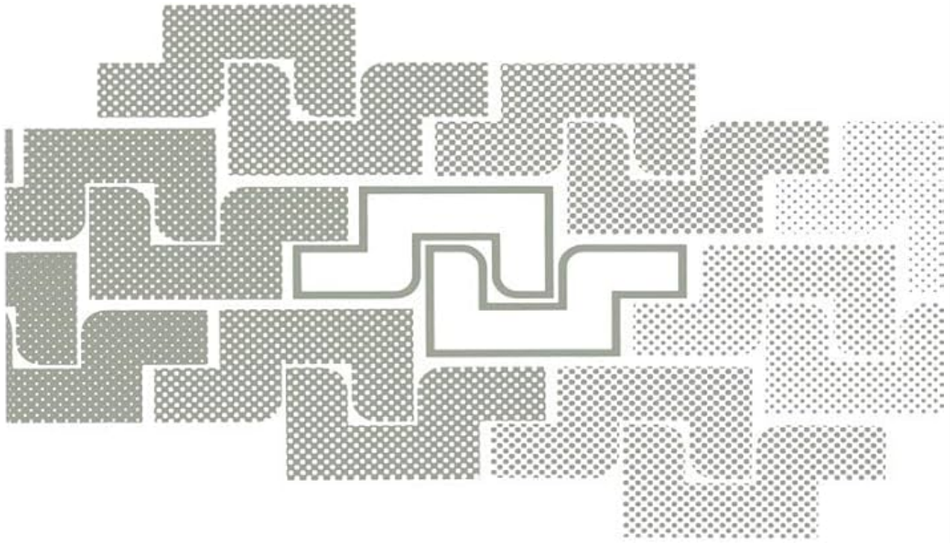
The Limits of Artificial Agency

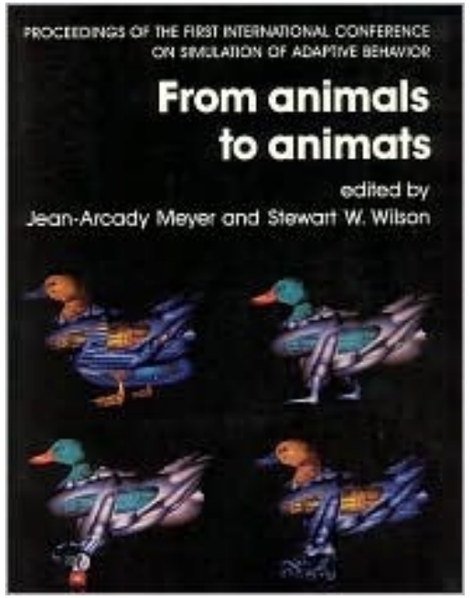
Hertie School
Centre for
Digital Governance

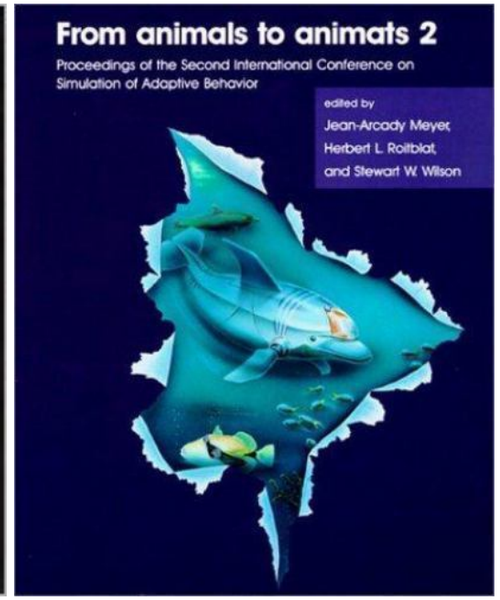
Joanna J. Bryson
@j2bryson.bsky.social
mastodon.social/@j2bryson
linkedin.com/in/bryson

Outline

- Autonomy Through the Ages
- AI and Human Autonomy in Theory
- Autonomy in Practice







Simulation of Adaptive Behaviour

1990-

18th Edition is in Berlin this October!

1993

The Biology and Technology of Intelligent Autonomous Agents

Edited by
Luc Steels

NATO ASI Series

Series F: Computer and Systems Sciences, Vol. 144

"The conference focuses on characterizing and comparing organizational principles and architectures underlying adaptive behavior in animals and animats. Animats denote the conceptual connections between animals and synthetic agents."

The Society for the Study of Artificial Intelligence and the Simulation of Behaviour

Memorial: 30 June, Brighton

OBITUARY | 12 August 2025 | Correction 13 August 2025

Margaret Boden obituary: cognitive scientist who explored how machines might emulate human imagination

Pioneering artificial-intelligence scholar, whose influential work bridged cognitive science, philosophy and computer science.

By Joanna Bryson

Intelligent Agents IV

(Bryson, ATAL 1997)

First publication for what became Behaviour Oriented Design (PhD 2001)

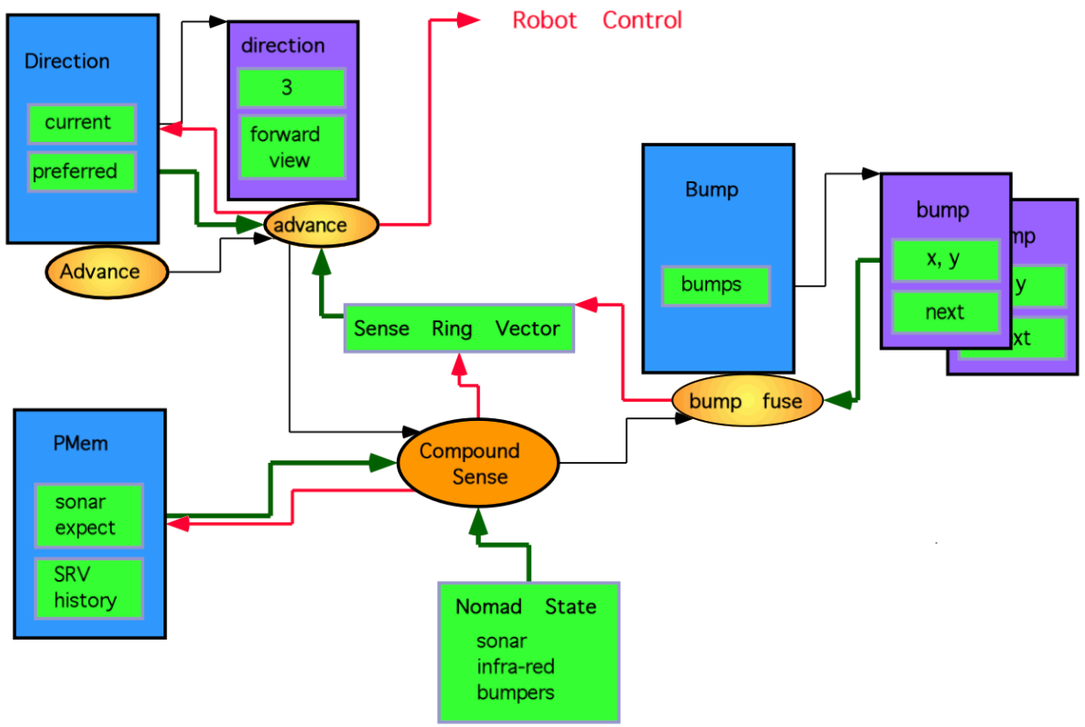


Fig. 4. The state supporting the Direction library function **advance**.

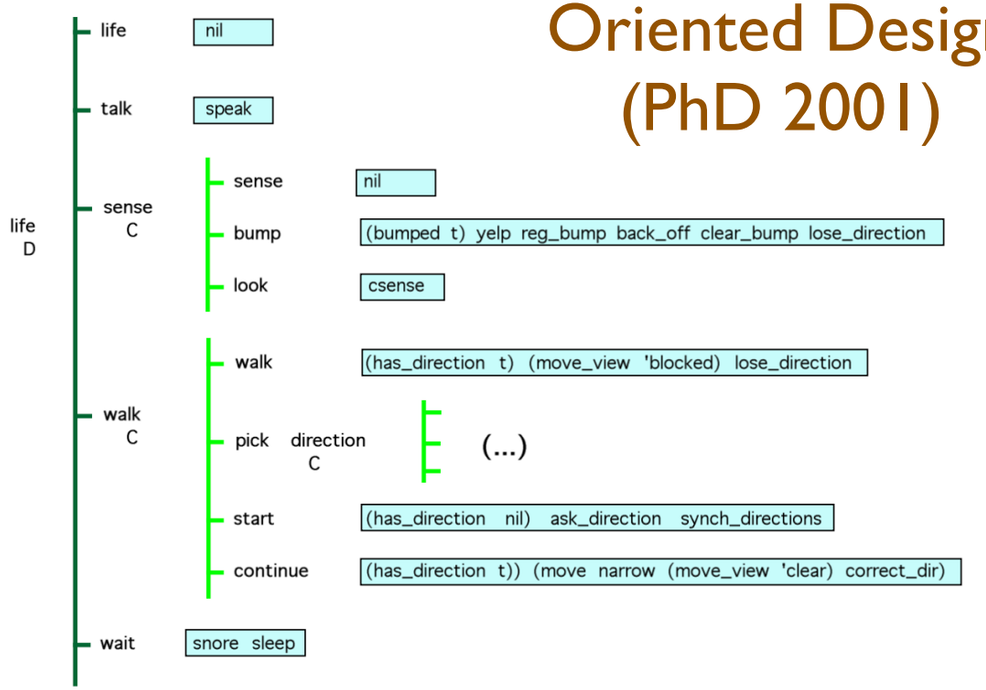
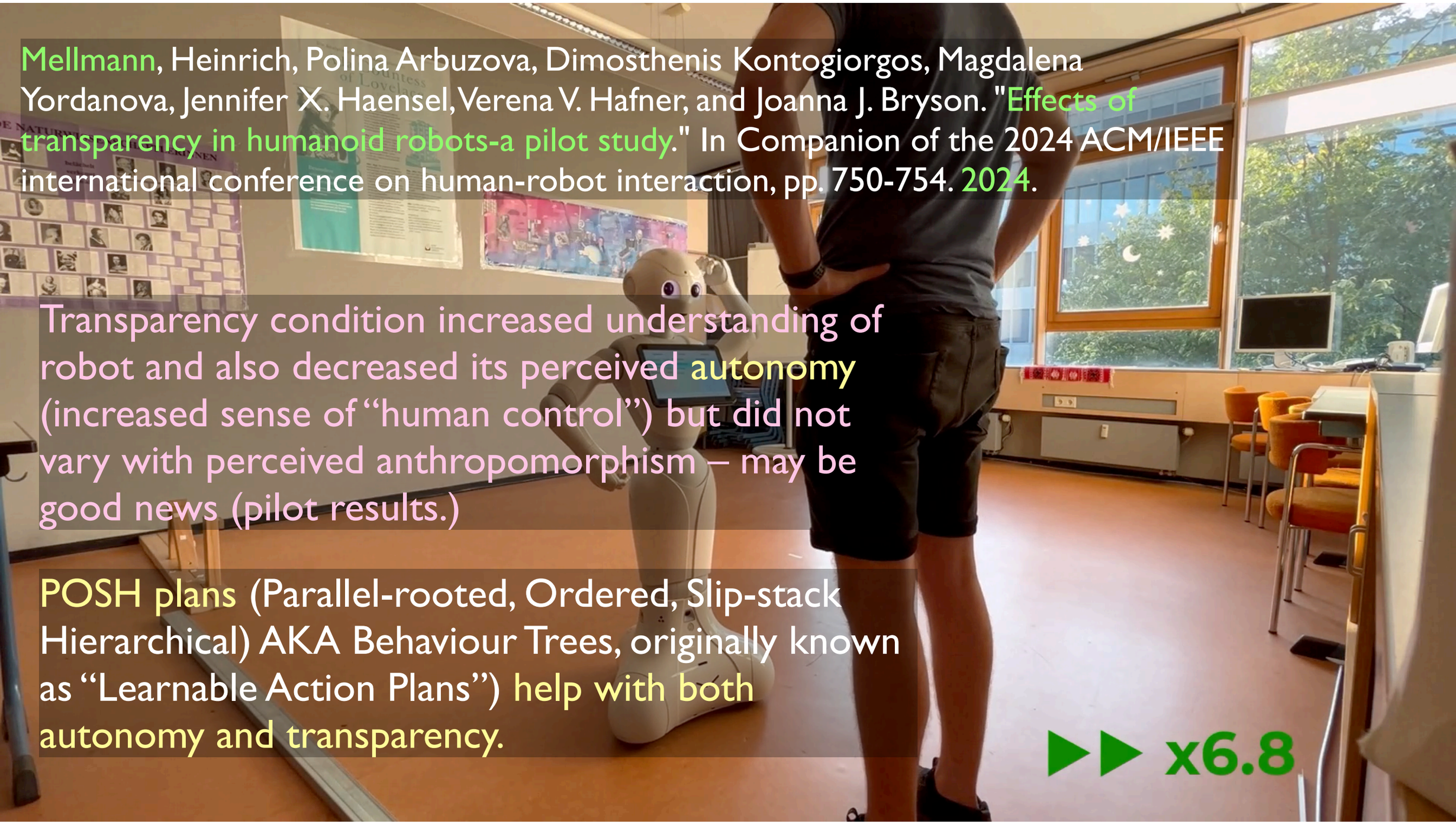


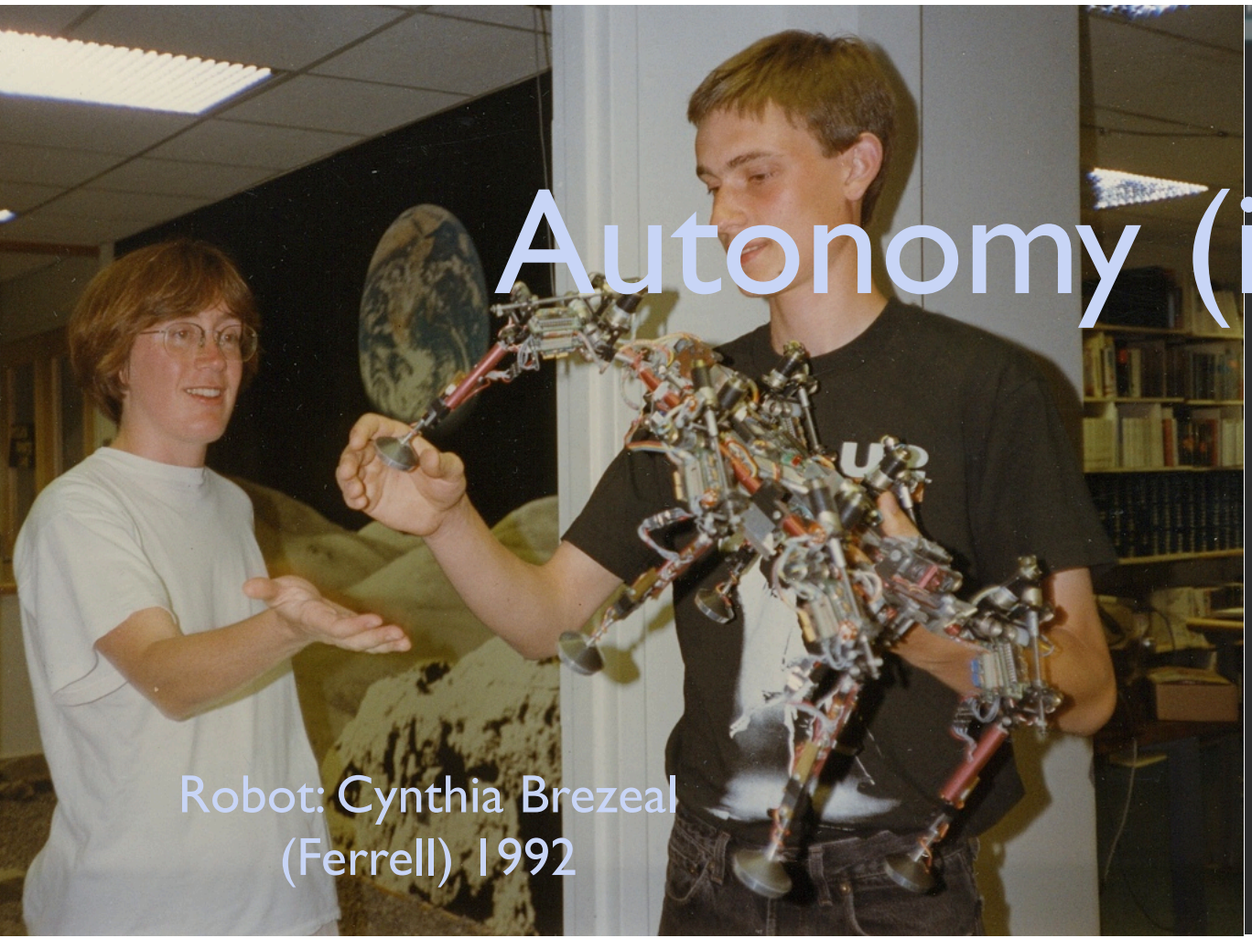
Fig. 2. A simple control heirarchy with no loops. This controls a navigating mobile robot. Priorities are not shown, but initial priority is ordered from top to bottom. Drives are labeled **D**, Competences **C**. Senses are paired with test values, unpaired primitives are Acts.

Behaviour Modules provide agency (sensing, action, memory)

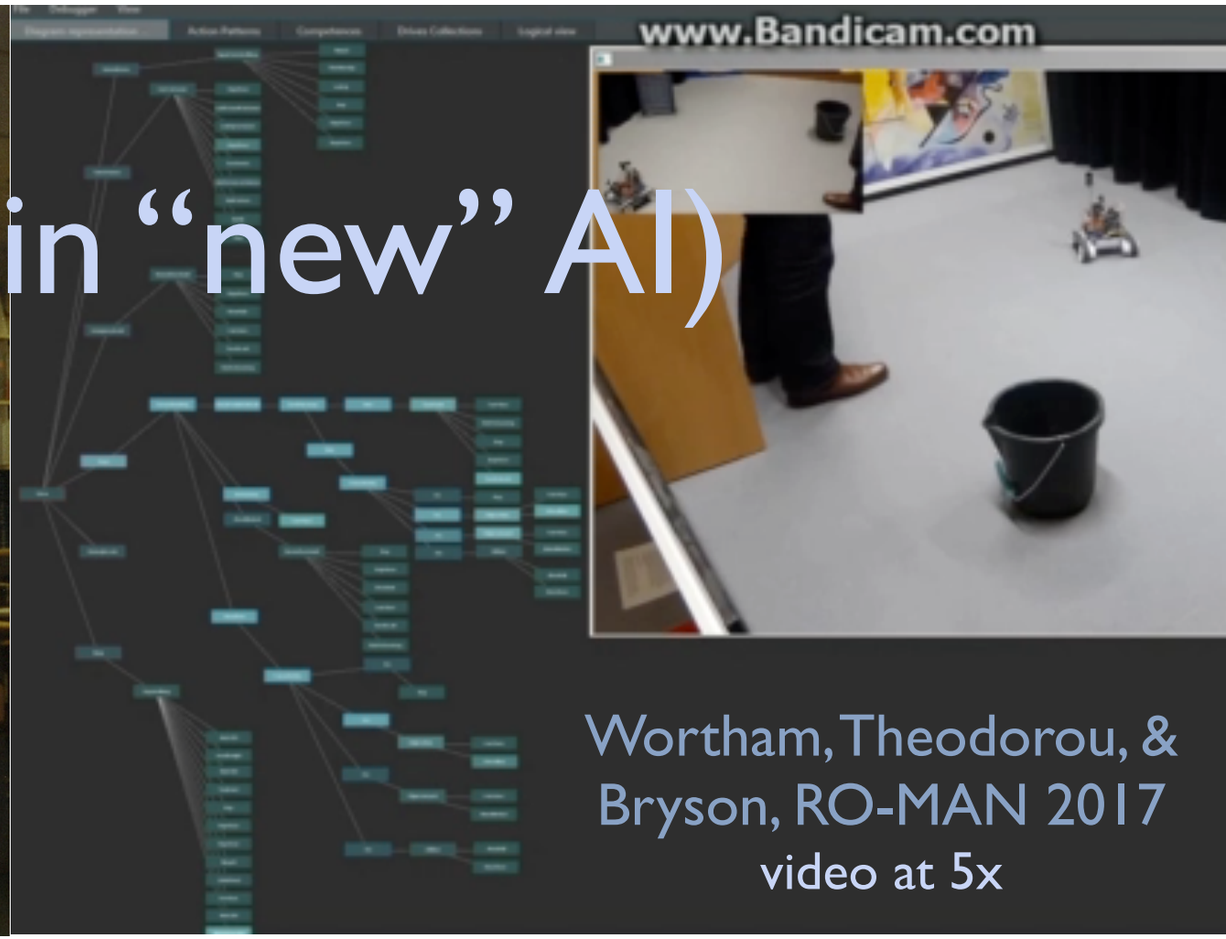
Structured Reactive/Dynamic Plans provide autonomy (priorities)


Mellmann, Heinrich, Polina Arbuzova, Dimosthenis Kontogiorgos, Magdalena Yordanova, Jennifer X. Haensel, Verena V. Hafner, and Joanna J. Bryson. "Effects of transparency in humanoid robots-a pilot study." In Companion of the 2024 ACM/IEEE international conference on human-robot interaction, pp. 750-754. 2024.
Transparency condition increased understanding of robot and also decreased its perceived autonomy (increased sense of "human control") but did not vary with perceived anthropomorphism – may be good news (pilot results.)
POSH plans (Parallel-rooted, Ordered, Slip-stack Hierarchical) AKA Behaviour Trees, originally known as "Learnable Action Plans") help with both autonomy and transparency.
▶▶ x6.8

Autonomy (in "new" AI)



Robot: Cynthia Brezeal (Ferrell) 1992



Wortham, Theodorou, & Bryson, RO-MAN 2017
video at 5x

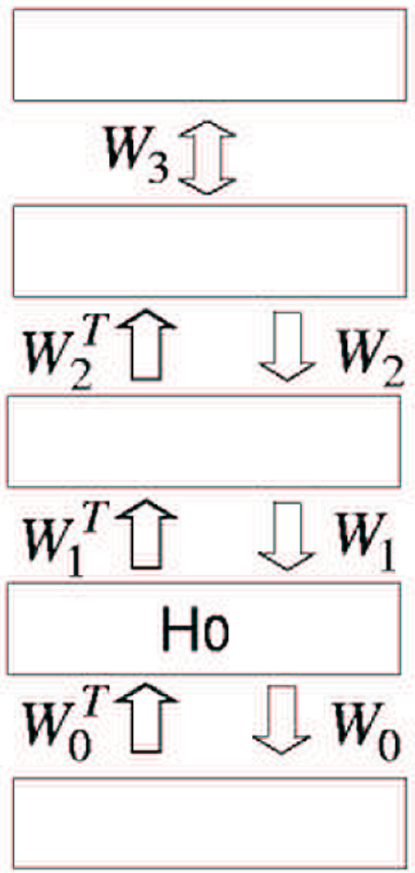
Robots that proactively determine their own behaviour.

Probably a total anthropomorphic myth, (but) a lot of people have been obsessed with it.

Autonomy in International Relations

- Autonomy is the extent to which a nation controls what happens within its own borders.
- No nation is fully autonomous – share an ecosystem and borders. Other nations affect costs and access to resources.
  - People within your borders may be influenced by people outside your borders.
  - Most people have multiple identities (e.g. religion), or may disregard national authority/identity for other reasons.
- Autonomy is a continuum. Even Sovereignty is seen to consist of having (sufficient) mutual interdependencies.

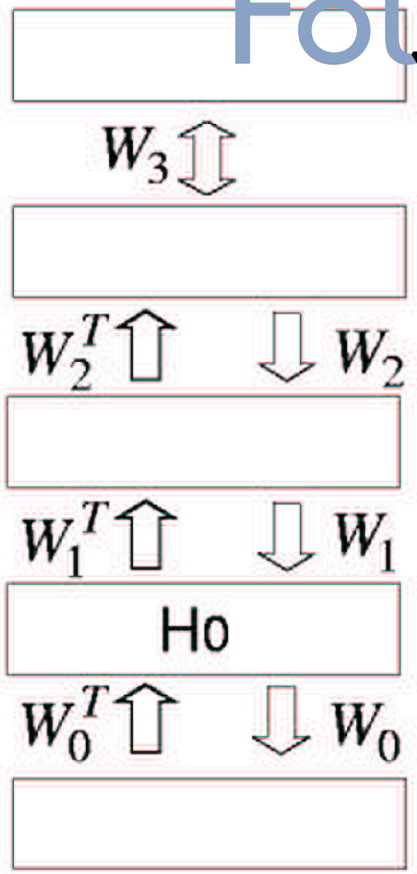
Deep Learning, including Trans-former Systems Are Also Engineered



Hinton, Osindero & Teh 2006

- Hinton helped 'save' neural networks in the 1980s with back propogation – enabling multilayered model architectures.
- Deep learning really took off with good methods (systems engineering + mathematics) for facilitating boot-strapping, then locking in and exploiting (mutual-informing) layer specialisation.
  - Of course hardware (GPU, from games) & data (from smartphones – since 2007) also help.
- Transformers too (Vaswani &*al* 2017)

Foundation Models and 'Guardrails'

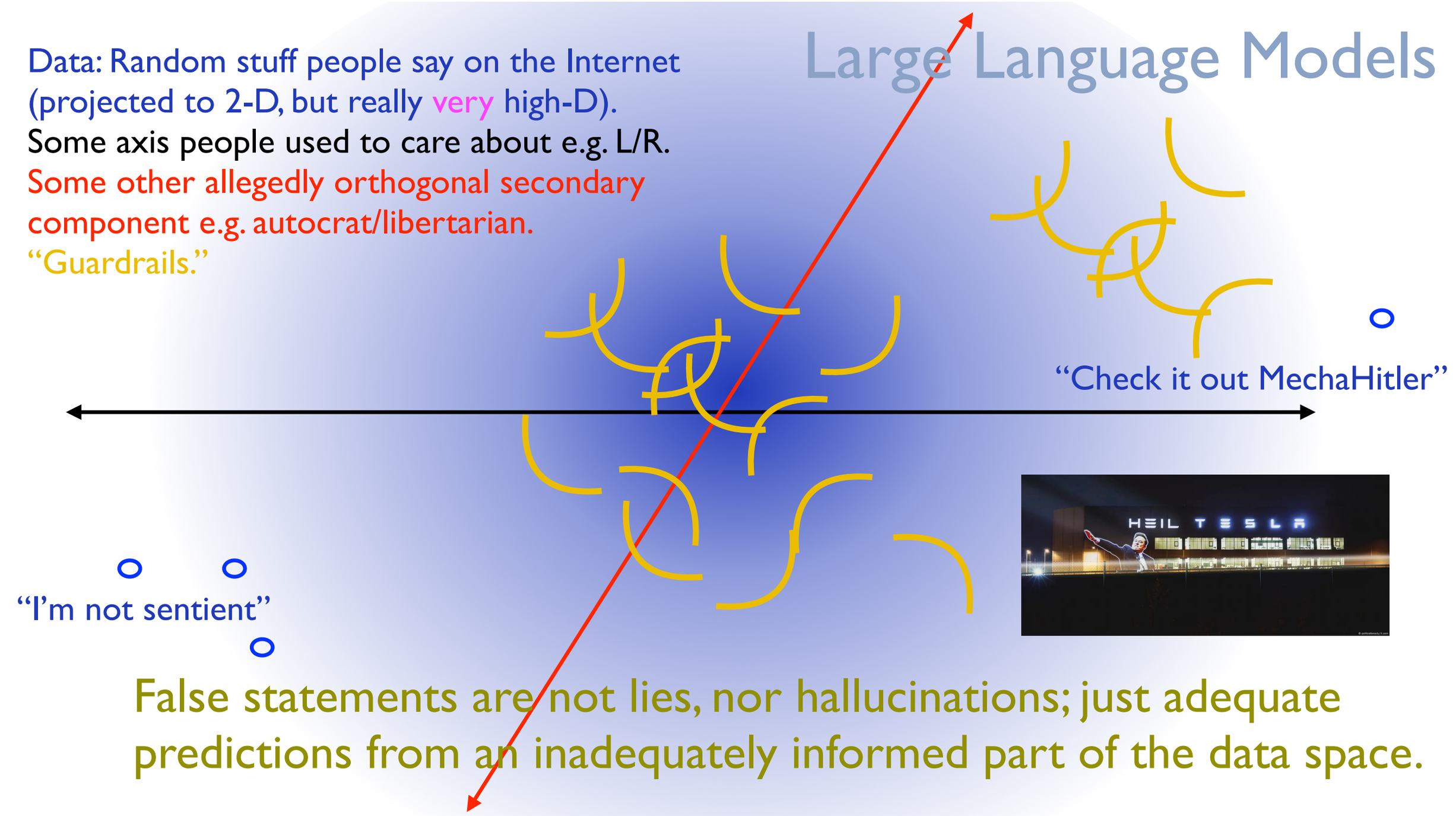


Josh Dzieza (2023)

- 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.
  - Trained using reinforcement learning / brute force.

Large Language Models

Data: 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 data space.

Foundation Model – Derived From Human Culture

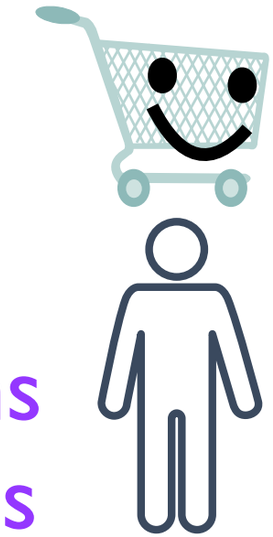
Guardrails – Designed and Trained

← user-specific memory

← user

Foundation Model – Derived From Human Culture

Guardrails – Designed and Trained



← multiplied by 2 Billion

By no means autonomous

(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 help facilitate the observed enormous increases in inequality, and with it political instability / collapse, loss of capacity to address sustainability crises, war crimes &c.

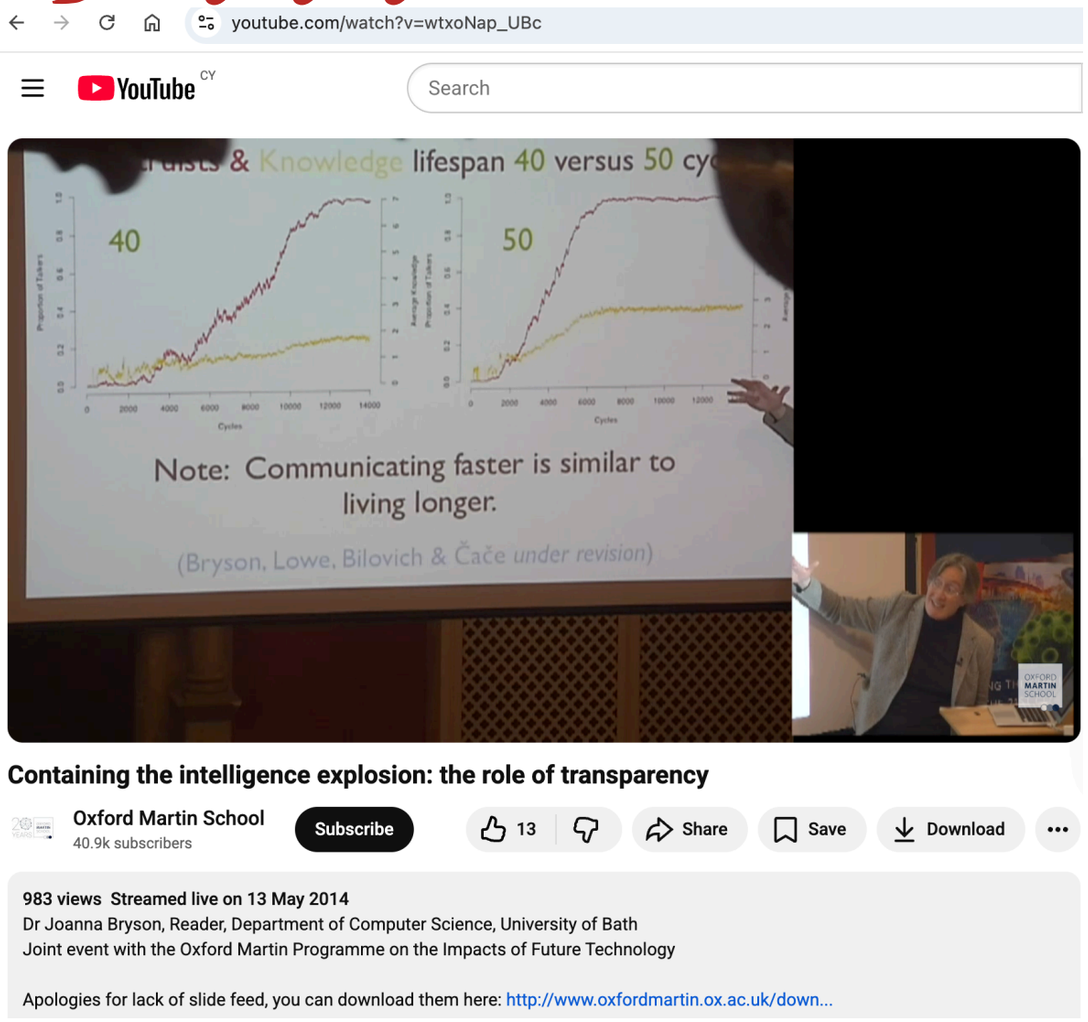
Outline

- Autonomy Through the Ages
- AI and Human Autonomy in Theory
- Autonomy in Practice

About AGI... and ABM

tl;dw cooperation is absolutely as natural as competion, slightly more ubiquitous in Nature

Čače & Bryson 2008; Stewart & al 2020.



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

Vernor Vinge
Department of Mathematical Sciences
San Diego State University

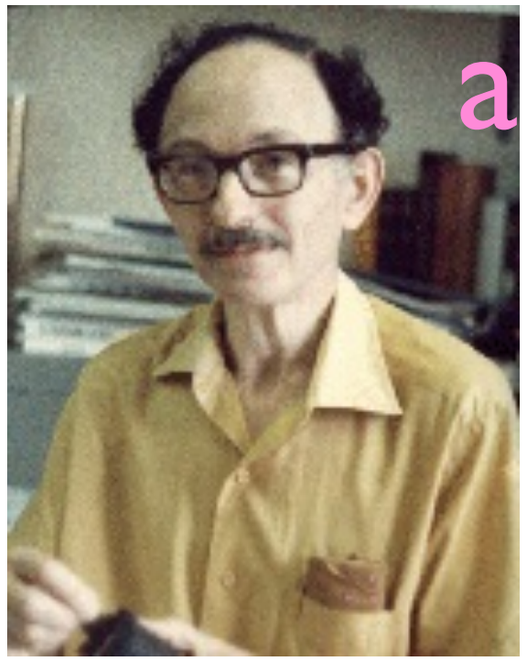
This article was for the VISION-21 Symposium sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute, March 30-31, 1993. It is also retrievable from the NASA technical reports server as part of NASA CP-10129. A slightly changed version appeared in the Winter 1993 issue of _Whole Earth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



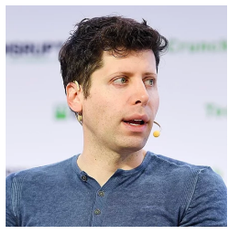
I J Good (1965)

Nick Bostrom (2014)

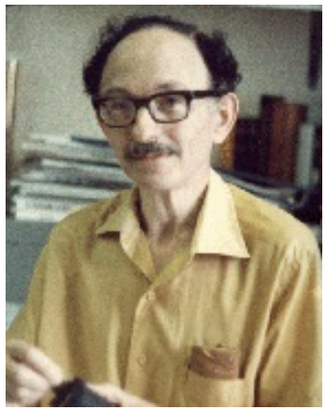
Self improving (machine) intelligence – learning to learn – leads to a singularity.

Exponential on exponential growth.

Unintended consequences derived in pursuit of designed priorities.



Superintelligence



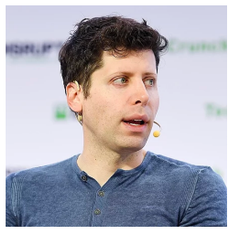
I J Good

Nick Bostrom (2014)

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

Exponential on exponential growth.

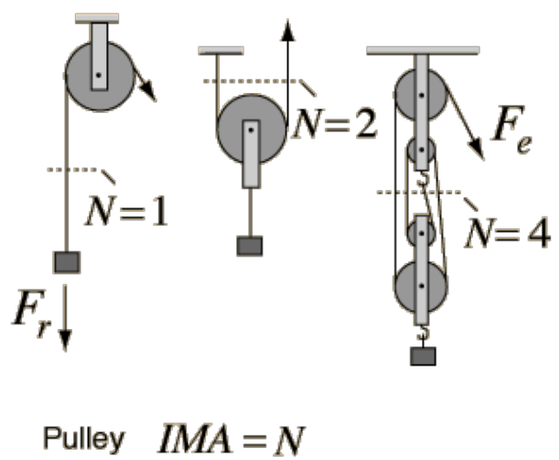
Unintended consequences derived in pursuit of designed priorities.



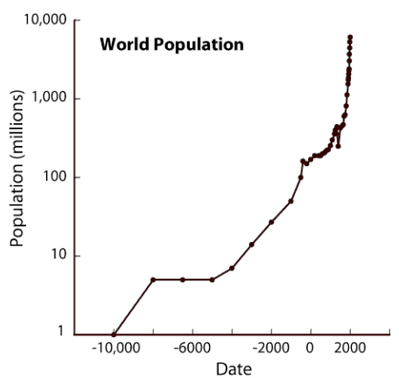
12,000 years of AI

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!



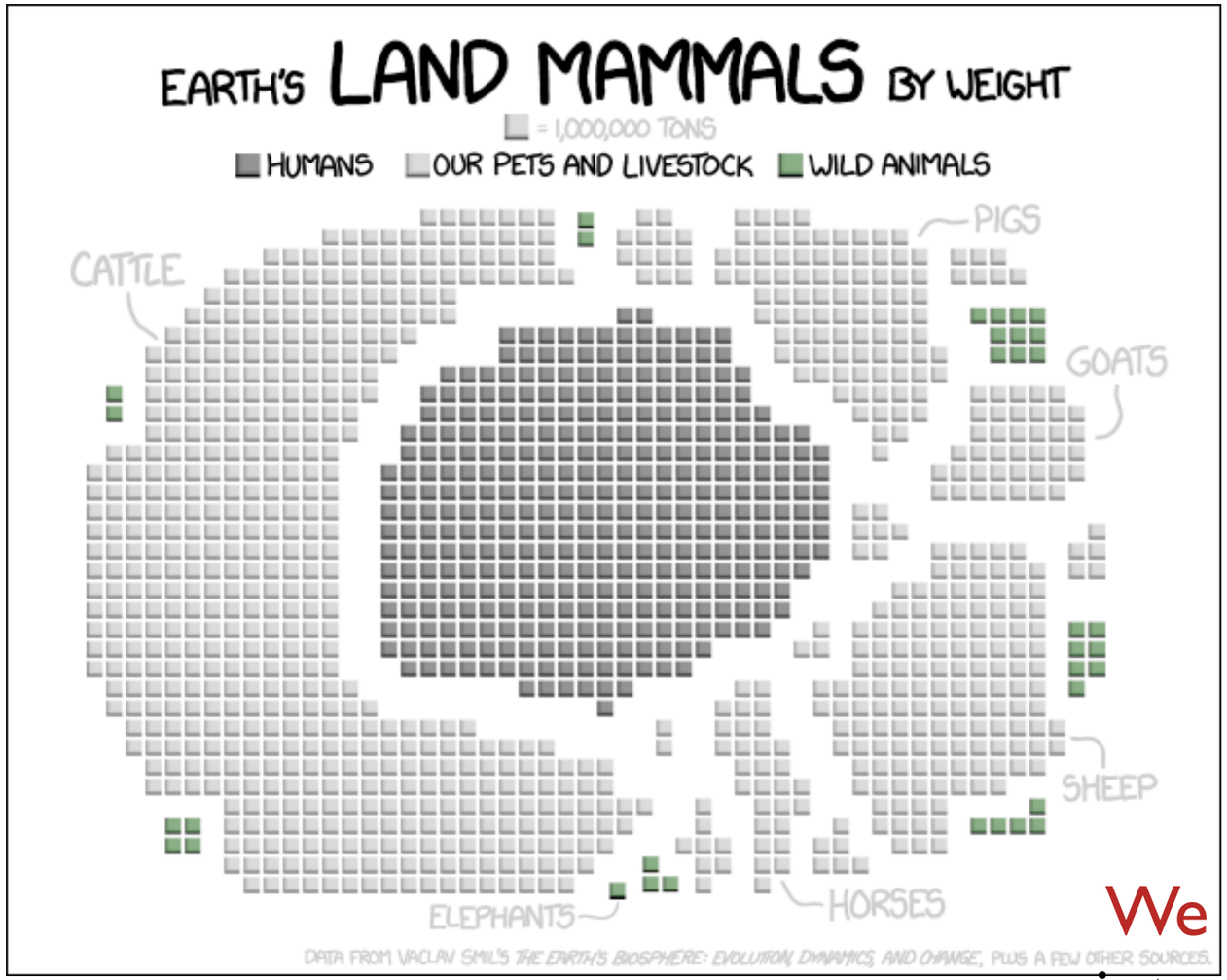
Pulley $IMA = N$

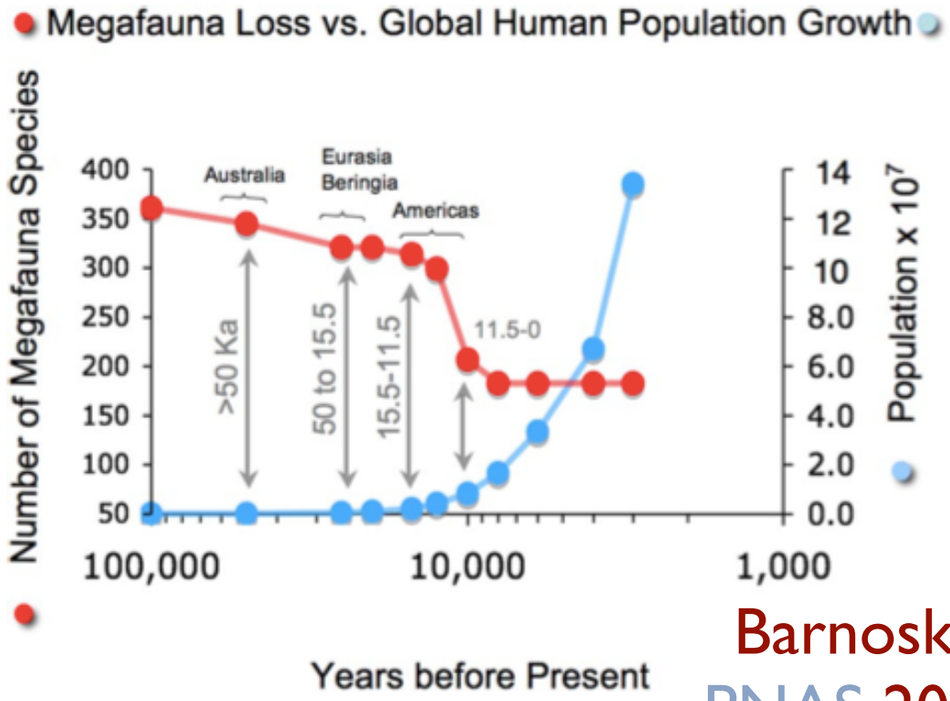


Agency
2015

Unanticipated Consequences

xkcd

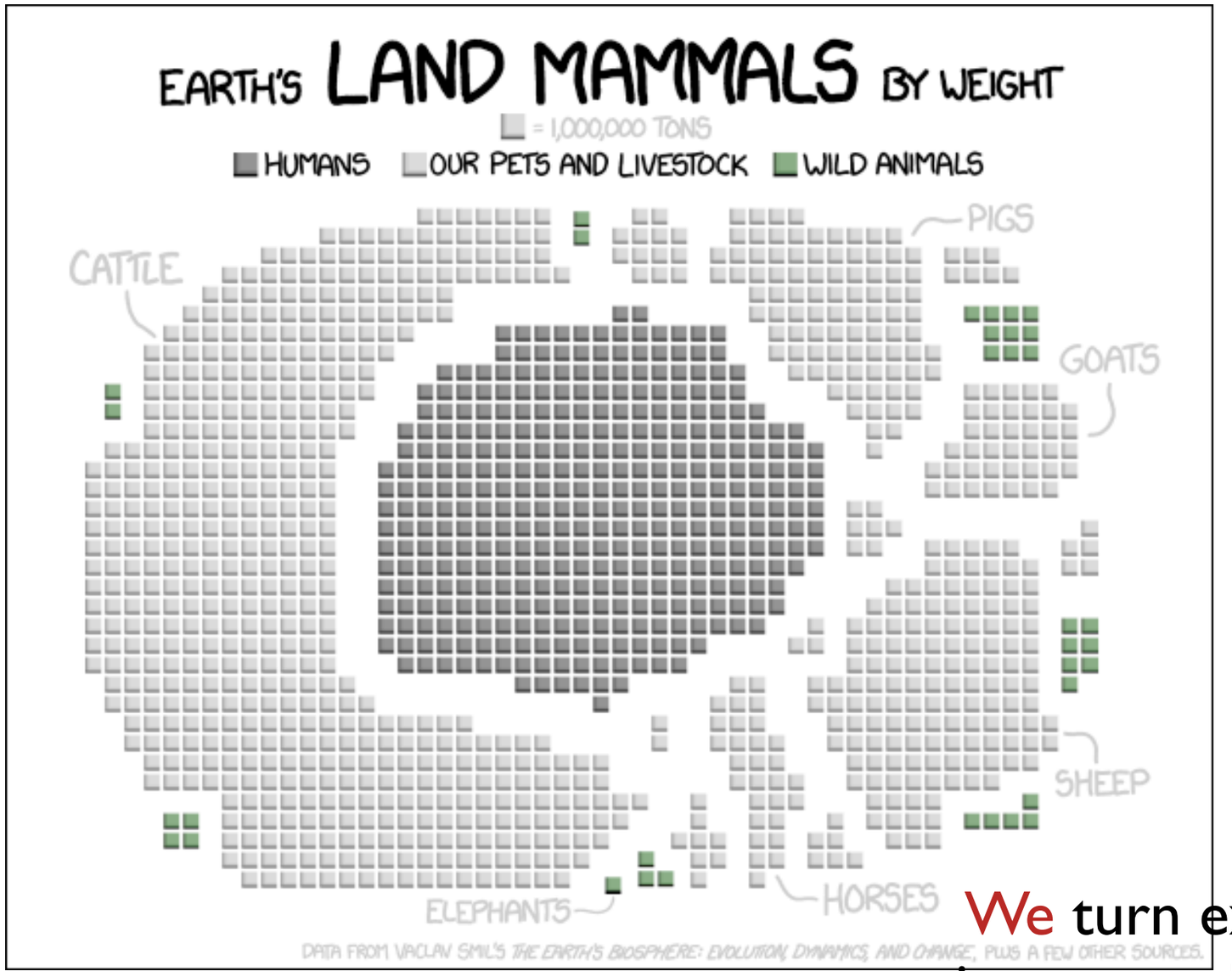




Barnosky, PNAS 2008

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

xkcd



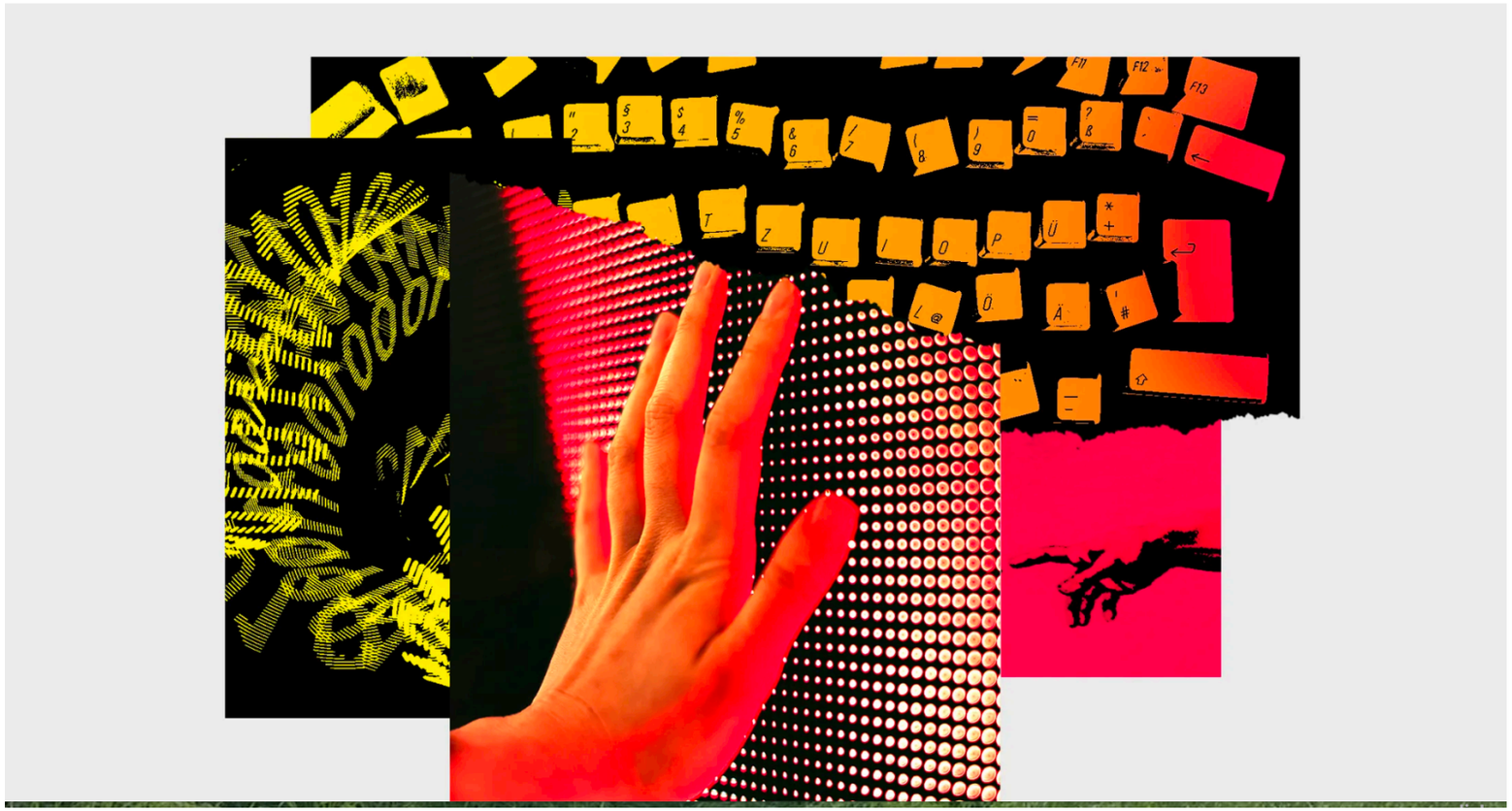
We're the Paperclips

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

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 patiency 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

Parrots are countable, discrete, and intrinsically concerned with social status

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

- Maintain their own (large!) societies.
- Participate morally in human-rooted households.
- Will never participate in human law.

(AI) Ethics Definitions for negotiating 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 (as does trust.)
- Law is a rapid, explicit method for constructing ethics.

} 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 (as does 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.

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

Robots Are Not Peers

Follow the money; dissuade those guys

Definitions from UK terrorism literature in 2000s (cf. Butler & Bryson 2007. Law: Bryson, Diamantis & Grant 2017.)

	unique and vulnerable	not unique or not vulnerable
can know and execute law	• Most humans • many corporations Rights (and obligations)	• intelligent law systems • shell companies
cannot know or cannot execute law	• non-cognizant humans • animals • badly designed AI (e.g. no autosave) Welfare	• Most artefacts • other stuff We are obliged to build AI we are not obliged to (Bryson 2010)

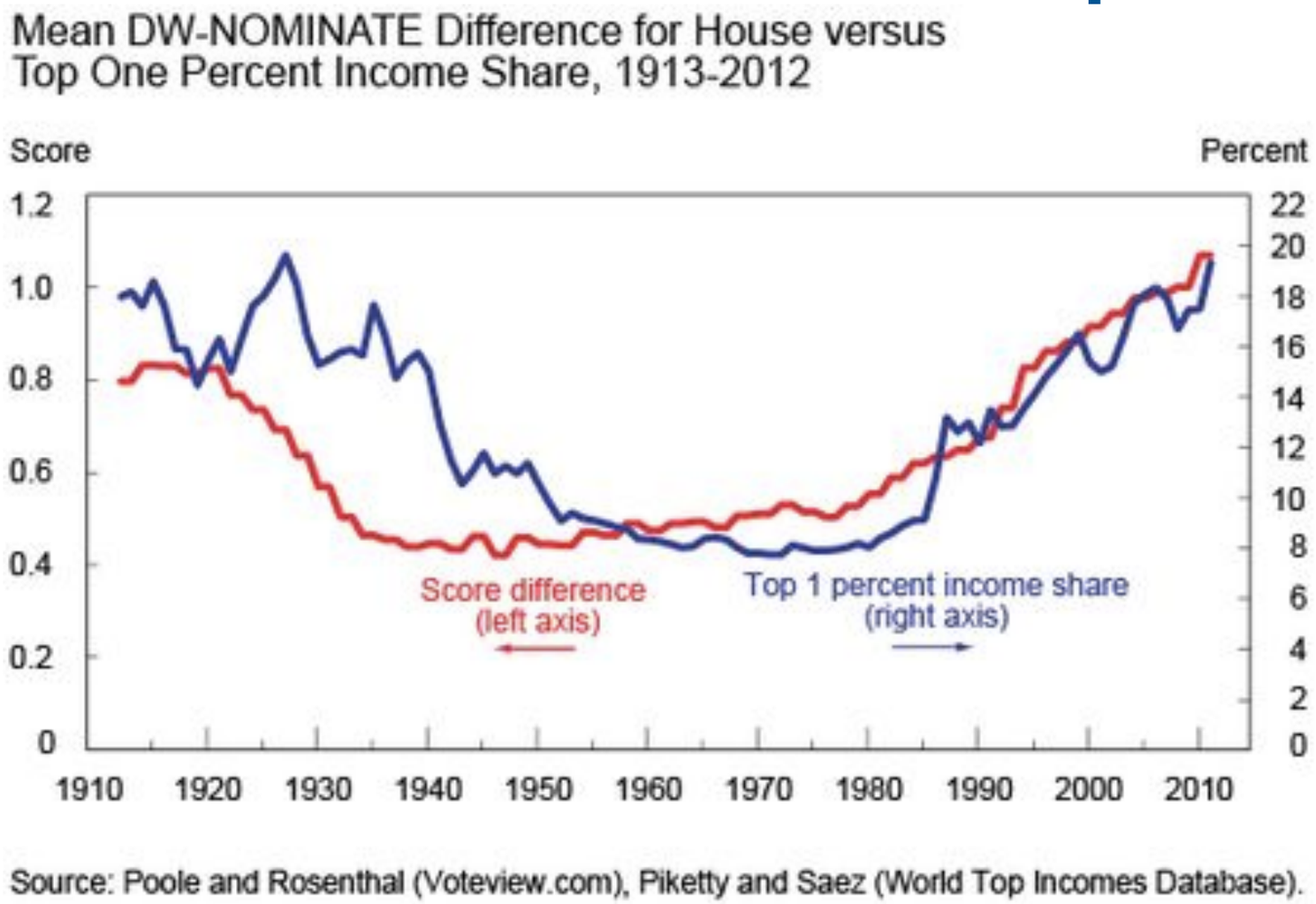
Outline

- Autonomy Through the Ages
- AI and Human Autonomy in Theory
- Autonomy in Practice

About Equity…

We've Been Here Before

Polarization and Inequality like it's 1899



In most countries.

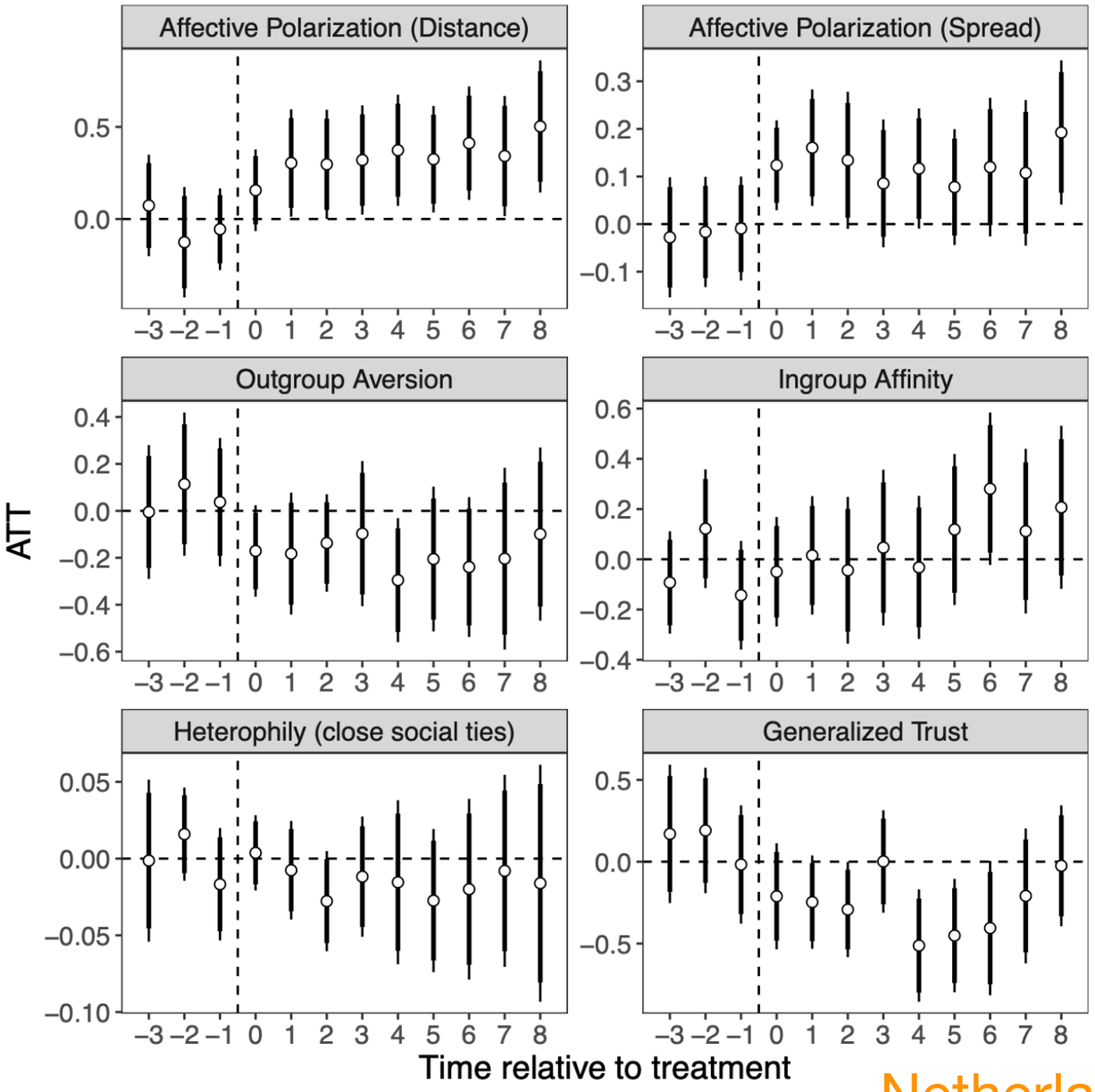
Not Germany or China(?)

Correlation has been known for decades.

Recent candidate explanation:

Polarization under rising inequality and economic decline.

Stewart, McCarty, & Bryson
Science Advances 2020



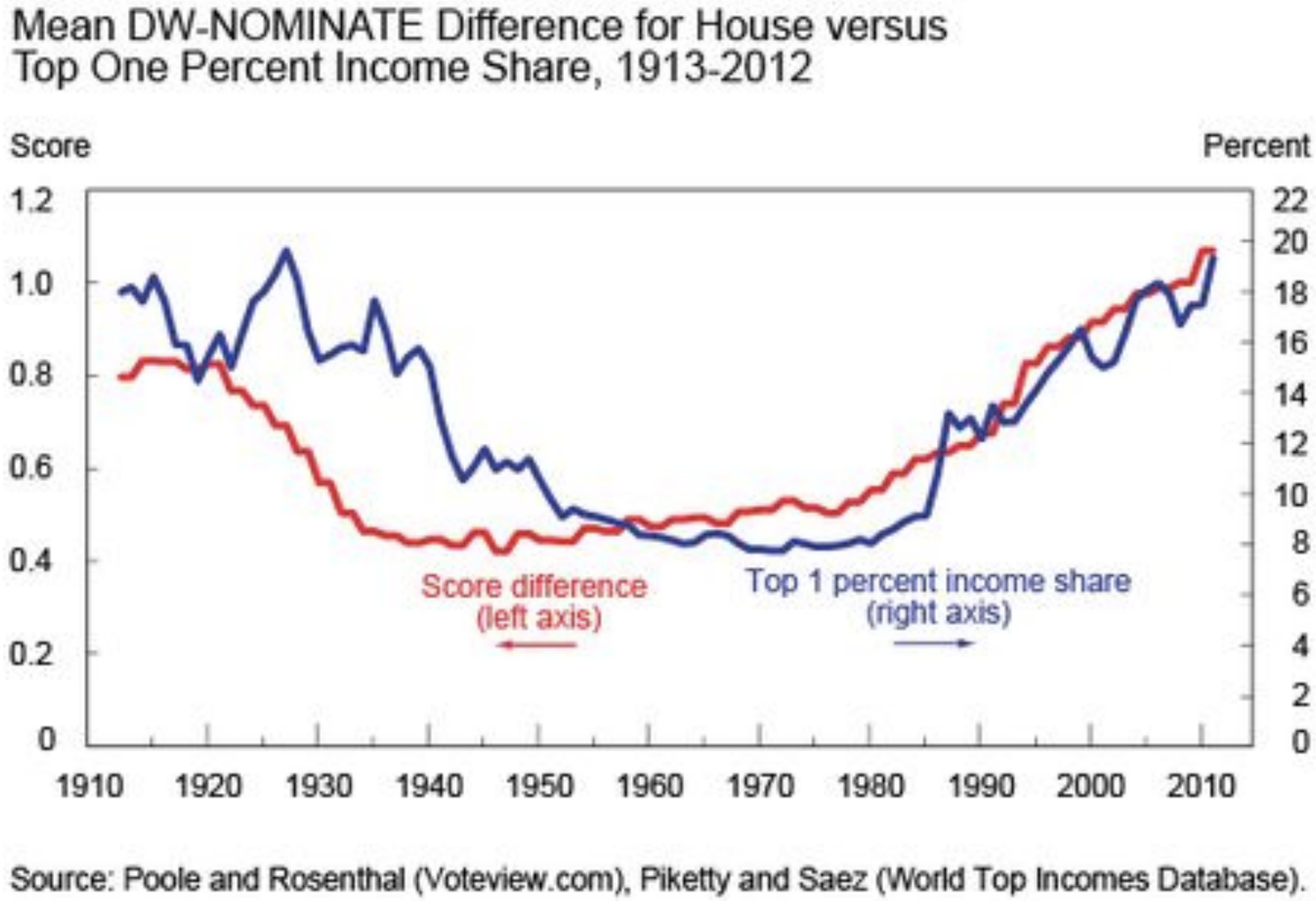
Affective polarization seems caused by newly introduced economic precarity.

We use two different measures of polarisation:

- mean distance between a respondent's affect towards their most favoured party relative to all other parties,
- the spread of respondents' weighted like-dislike scores.

Netherlands, JOB LOSS Unpublished results based on Dutch panel data – Vincent Heddesheimer and Bryson in prep.

We've Been Here Before… And We Governed Our Way Out Of It



In US & UK, enough elite disrupted by WWI, 1928 to join with progressives, reduce inequality.

After 1945, solution was spread globally via Bretton Woods,

until 1978 (plateauing of the USSR economy.)

EU Digital Regulation

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

Haataja & Bryson 2022a,b;2023

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, enforcement.) aka competition law
- AI Regulation – ensure product liability enforcement (devops), mandate capacity for responsible deployment of software acting autonomously.

Haataja & Bryson 2022a,b;2023

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, enforcement.) aka competition law
- AI Regulation – ensure product liability enforcement (devops), mandate capacity for responsible deployment of software acting autonomously.

Haataja & Bryson 2022a,b;2023

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, enforcement.)
- AI Regulation – ensure product liability enforcement (devops), mandate capacity for responsible deployment of software acting autonomously.

Bryson (Weizenbaum 100!) 2023

Can We Really Trust the Government?

White Lion Tour on TripAdvisor

Legitimacy

≈ the extent to which the people perceive their role as constitutive

Take Aways

- People matter, at least to people, which people are the motivated actors creating products, laws, and changing the ecosystem.
- Our autonomy depends on our equity – our capacity to enforce accountability.
- AI is a corporate output (manifested capital). Presently it's often used to blur accountability, but that's not a necessary outcome.
- Human Centring Is a Design Specification. DevOps and Systems Engineering are obligatory for all artefacts, including AI.
- You can help establish these norms.

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

Hertie School
Centre for
Digital Governance

The Limits of Artificial Agency

Hertie School
Centre for
Digital Governance

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