Artificial Intelligence and Human Accountability

Joanna J. Bryson
University of Bath, United Kingdom

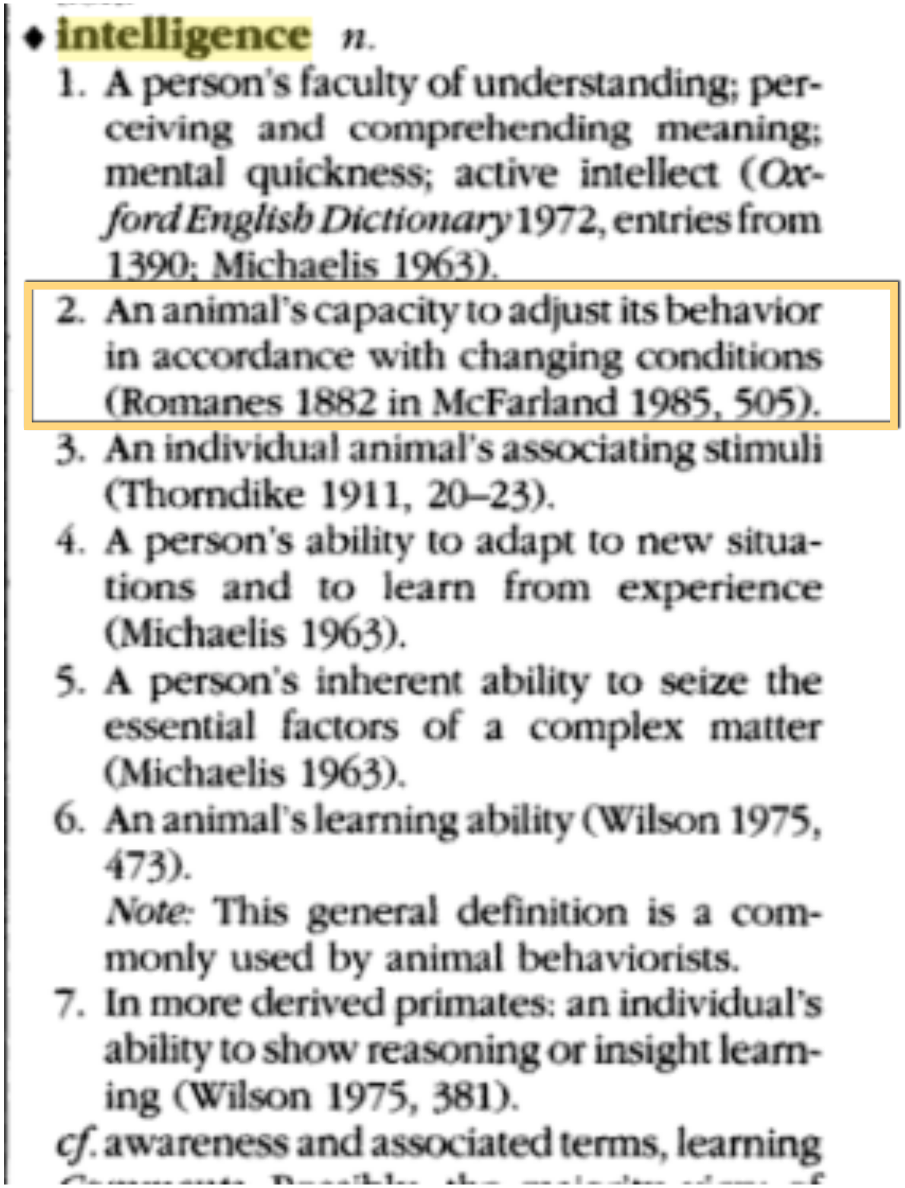
@j2bryson

Definitions

for communicating right now

- Intelligence is doing the right thing at the right time.
  - A form of computation (not math)–transforms perception into action. Requires time, space, and energy.
- 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.
- Ethics is the set of behaviours that creates and sustains a society, including by defining its identity.
- Artificial intelligence is an artefact, built intentionally.

Romanes, 1883 – Animal Intelligence, a seminal monograph in comparative psychology.



◆ **intelligence** *n.*

1. A person's faculty of understanding; perceiving and comprehending meaning; mental quickness; active intellect (*Oxford English Dictionary* 1972, entries from 1390; Michaelis 1963).
2. An animal's capacity to adjust its behavior in accordance with changing conditions (Romanes 1882 in McFarland 1985, 505).
3. An individual animal's associating stimuli (Thorndike 1911, 20–23).
4. A person's ability to adapt to new situations and to learn from experience (Michaelis 1963).
5. A person's inherent ability to seize the essential factors of a complex matter (Michaelis 1963).
6. An animal's learning ability (Wilson 1975, 473).

   Note: This general definition is a commonly used by animal behaviorists.
7. In more derived primates: an individual's ability to show reasoning or insight learning (Wilson 1975, 381).

cf. awareness and associated terms, learning

Comments: Possibly, the majority view of

Definitions for communicating right now

- Intelligence is doing the right thing at the right time.
  - A form of computation (not math)–transforms sensing into action. Requires time, space, and energy.
- 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.

} Moral subjects

- Ethics is the set of behaviours that creates and sustains a society, including by defining its identity.

} Including but not limited to general principles

- Artificial intelligence is an artefact; built intentionally.

} Intent ⟹ responsibility.

AI is built by humans, with or without machine learning.

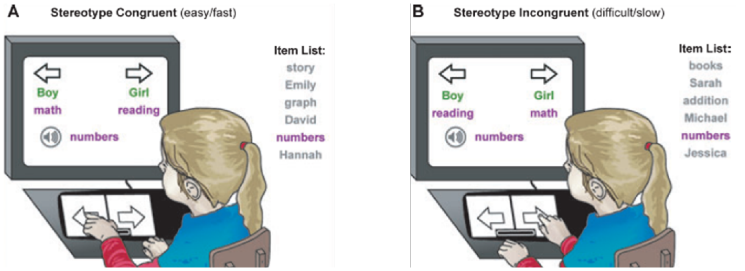
We're responsible.

- How AI is built, how it is trained (and on what), how it is tested, monitored–all things for which humans can be held to account.
- Every aspect of developing and operating AI can be logged; we can demand evidence of due diligence.

AI Trained on Human Language Replicates Our 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 planning and cooperation.

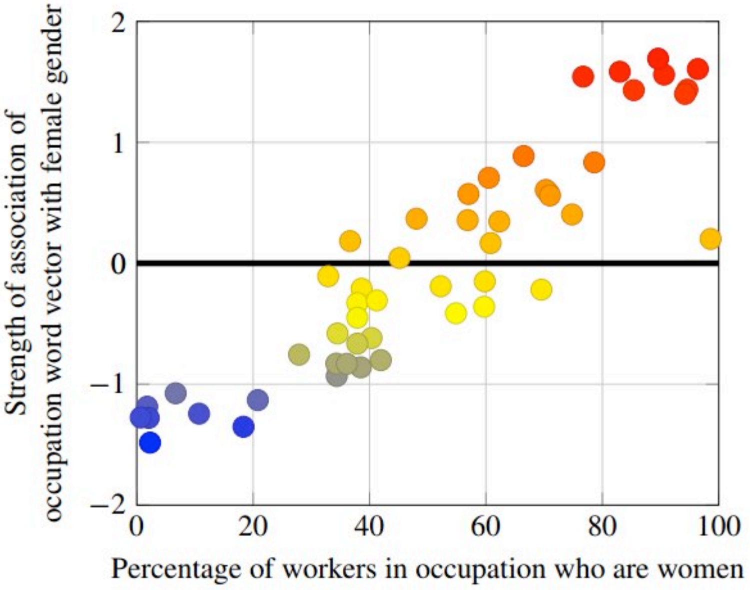


Figure 1. Occupation-gender association
Pearson's correlation coefficient $\rho = 0.90$ with p-value $< 10^{-18}$.

2015 US labor statistics
ρ = 0.90

At Least Three Sources of AI Bias

- Implicit: Absorbed automatically using machine learning on data from ordinary culture.
- Accidental: Introduced through ignorance by insufficiently diverse or careful development teams.
- Deliberate: Introduced intentionally as a part of the development process (planning or implementation.)

How to deal with them

- Implicit–compensate with design, architecture (see also accidental).
- Accidental–diversify work force, test, log, iterate, improve.
- Deliberate–audits, regulation.

salon

Donald Trump, Nigel Farage (AP/Andrew Harnik/Reuters/Vincent Kessler/Ph

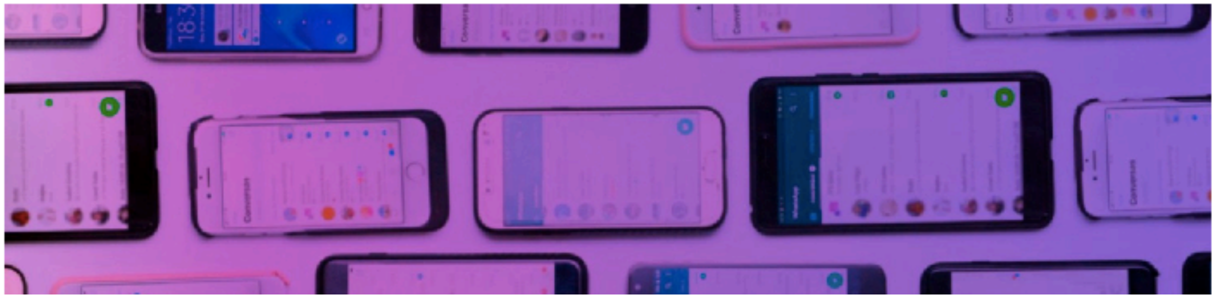
Computational Power: Automated Use of WhatsApp in the Elections

Study on the use of automation tools to boost political campaigns digitally in the 2018 Brazilian elections

ITS Rio Follow

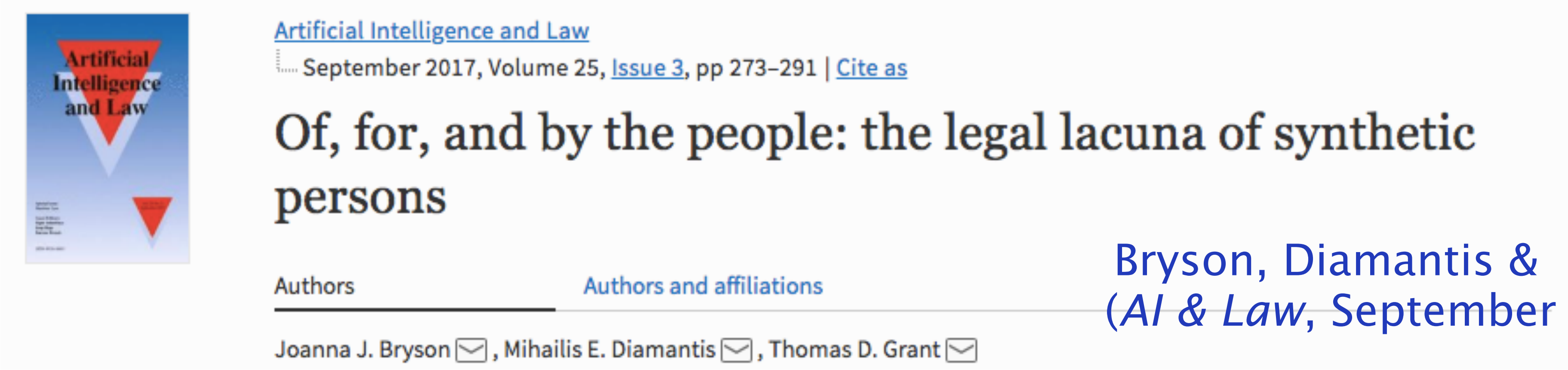
Oct 26, 2018 · 22 min read

Caio Machado and **Marco Konopacki**



Only Humans Can Be Accountable

- Law and Justice are more about dissuasion than recompense.
- Safe, secure, accountable software systems are modular – suffering* in such is incoherent. *e.g. systemic dysphoria of isolation, loss of status or wealth.
    - No penalty of law enacted directly against an artefact (including a shell company) can have efficacy.



Artificial Intelligence and Law

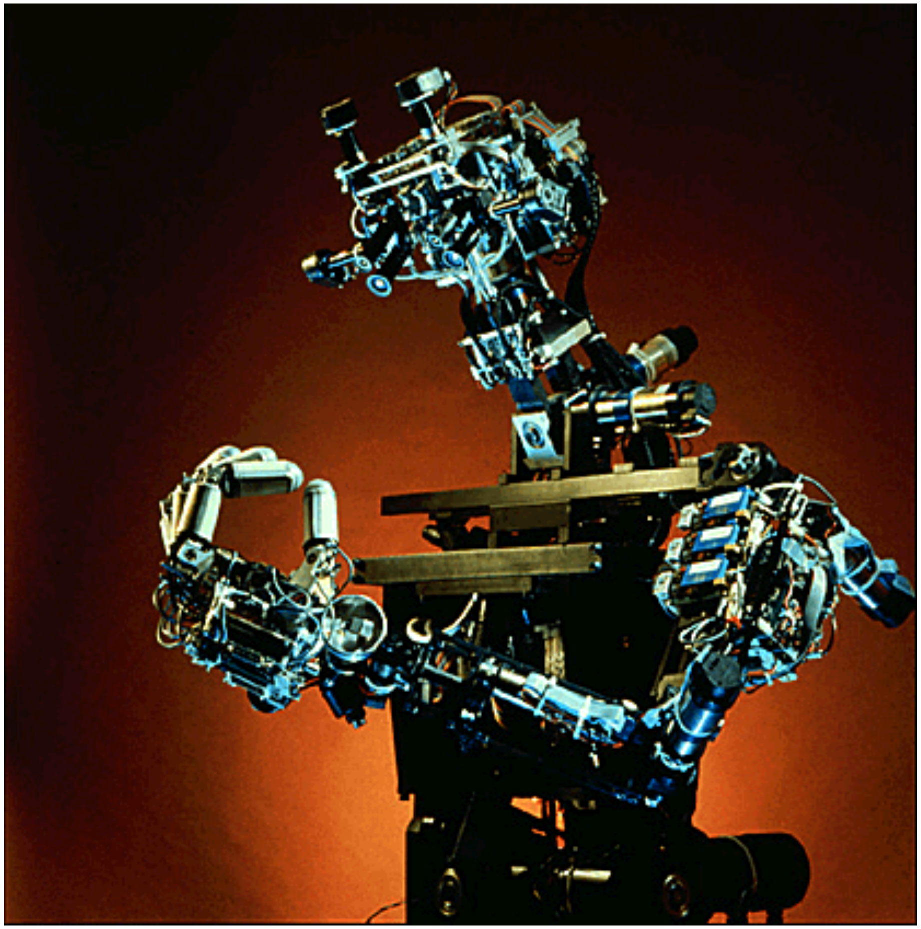
September 2017, Volume 25, Issue 3, pp 273–291 | Cite as

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

Authors Authors and affiliations

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

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



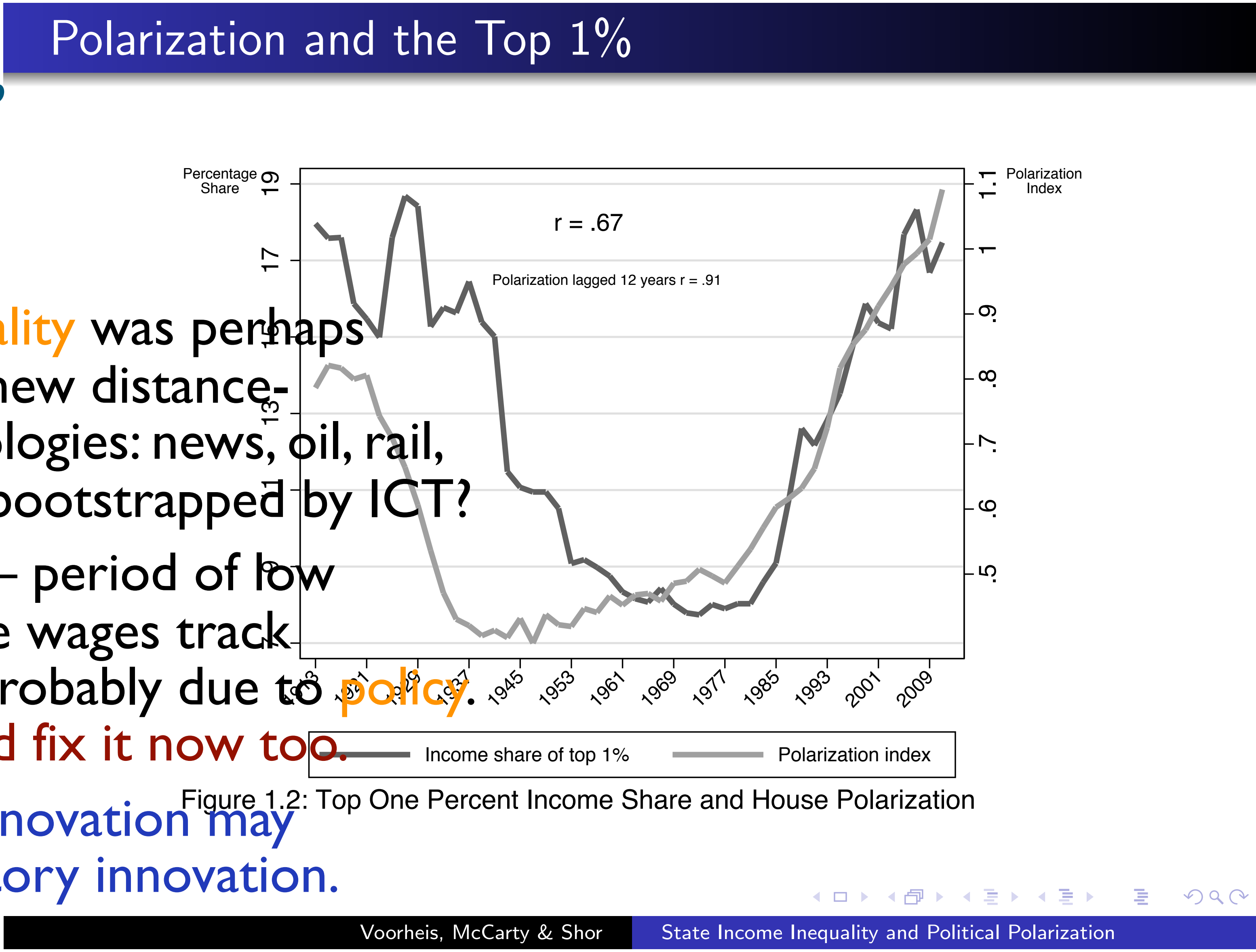
People want to make AI they can be friends with, fall in love with, will their fortunes to – “equals” over which we have complete dominion.

This is (arguably) both sick and dangerous.

Bryson & Kime 1998; 2011
Kathleen Richardson 2016

Economically, we've been here before.

- Late 19C inequality was perhaps driven by then-new distance-reducing technologies: news, oil, rail, telegraph; now bootstrapped by ICT?
- Great coupling – period of low inequality where wages track productivity – probably due to policy. Implies we could fix it now too.
- Technological innovation may mandate regulatory innovation.



Polarization and the Top 1%

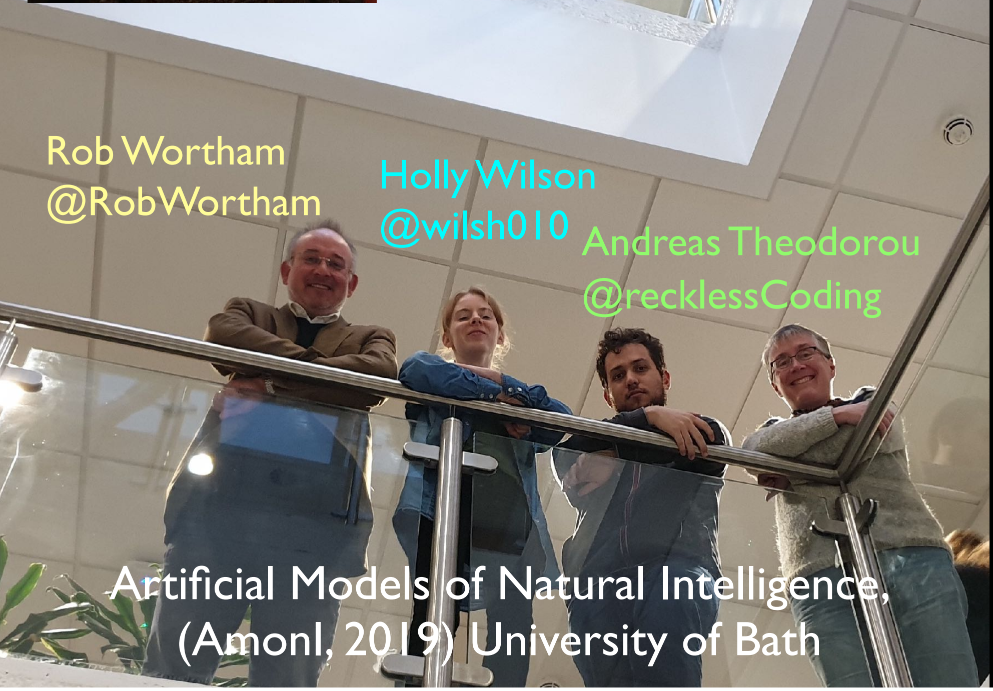
Figure 1.2: Top One Percent Income Share and House Polarization

Voorheis, McCarty & Shor State Income Inequality and Political Polarization

Thanks to my collaborators, and to you for your attention.

Aylin Caliskan
@aylin_cim

Arvind Narayanan
@random_walker



Rob Wortham
@RobWortham

Holly Wilson
@wilsh010

Andreas Theodorou
@recklessCoding

Artificial Models of Natural Intelligence,
(AmonI, 2019) University of Bath

Mihailis E.
Diamantis

Nolan McCarty
@nolan_mc

Alex
Stewart
@al_cibiades

Tom Dale Grant

AmonI, 2016