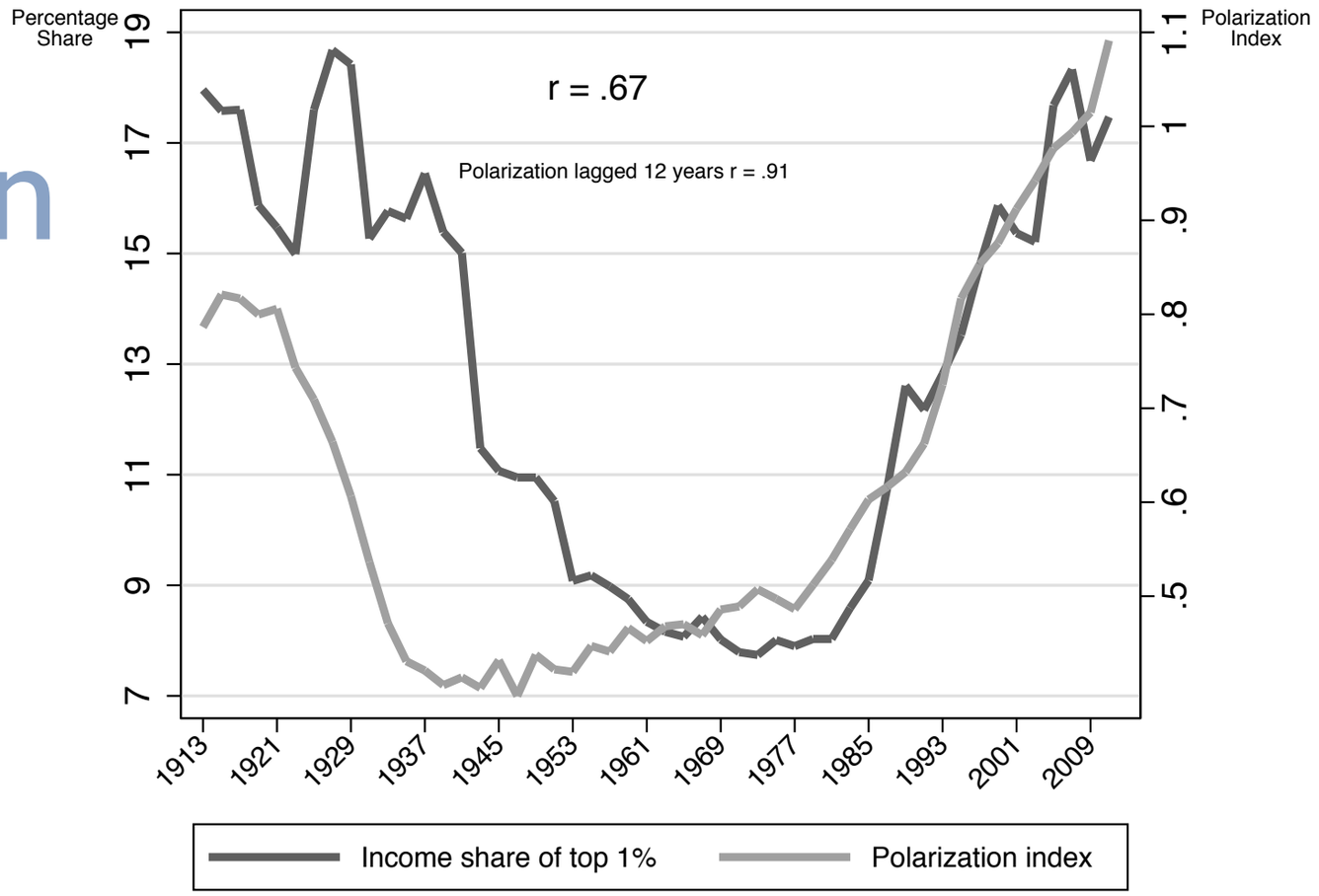
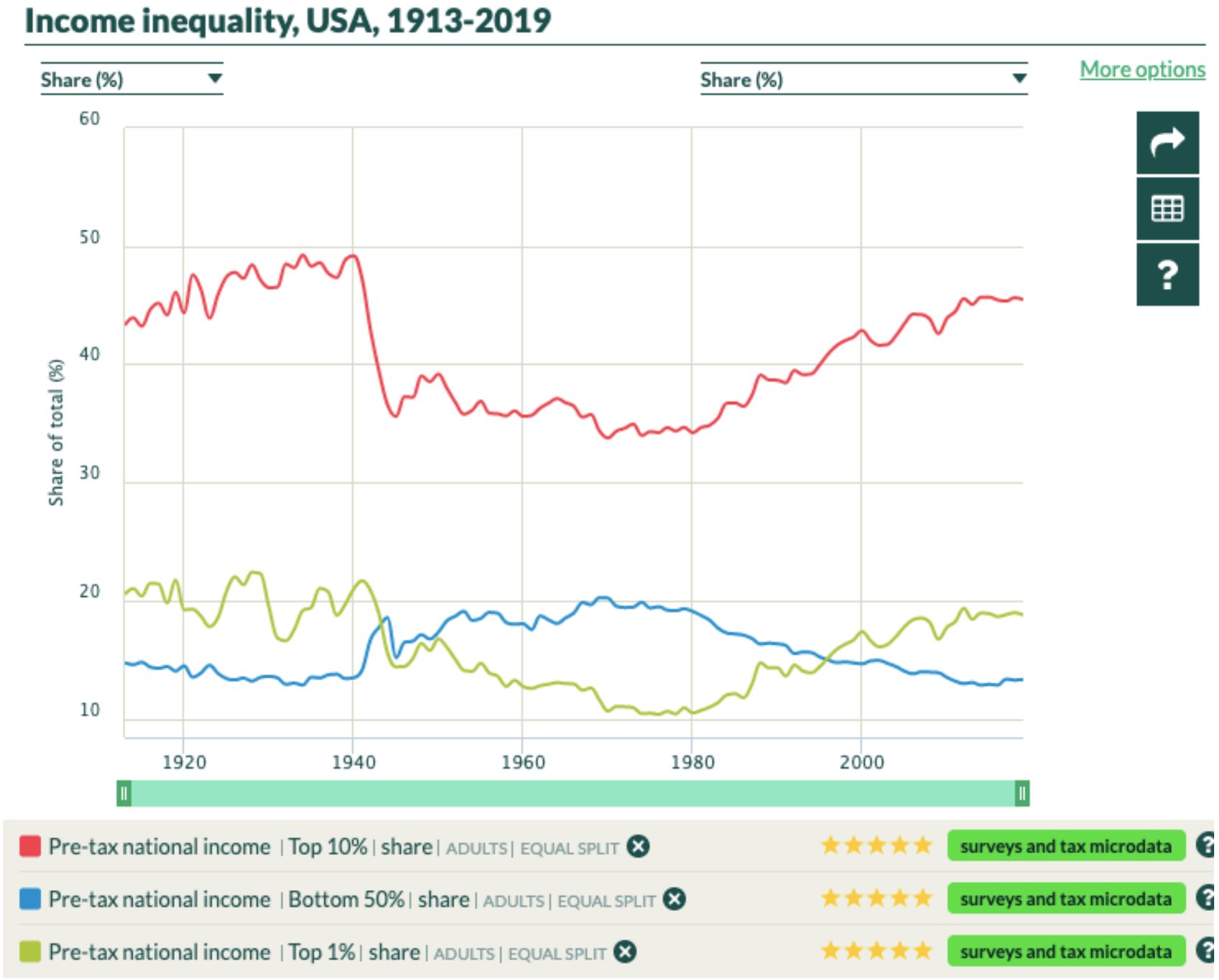


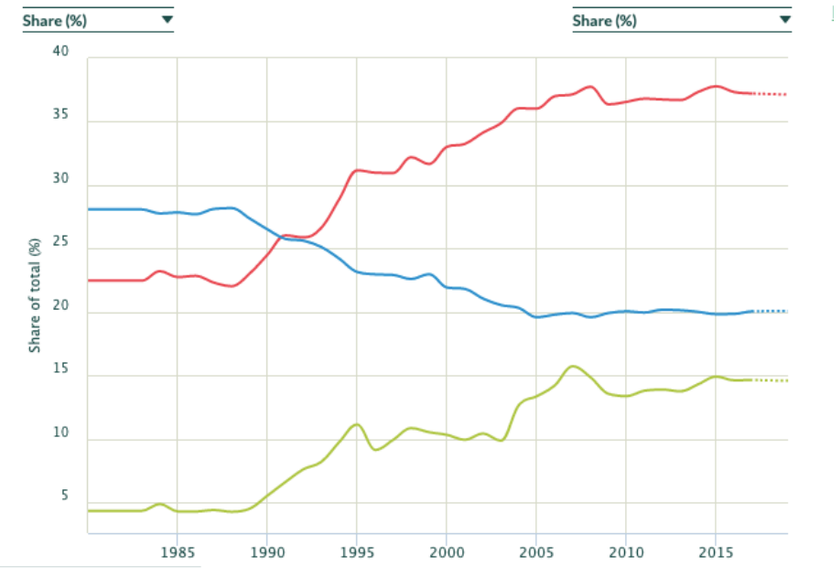
Figure 1.2: Top One Percent Income Share and House Polarization

U.S.A.

http://wid.world/country/usa

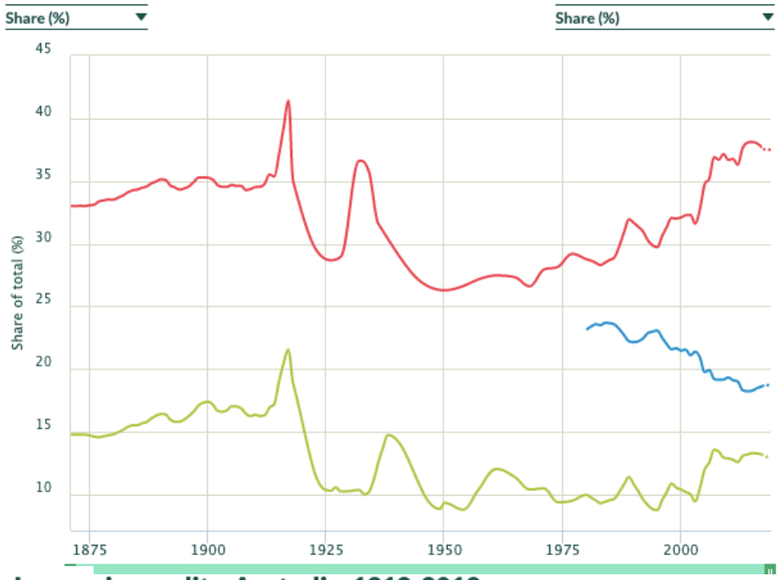


Income inequality, Poland, 1980-2019



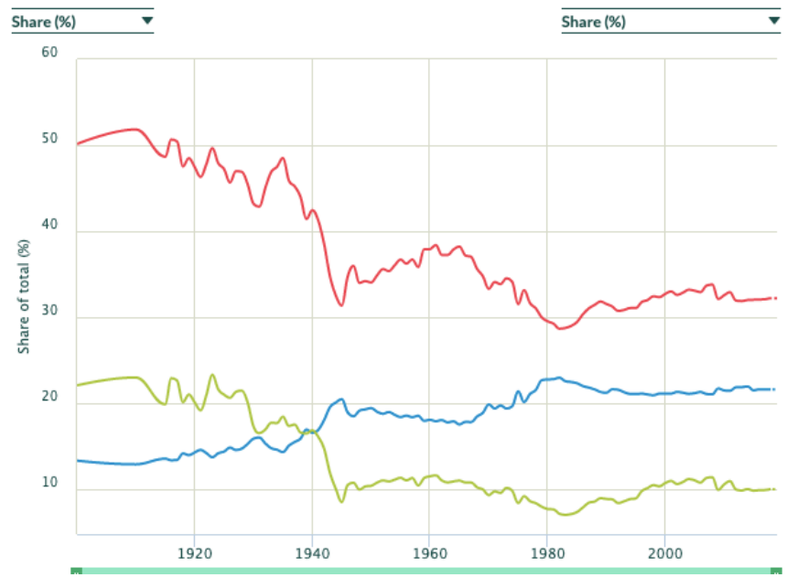
Pre-tax national income | Top 10% | share | ADULTS | EQUAL SPLIT
Pre-tax national income | Bottom 50% | share | ADULTS | EQUAL SPLIT
Pre-tax national income | Top 1% | share | ADULTS | EQUAL SPLIT

Income inequality, Germany, 1871-2019



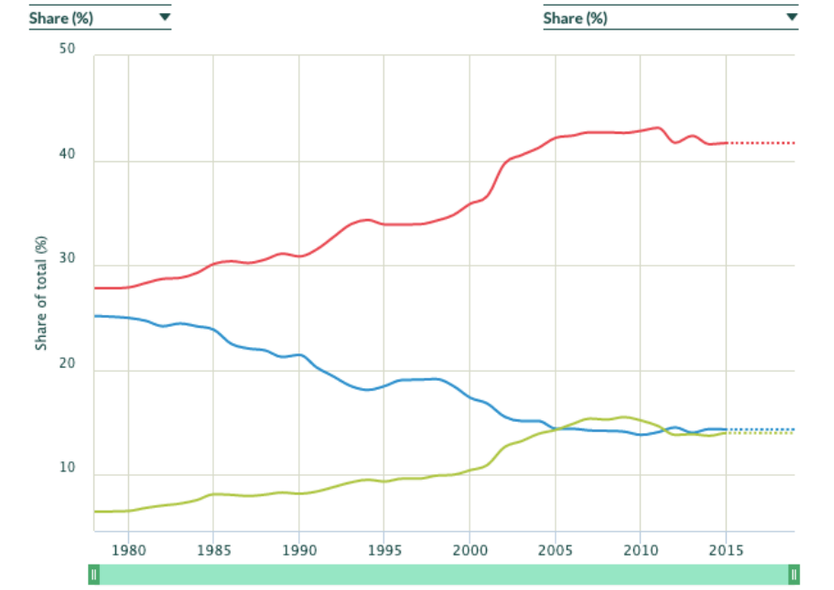
Pre-tax national income | Top 10% | share | ADULTS | EQUAL SPLIT
Pre-tax national income | Bottom 50% | share | ADULTS | EQUAL SPLIT
Pre-tax national income | Top 1% | share | ADULTS | EQUAL SPLIT

Income inequality, France, 1900-2019



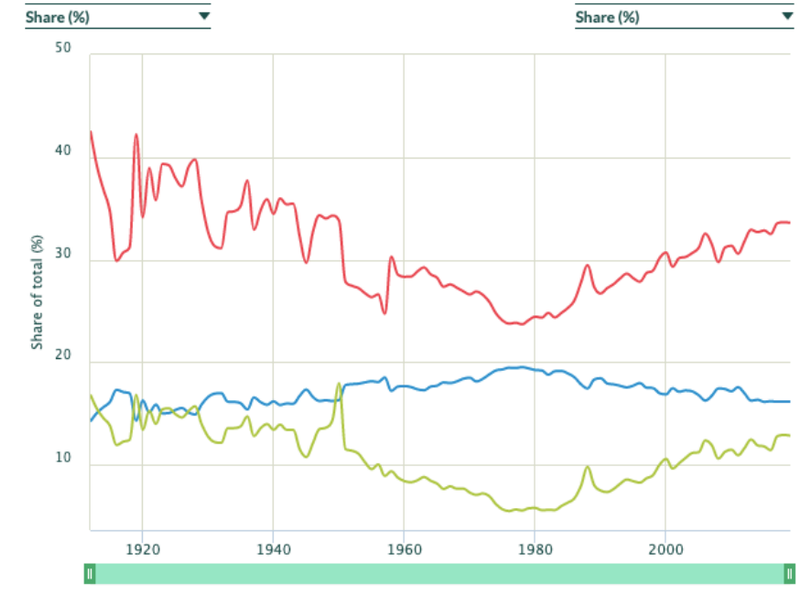
Pre-tax national income | Top 10% | share | ADULTS | EQUAL SPLIT
Pre-tax national income | Bottom 50% | share | ADULTS | EQUAL SPLIT
Pre-tax national income | Top 1% | share | ADULTS | EQUAL SPLIT

Income inequality, China, 1978-2019



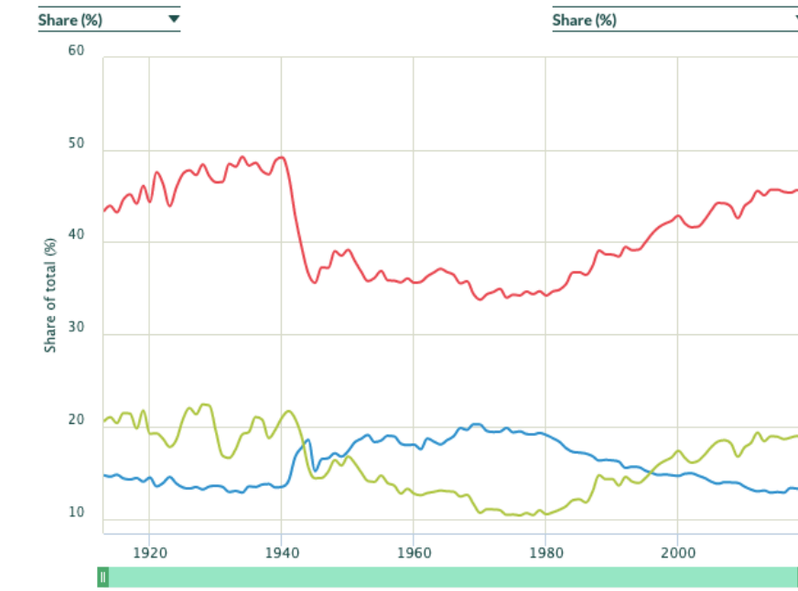
Pre-tax national income | Top 10% | share | ADULTS | EQUAL SPLIT
Pre-tax national income | Bottom 50% | share | ADULTS | EQUAL SPLIT
Pre-tax national income | Top 1% | share | ADULTS | EQUAL SPLIT

Income inequality, Australia, 1912-2019



Pre-tax national income | Top 10% | share | ADULTS | EQUAL SPLIT
Pre-tax national income | Bottom 50% | share | ADULTS | EQUAL SPLIT
Pre-tax national income | Top 1% | share | ADULTS | EQUAL SPLIT

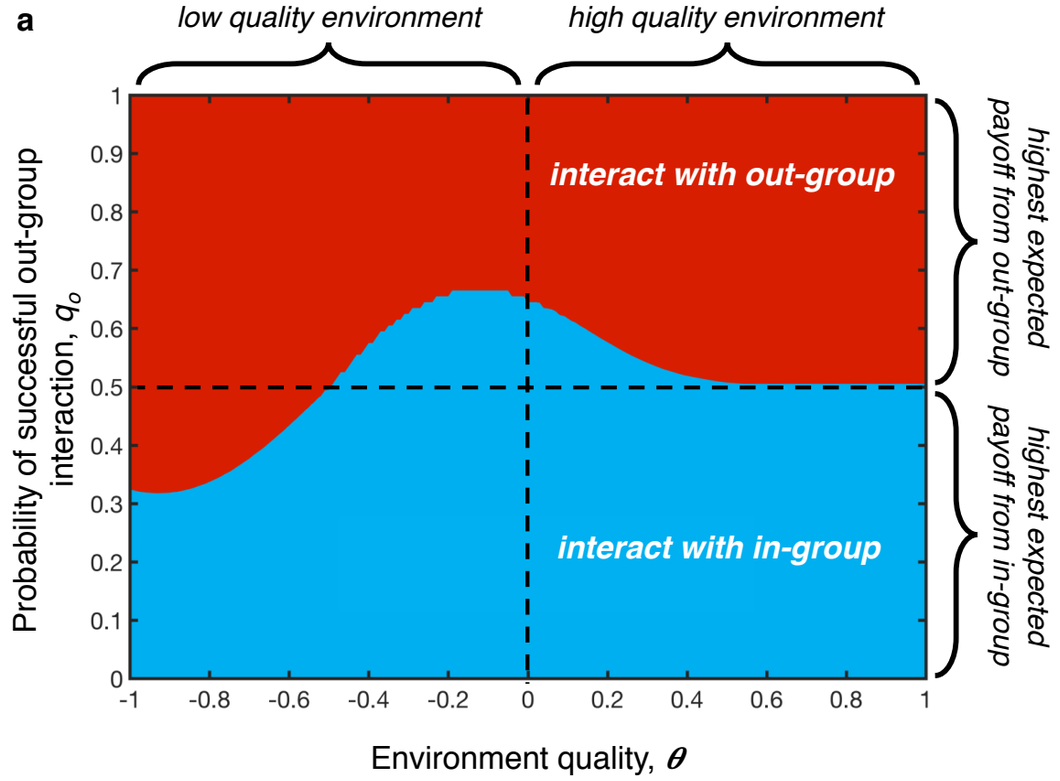
Income inequality, USA, 1913-2019



Pre-tax national income | Top 10% | share | ADULTS | EQUAL SPLIT
Pre-tax national income | Bottom 50% | share | ADULTS | EQUAL SPLIT
Pre-tax national income | Top 1% | share | ADULTS | EQUAL SPLIT

Why and When Inequality Causes Polarisation

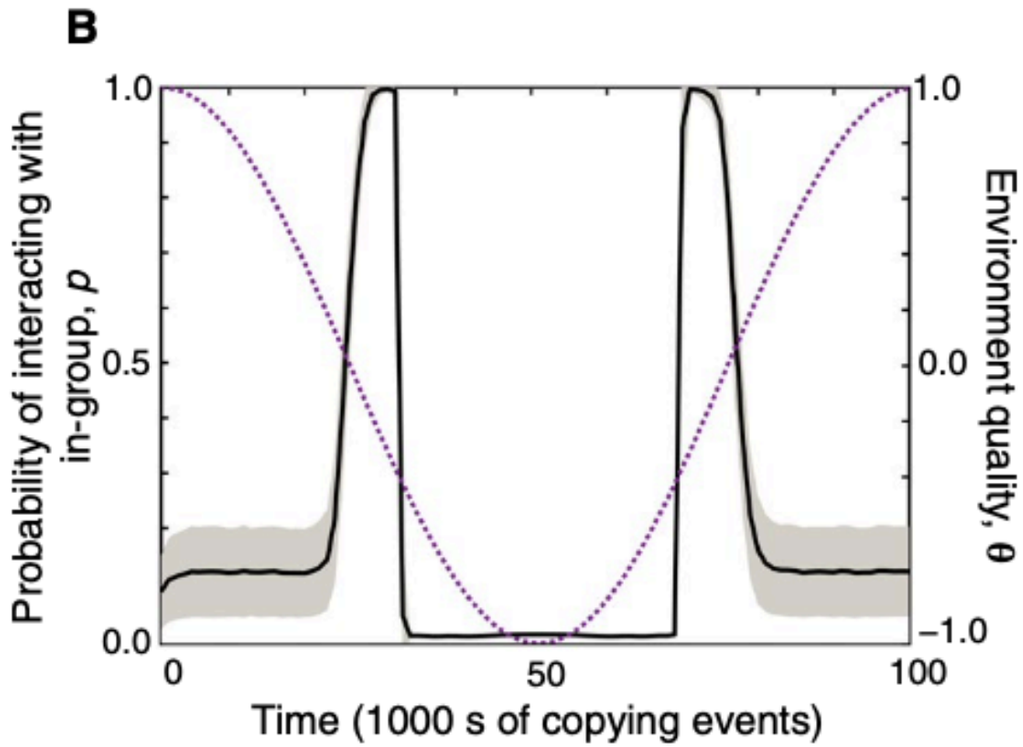
- Model assumptions: In-group cooperation has lower risk but outgroup diversity has higher expected outcome even so.
- Model outcomes: when an economy offers poor support, avoiding risk can become more important than maximizing revenue. Inequality triggers this when it creates false scarcities.
- Caused by discontinuity, e.g. fear of bankruptcy, foreclosure, divorce, losing children, starvation, etc.



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

Results: Recovery

If one person can choose who to cooperate with.



The Limits of Transparency

1. Combinatorics
2. Polarisation
3. Multiple, Conflicting Goals

If you're good with digital technology, transparency should be easy.

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

Why can't the world's leading communication company communicate even internally?

- Presumably, they can (up to the limits of time, space & energy).
- But what if some actors had the highest priority of maintaining agency (so their company could act),
  - and maintaining "first mover" advantage (being biggest) seemed to them existentially necessary (fear of "kill zones".)
- And other actors were hired to ensure ethical integrity.
- Apparent breakdown of transparency might be a logical impasse.

What can be done?

1. Combinatorics
2. Polarisation
3. Multiple, Conflicting Goals

What can be done?

- Combinatorics: Cooperate, build computers, ultimately intractable.
- Polarisation: Reduce vulnerability through adequate infrastructure.
- Multiple conflicting goals:
  - Iterative design – what my PhD dissertation was about (for AI).
  - What governance and politics are all about (social sciences FTW).
- Breath – it's a form of regulation.
  - Perpetuation benefits from diversity and oscillations.
  - Also ultimately intractable, but life has been going for billions of years.

Outline

- AI and Ethics
- Human Cooperation and AI

Conclusions

- Artificial Intelligence is not a component of human peers. Rather it's a component of designed systems, which are themselves components of the entities that develop and own them, and maybe hackers too.
- Fortunately, such systems can be made transparent and accountability can be traced through them, but only with adequate regulation and enforcement.
- Cooperation and focussing on long term sustainability is a perfectly sensible strategy for security, which does not exclude growth in well being.

Helena Malikova

Nolan McCarty

Alex Stewart

Tom Dale Grant

Mihailis E. Diamantis

Arvind Narayanan

Aylin Caliskan