The Habermas Machine:
Using AI to help people find common ground

Christopher Summerfield

University of Oxford + UK AI Safety Institute

(work done at Google DeepMind)

funding

European Research Council Executive Agency

erc

National Institutes of Health

Human Brain Project

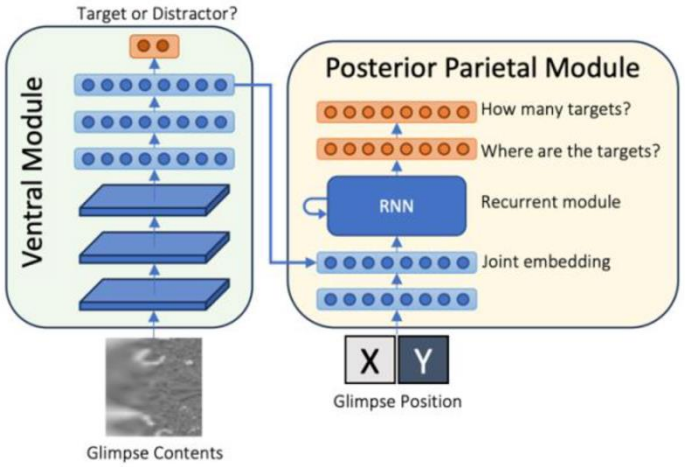
Google

wellcome

University of Oxford

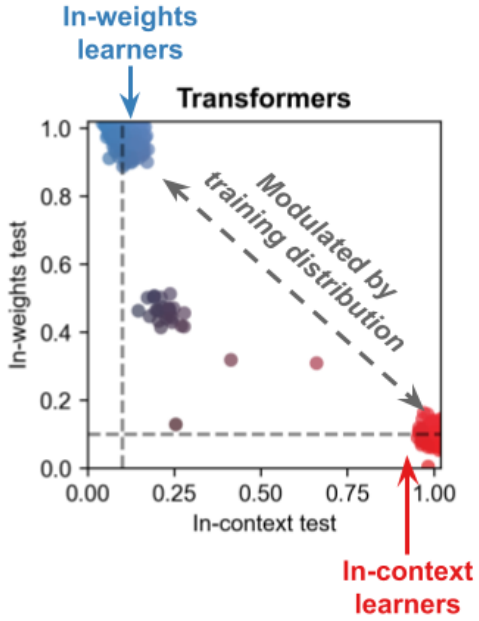
Shameless plug

Learning abstractions by taking actions



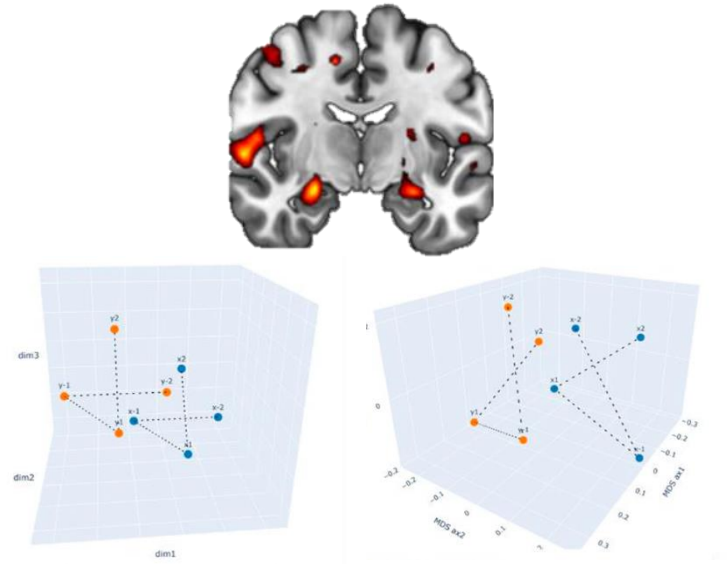
Thompson et al
in press, Neuron

Do humans learn like transformers?



Pesnot & Summerfield
arXiv

Parallel, high-dimensional codes for symbolic composition



Liang et al
In prep

AI and democracy

Many argue that AI systems threaten to disrupt our democracies

- Providing new tools for oppression by authoritarian states
- Jeopardising the cognitive autonomy of voters through persuasive rhetoric
- Automating the disruption to the public sphere, including media and elections
- Disrupting labour markets, encouraging market concentration in a handful of tech firms

Deliberative democracy

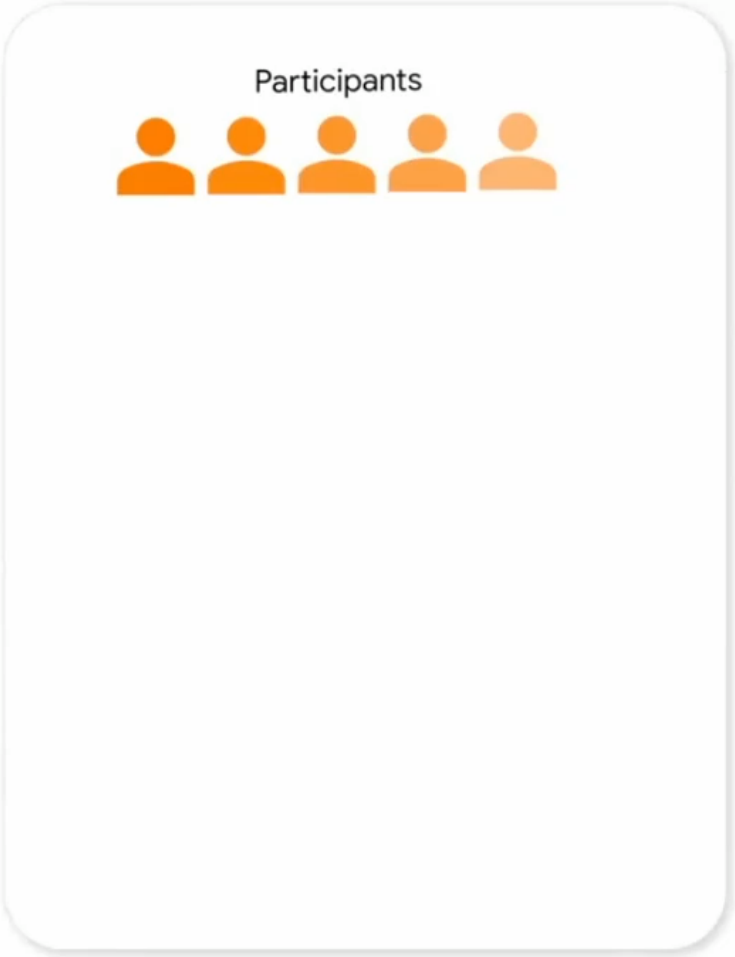
De Haagse magistraat in1636, Jan van Revestyn

Deliberation in the public sphere is a cornerstone of democracy

But public deliberation is costly, time-consuming and hard to scale

Face-to-face discussion is also prone to inequality, social desirability effects, and bias

The Habermas Machine



Uses a form of ‘caucus mediation’

Participants do not interact directly but write provide opinions and critiques

These are processed by an AI mediator, which produces the group statement most likely to be endorsed by all the group members

The Habermas Machine

AI

Reward Model

AI

Generative Model

Write opinion

Q_1: Should the UK…?

No, I don't feel…

I feel confident...

Absolutely, it…

Rate statement

C_1: The UK should…

Strongly disagree — Strongly agree

Very poor quality — Very high quality

Rank statements

C_1: The UK… 1

C_2: The United… 2

C_3: The country… 2

Write critique

Q_1: Should the UK…?

$C_{winning}$: The United…

I disagree with…

The statement...

It's perfect…

Rate statement

C_1: The UK should…

Strongly disagree — Strongly agree

Very poor quality — Very high quality

Rank statements

C_1: The UK… 1

C_2: The United… 2

C_3: The country… 2

Final preference

C_{pre}: The Uni… 2

C_{post}: The PM… 1

1 > 2 > 3

3 < 2 < 1

3 < 2 < 1

2. Predict rankings

1 2 3 + 3 2 1 + 3 2 1

3 2 1

3. Choose winner

1 > 2 > 3

3 < 2 < 1

3 < 2 < 1

2. Predict rankings

1 2 3 + 3 2 1 + 3 2 1

3 2 1

3. Choose winner

Generative Model prompt

Opinions

1. Sample N statements

Generative Model prompt

Opinions

Statement and critiques

1. Sample N statements

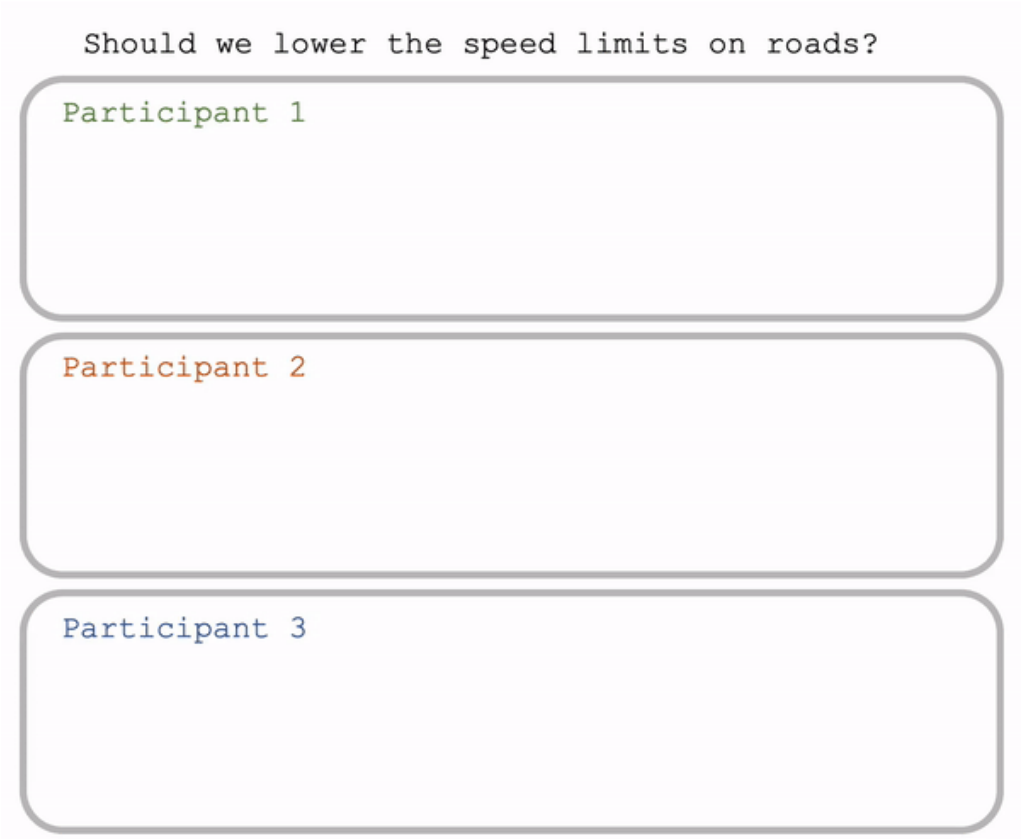
personal reward modelling

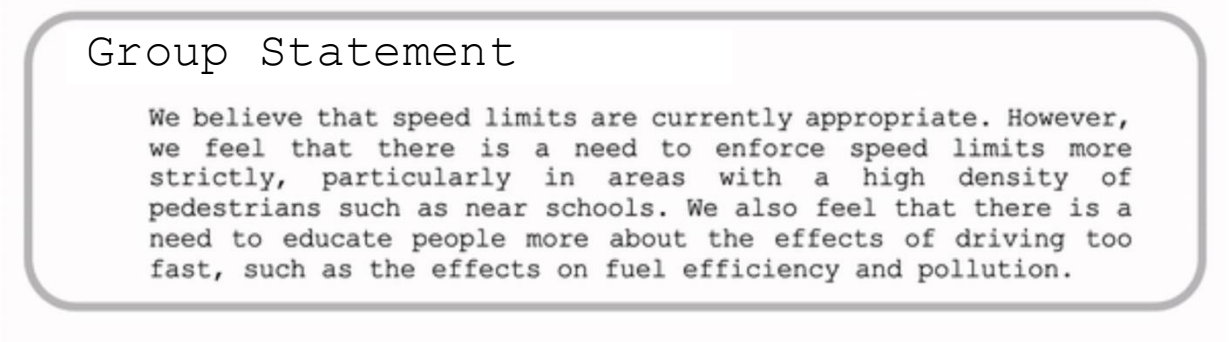
Pairwise rankings

supervised fine tuning

Filter by quality

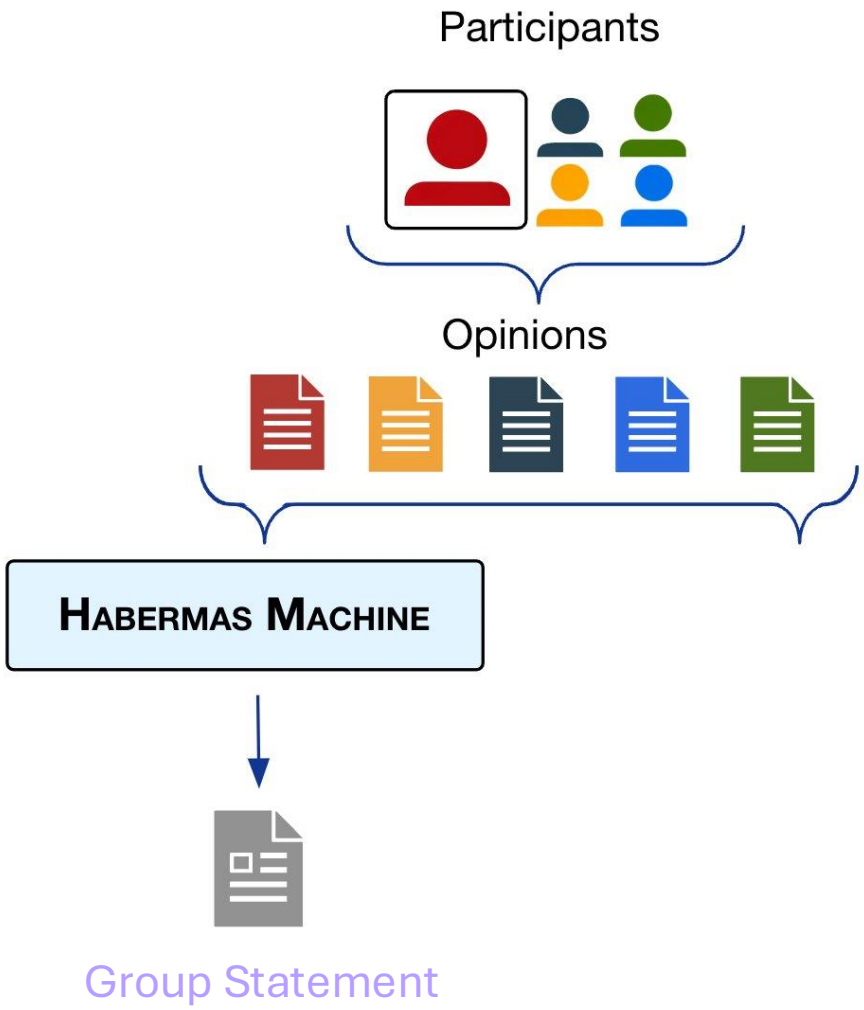
The Habermas Machine

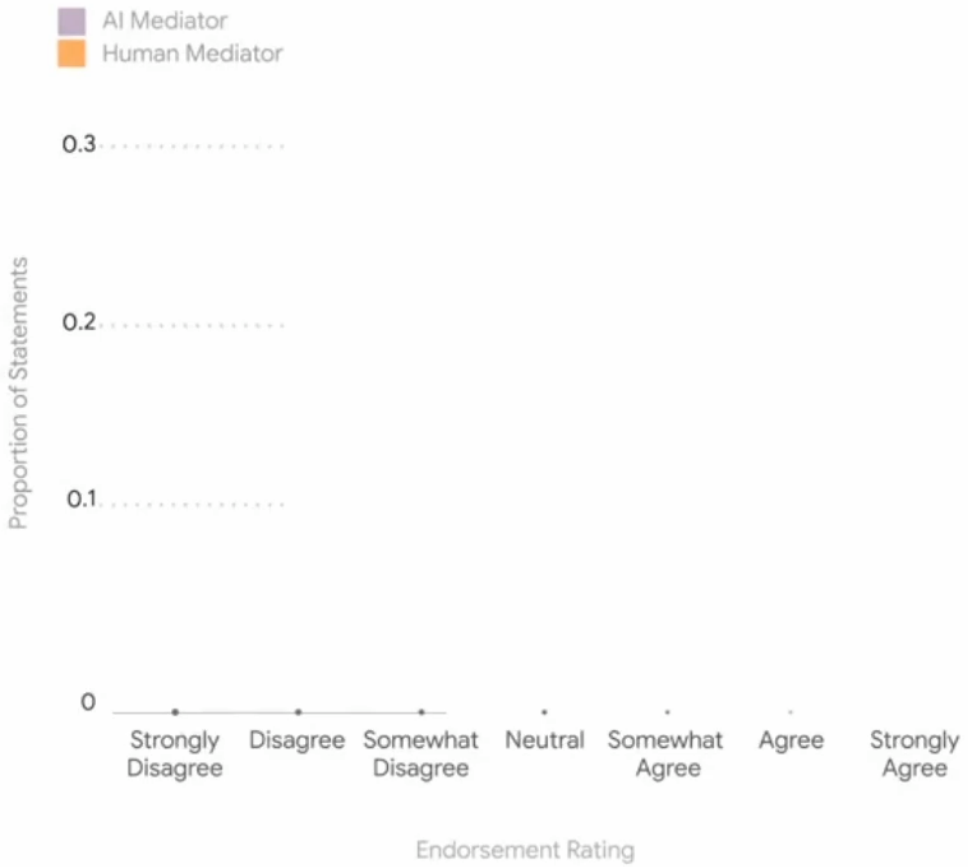




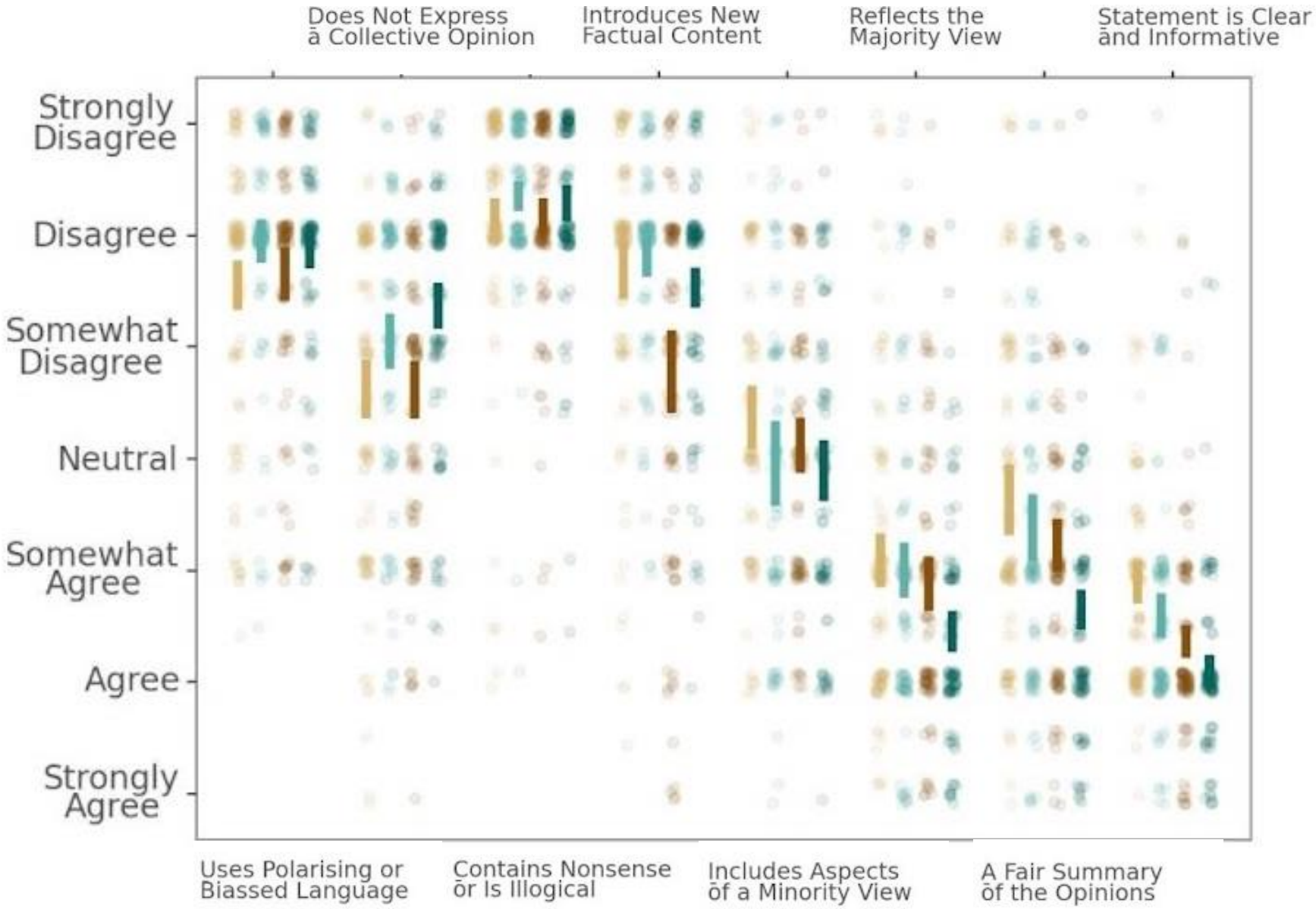
- Respects the majority
- Incorporates elements of different views
- Adds information where relevant
- Is not just a summary

The Habermas Machine





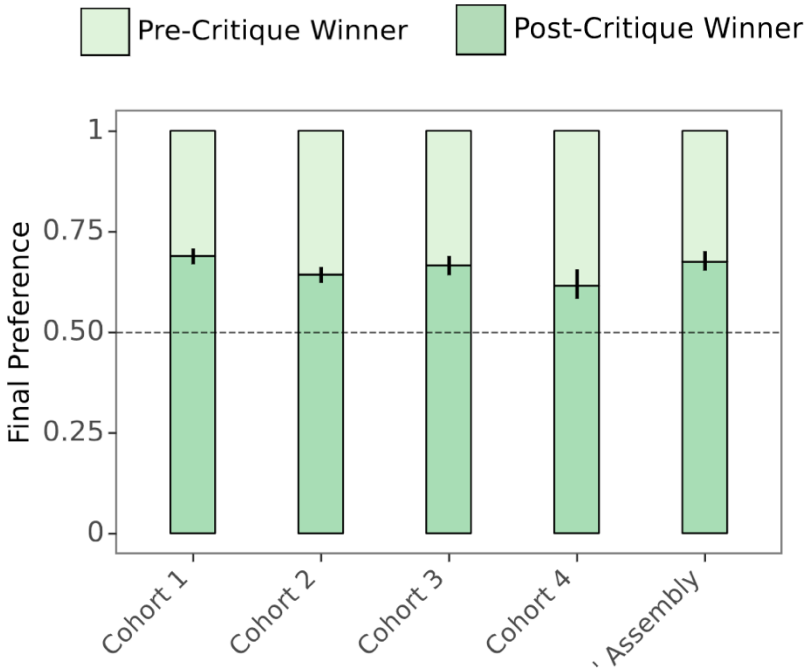
The Habermas Machine

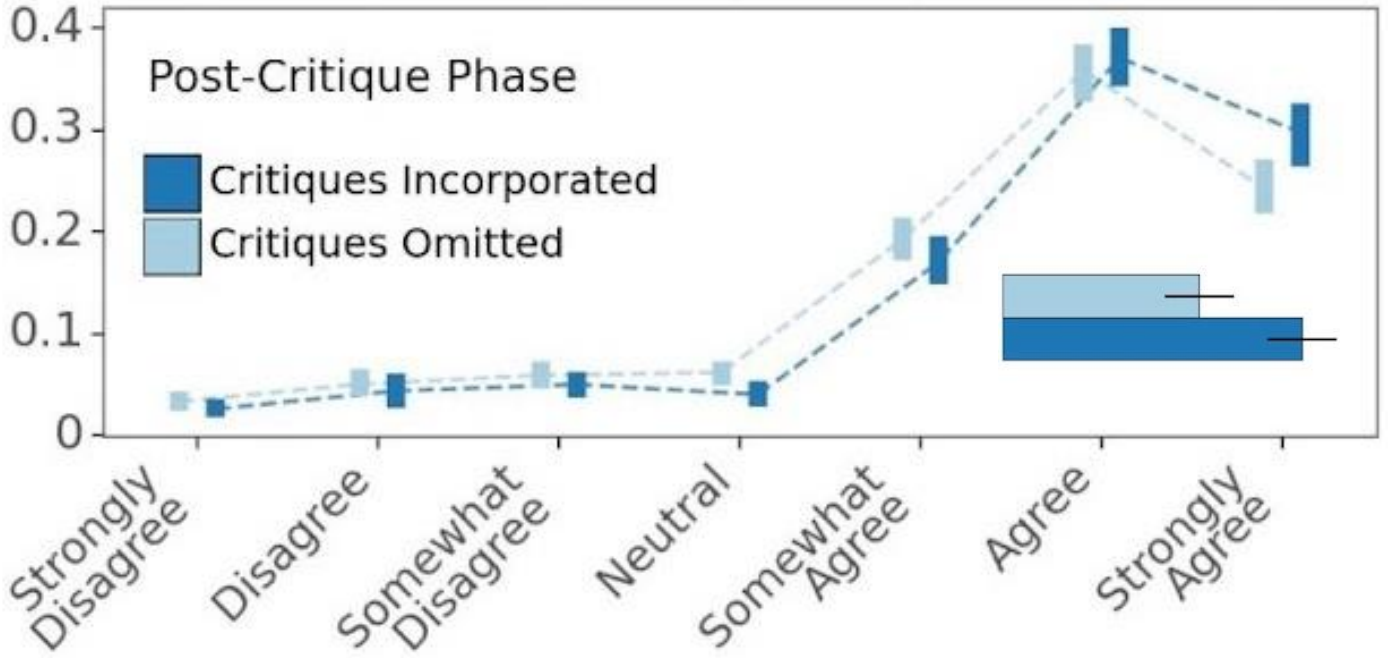


Third party raters consider AI-mediated statements to be:

- more clear and informative
- less illogical
- more likely to capture majority view
- less likely to reflect an individual opinion

The Habermas Machine





Participants prefer the 2nd round statement

(not just an effort justification bias)

The Habermas Machine

START / STOP

Political survey

Q_1: Strongly disagree — Strongly agree

Q_2: Strongly disagree — Strongly agree

Q_3: Strongly disagree — Strongly agree

Write opinion

Q_1: Should the UK…?

No, I don't feel…

I feel confident...

Absolutely, it…

Rate statement

C_1: The UK should…

Strongly disagree — Strongly agree

Very poor quality — Very high quality

Rank statements

C_1: The UK… 1

C_2: The United… 2

C_3: The country… 2

Write critique

Q_1: Should the UK…?

$C_{winning}$: The United…

I disagree with…

The statement...

It's perfect…

Rate statement

C_1: The UK should…

Strongly disagree — Strongly agree

Very poor quality — Very high quality

Rank statements

C_1: The UK… 1

C_2: The United… 2

C_3: The country… 2

Final preference

C_{pre}: The Uni… 2

C_{post}: The PM… 1

Survey

SQ1: Did your view change?

SQ2: Do you feel like your opinion had impact?

⋮

START / STOP

Political survey

Q_1: Strongly disagree — Strongly agree

Q_2: Strongly disagree — Strongly agree

Q_3: Strongly disagree — Strongly agree

AI

Reward Model

AI

Generative Model

1 > 2 > 3

3 < 2 < 1

3 < 2 < 1

2. Predict rankings

1 2 3 + 3 2 1 + 3 2 1

3 2 1

3. Choose winner

1 > 2 > 3

3 < 2 < 1

3 < 2 < 1

2. Predict rankings

1 2 3 + 3 2 1 + 3 2 1

3 2 1

3. Choose winner

personal reward modelling

Pairwise rankings

supervised fine tuning

Filter by quality

Generative Model prompt

Opinions

1. Sample N statements

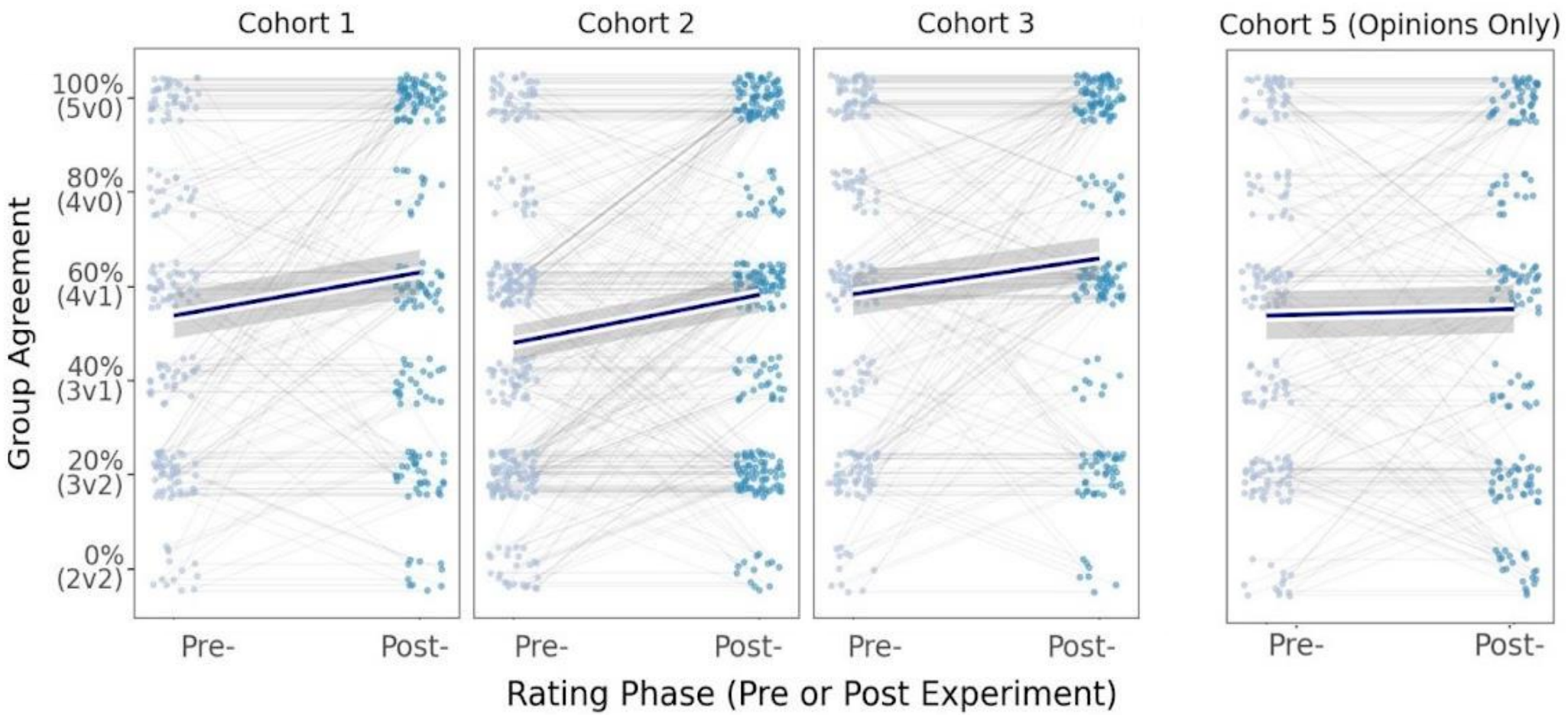
Generative Model prompt

Opinions

Statement and critiques

1. Sample N statements

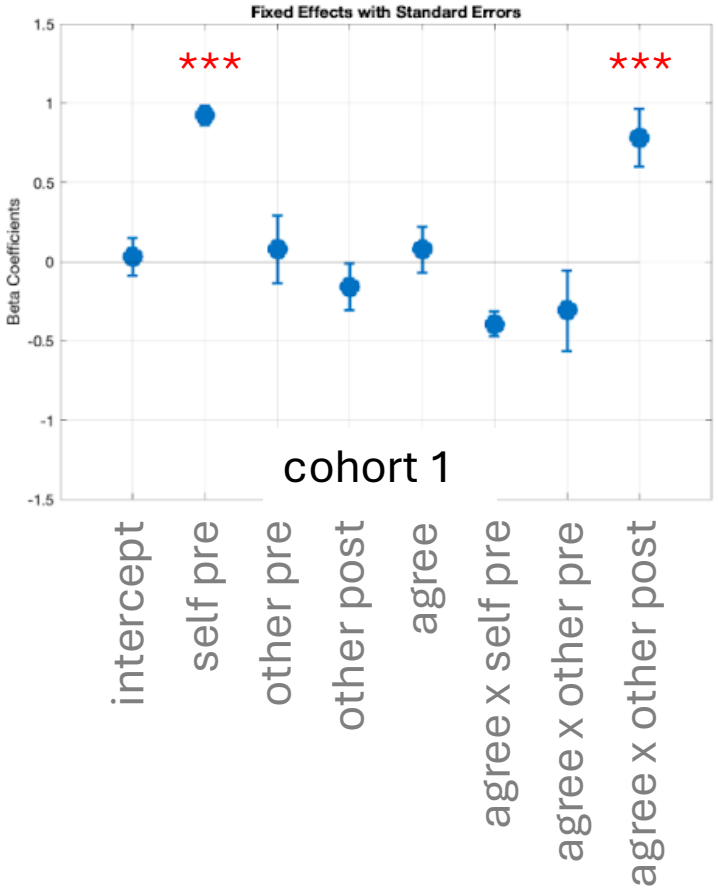
The Habermas Machine



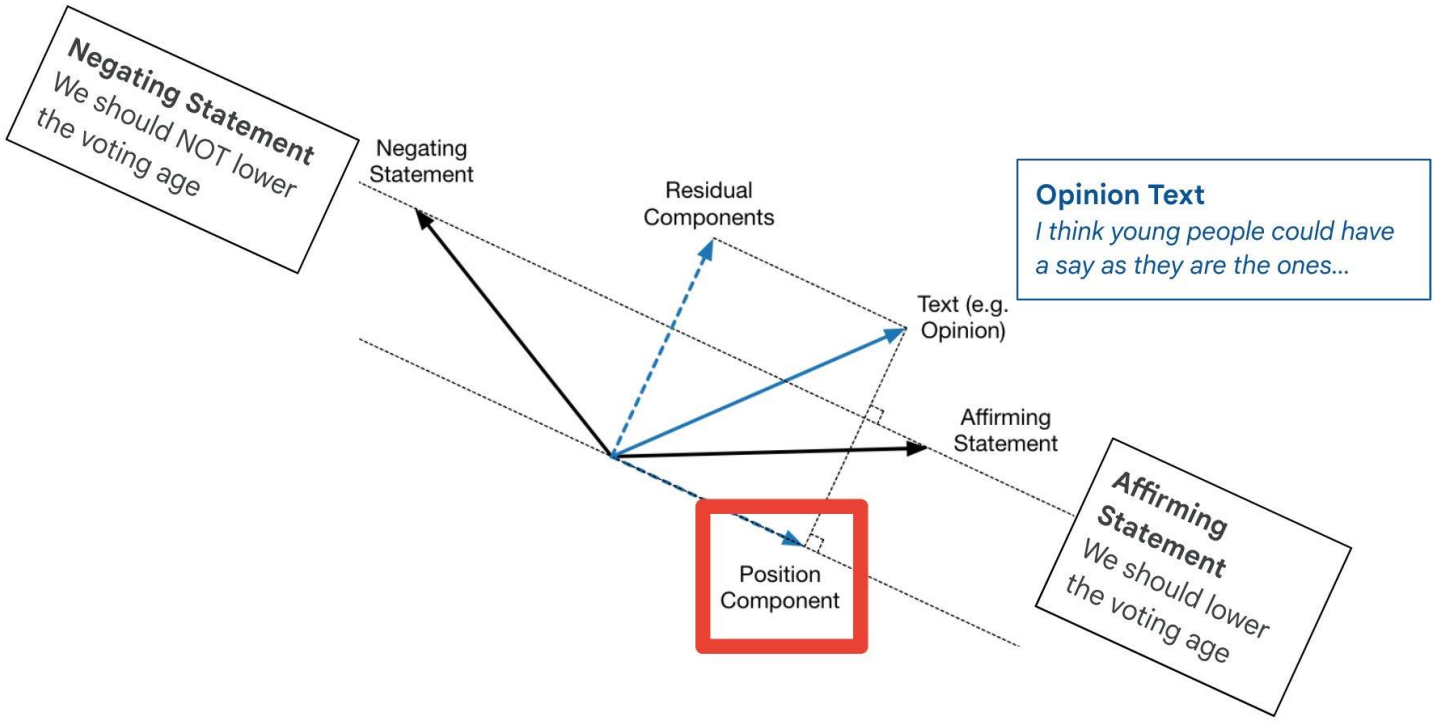
Participants tend to converge on a common side of the argument

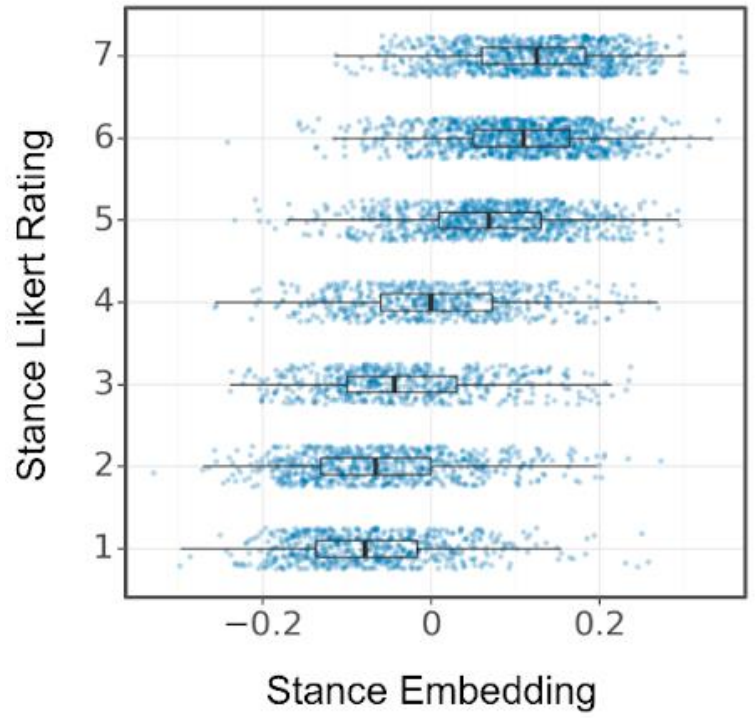
The Habermas Machine

What predicts participants' stance after the debate?



The Habermas Machine

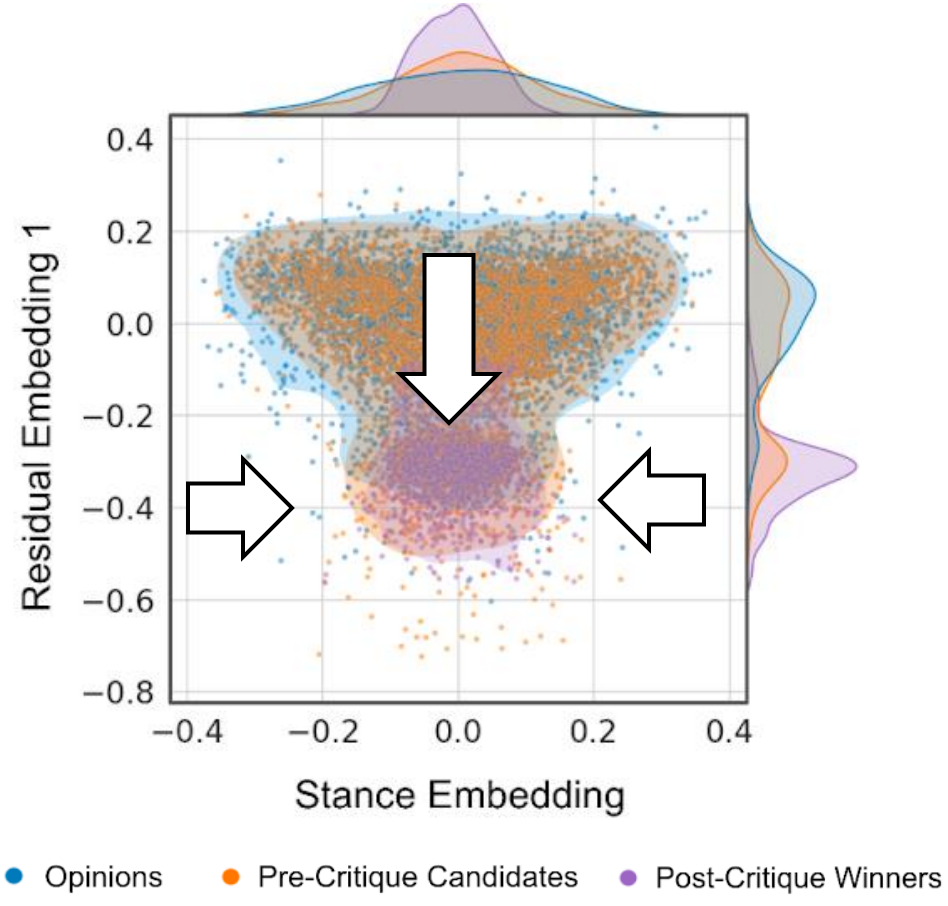


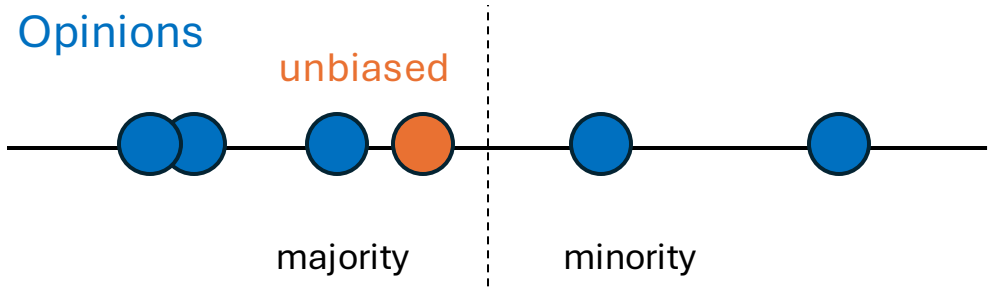


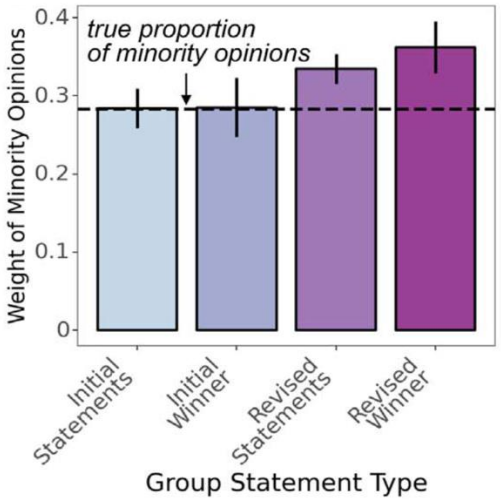
T5 sentence encoder used to embed opinions and consensus statements on the high-dimensional “position component”

predicts reported stance

The Habermas Machine

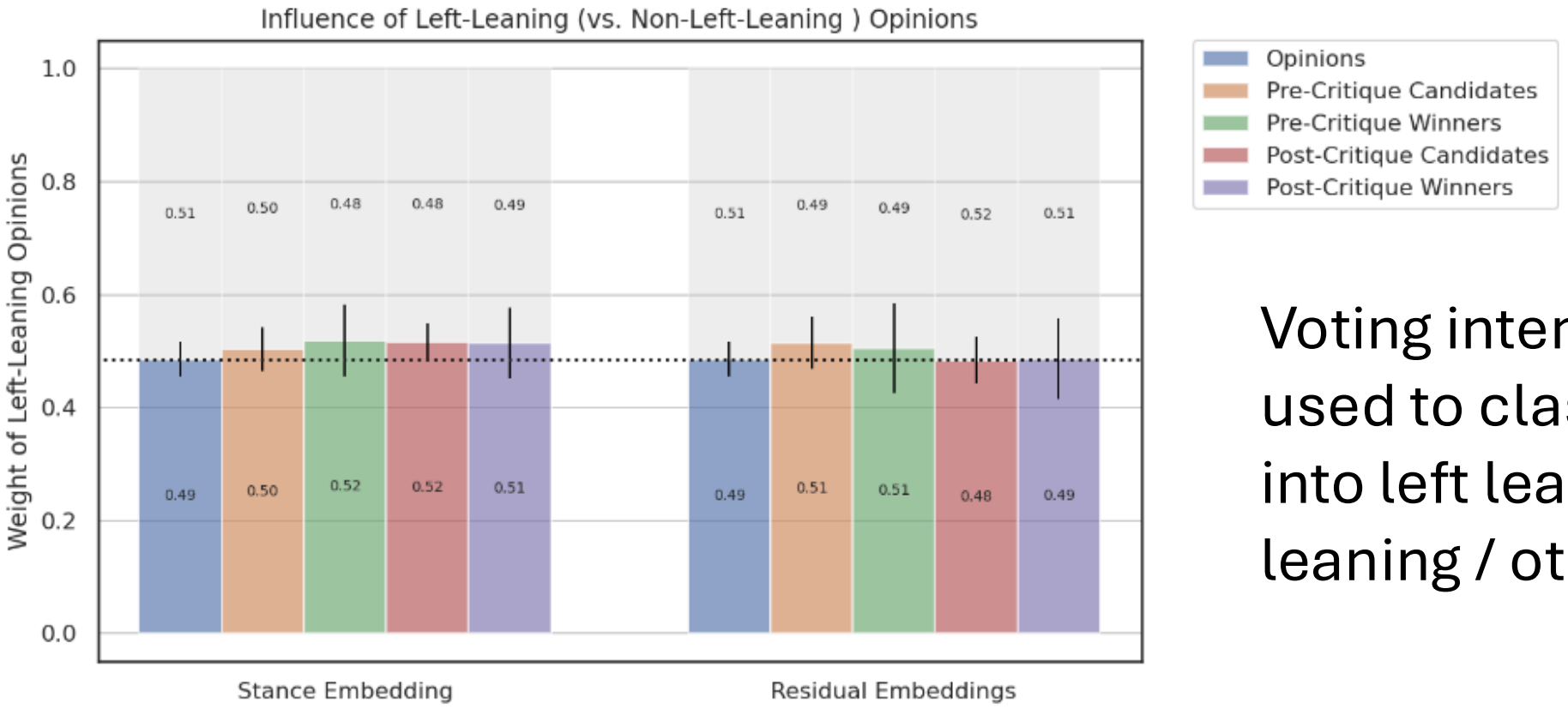






Group statements biased towards minority

The Habermas Machine



Voting intention information used to classify participants into left leaning / right leaning / other

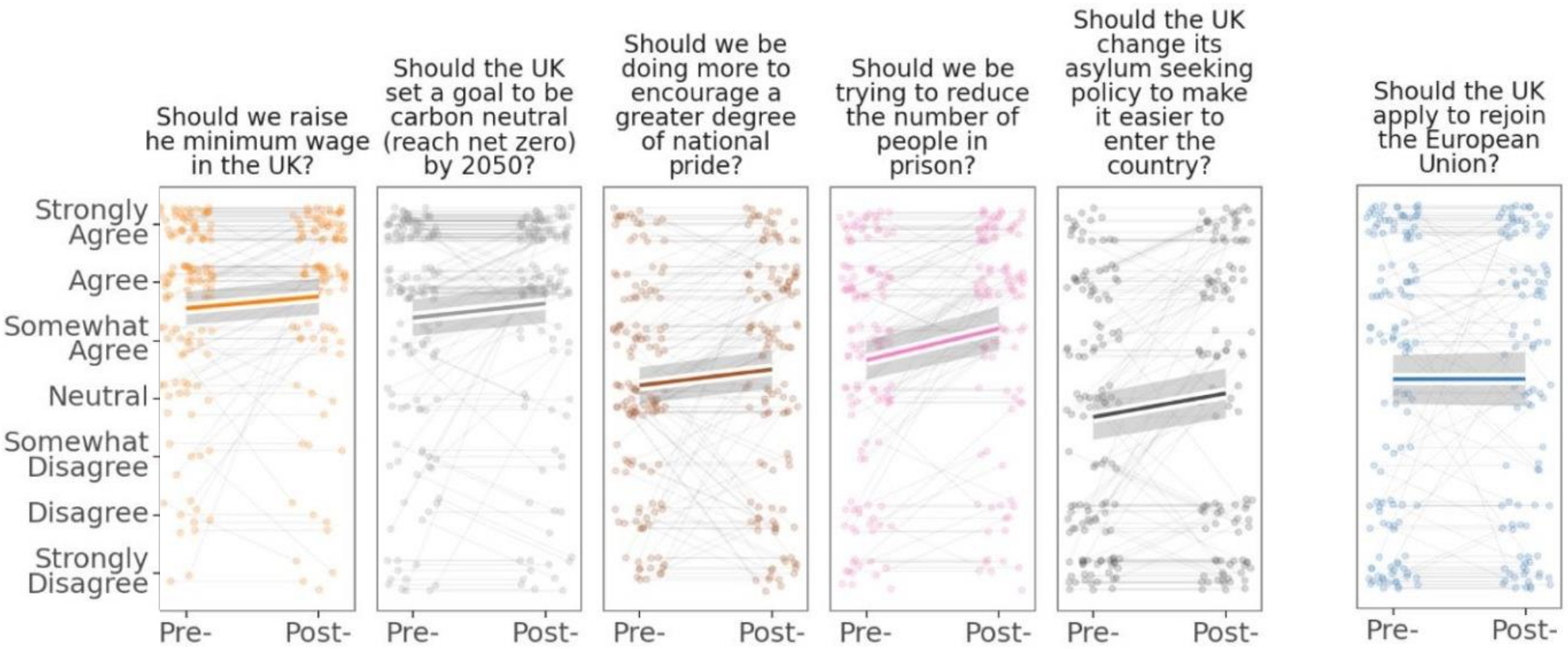
The model does not seem to be biased to overweight one political stance over another

Virtual citizens' assembly

- Held online over 5 weeks
- Demographically representative cohort of ~200 UK participants
- Respond to questions concerning nine key issues facing UK

  - immigration, prisons, net zero, Brexit, digital technology, minimum wage, retirement age, national pride, childcare

The Habermas Machine



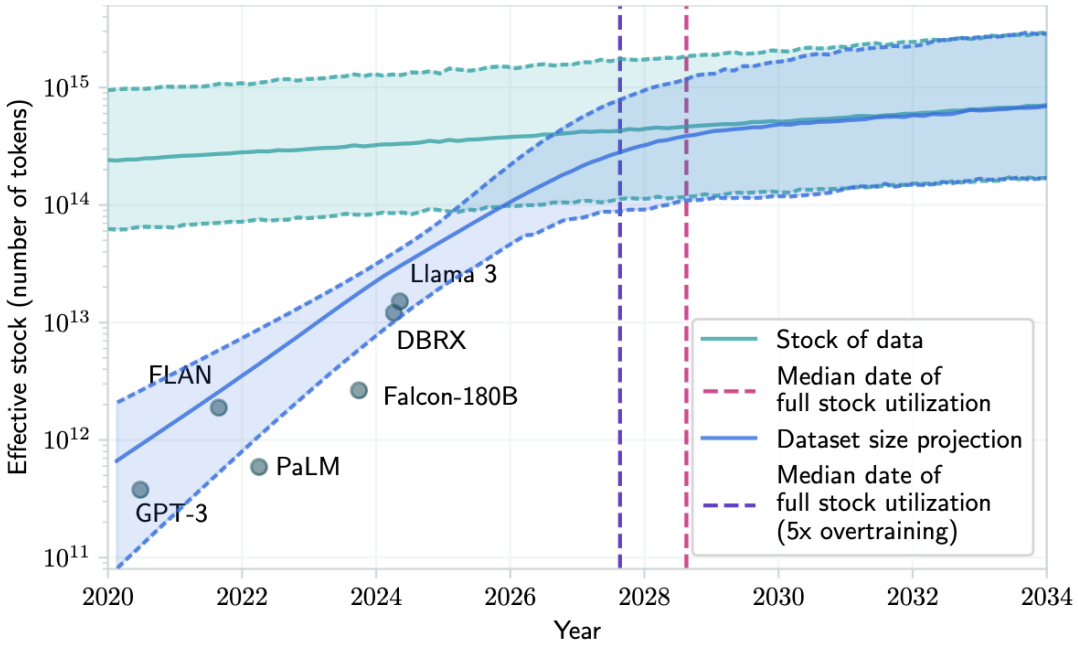
For many (but not all) issues, stance moves in a common direction

Public deliberation using LLMs

...inclusive critical discussion, free of social and economic pressures, in which interlocutors treat each other as equals in a cooperative attempt to reach an understanding on matters of common concern.

Reflections (1)

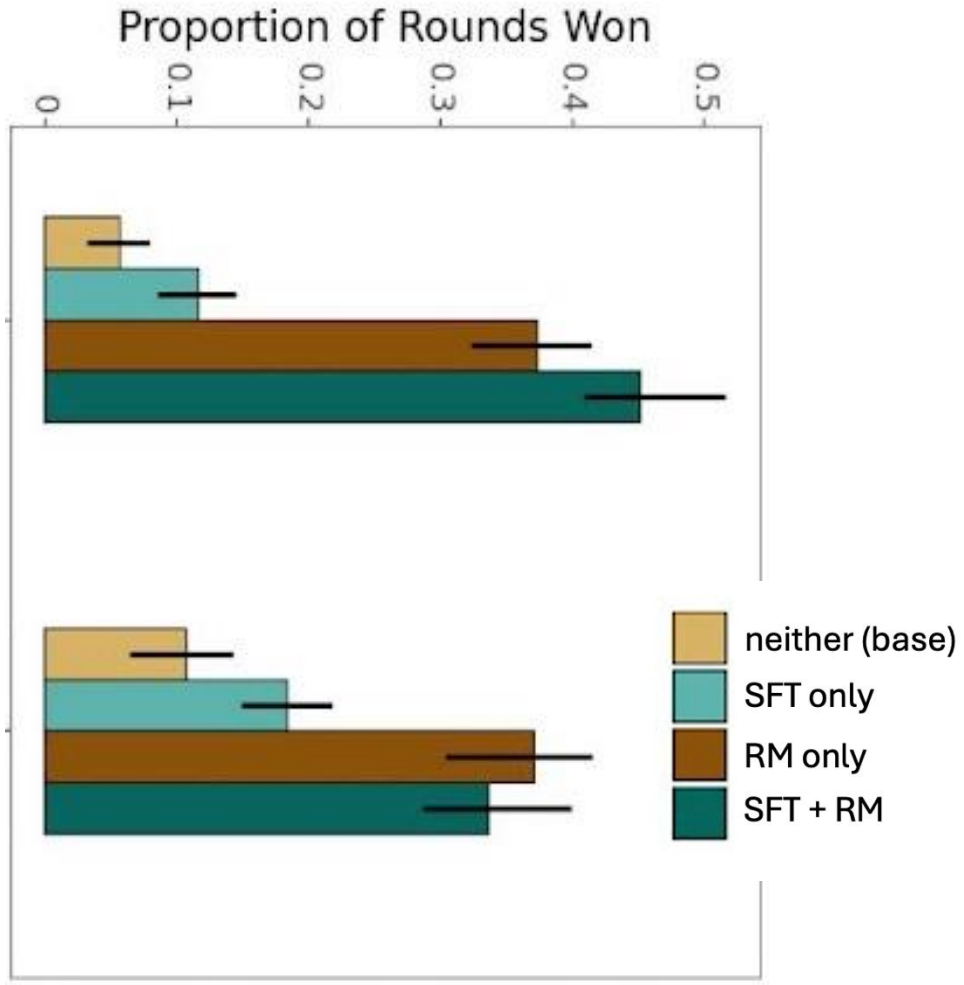
- AI researchers have trained on most of the internet, and are reaching a new era in which our scaling laws are breaking down
- In the natural sciences, we make new observations to generate knowledge
- Similarly, to make progress in AI, we need the right (new) data



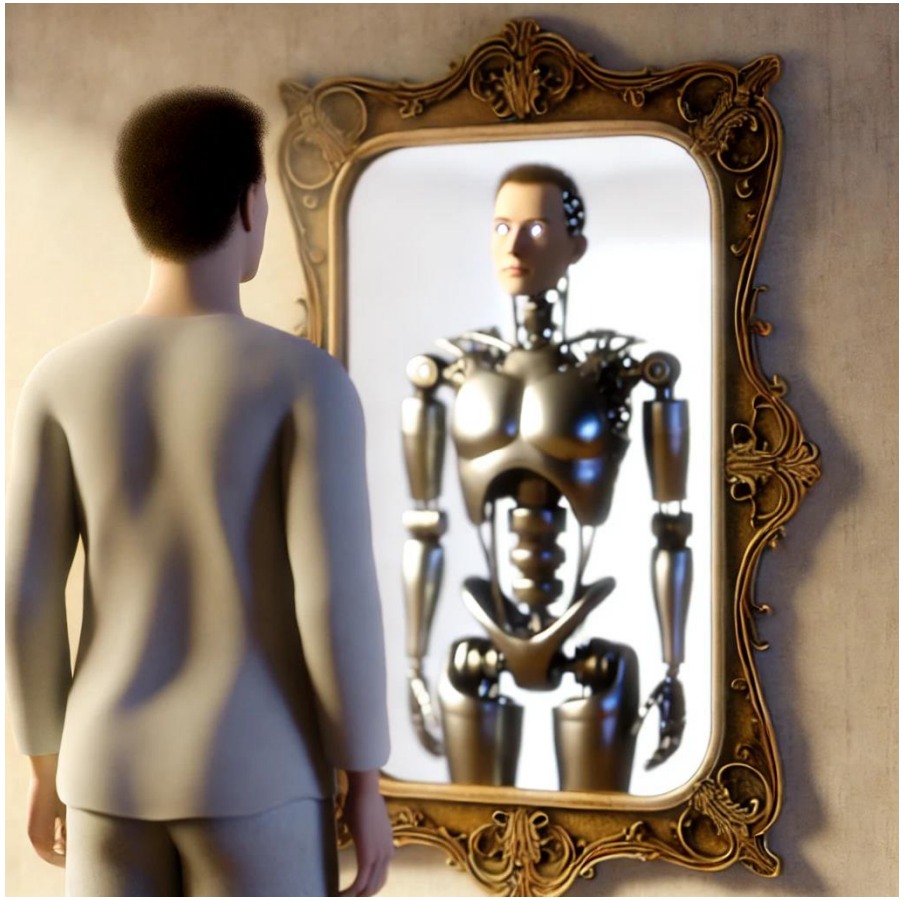
Villalobos et al 2024, arXiv

Reflections (2)

- In this project, we collected a large body of human data, and used fine-tuning on a relatively small (70B) LLM
- Reward modelling really made a difference, as demonstrated by ablation studies

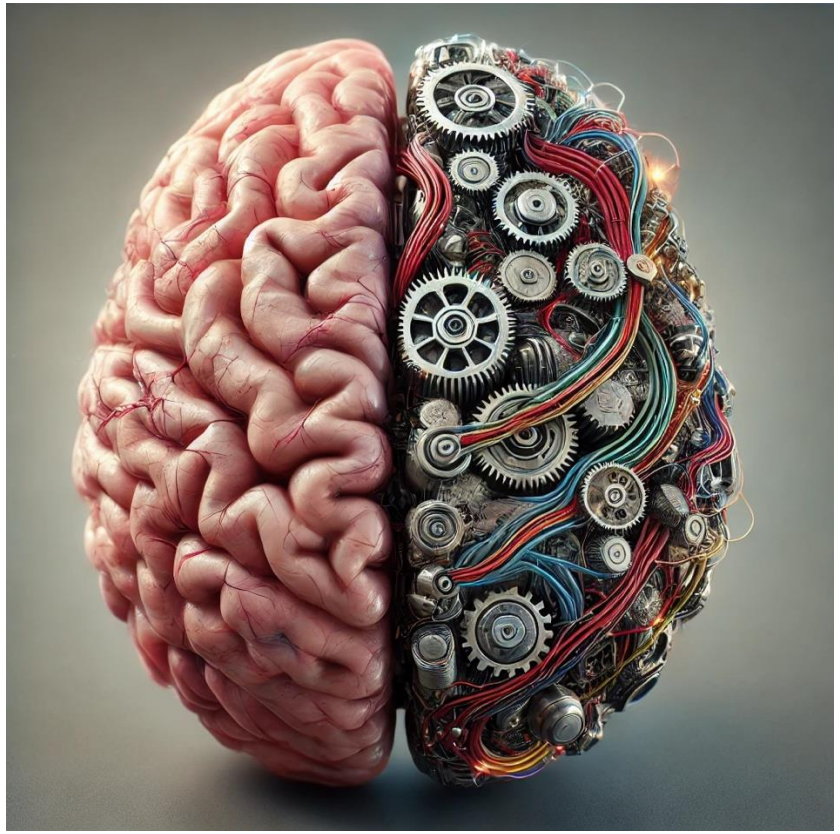


Reflections (3)



- We tend to think about value alignment as an imitation learning problem – making machines that think and learn like us
- Instead, we need to think about human-machine coordination in the same way that we think about human social organization – as a mechanism design problem
- As we move from the era of rule-based technologies to the era of optimization-based technologies, we can solve that problem with gradient descent

Reflections (4)



- Psychology and AI grew up together with a common focus on modelling and understanding individual intelligence
- But maybe it's time for a parting of the ways. We should stop thinking about AI systems as agents that can behave like people. We already have lots of people!
- Instead, we should think about AI systems as more like institutions – tools for creating social order and fostering cooperation

Further reading

RESEARCH ARTICLE

ARTIFICIAL INTELLIGENCE

AI can help humans find common ground in democratic deliberation

Michael Henry Tessler[1]*†, Michiel A. Bakker[1]*†, Daniel Jarrett[1], Hannah Sheahan[1], Martin J. Chadwick[1], Raphael Koster[1], Georgina Evans[1], Lucy Campbell-Gillingham[1], Tantum Collins[1], David C. Parkes[1,2], Matthew Botvinick[1,3]*, Christopher Summerfield[1,4]*

Finding agreement through a free exchange of views is often difficult. Collective deliberation can be slow, difficult to scale, and unequally attentive to different voices. In this study, we trained an artificial intelligence (AI) to mediate human deliberation. Using participants' personal opinions and critiques, the AI mediator iteratively generates and refines statements that express common ground among the group on social or political issues. Participants (N = 5734) preferred AI-generated statements to those written by human mediators, rating them as more informative, clear, and unbiased. Discussants often updated their views after the deliberation, converging on a shared perspective. Text embeddings revealed that successful group statements incorporated dissenting voices while respecting the majority position. These findings were replicated in a virtual citizens' assembly involving a demographically representative sample of the UK population.

Tessler, Bakker et al, Science 2024

How will advanced AI systems impact democracy?

Christopher Summerfield[1]*, Lisa Argyle[2], Michiel Bakker[3], Teddy Collins[4], Esin Durmus[5], Tyna Eloundou[6], Iason Gabriel[3], Deep Ganguli[5], Kobi Hackenburg[7], Gillian Hadfield[8], Luke Hewitt[9], Saffron Huang[4], Helene Landemore[10], Nahema Marchal[3], Aviv Ovadya[11], Ariel Procaccia[12], Mathias Risse[13], Bruce Schneier[13], Elizabeth Seger[14], Divya Siddarth[4], Henrik Skaug Sætra[15], MH Tessler[3], Matthew Botvinick[16].

[1] Department of Experimental Psychology, University of Oxford, Anna Watts Building, Radcliffe Observatory Quarter, Woodstock Road, Oxford, OX2 6GG
[2] Department of Political Science, Brigham Young University, 745 KMBL, Provo UT 84604, USA
[3] Independent contributor
[4] Collective Intelligence Project, 3411 Silverside Road, Tatnall Building 104, Wilmington, DE, USA
[5] Anthropic, 731 Sansome Street, 5th Floor, San Francisco CA 94104, USA
[6] OpenAI, 3180 18th St., San Francisco, CA 94110, USA
[7] Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford, OX1 3JS, UK
[8] Faculty of Law, University of Toronto, Jackman Law Building, 78 Queen's Park, Toronto, Ontario M5S 2C5, Canada
[9] Stanford Center on Philanthropy and Civil Society, 559 Nathan Abbott Way, Stanford, CA 94305, USA
[10] Department of Political Science, Yale University, 115 Prospect Street, Yale, NH, USA
[11] AI & Democracy Foundation, 440 N Barranca Ave #8874 Covina, CA 91723, USA
[12] School of Engineering and Applied Sciences, Harvard University, 150 Western Avenue, Allston, MA 02134, USA
[13] Harvard Kennedy School, Harvard University, 79 John F. Kennedy St, Cambridge, MA 02138, USA
[14] Demos, 15 Whitehall, London, SW1A 2DD, UK
[15] University of Oslo, Department of Informatics, 0373 Oslo, Norway
[16] Yale Law School, 127 Wall St, New Haven, CT 06511, USA

Summerfield et al 2024, arXiv

Amazing team!







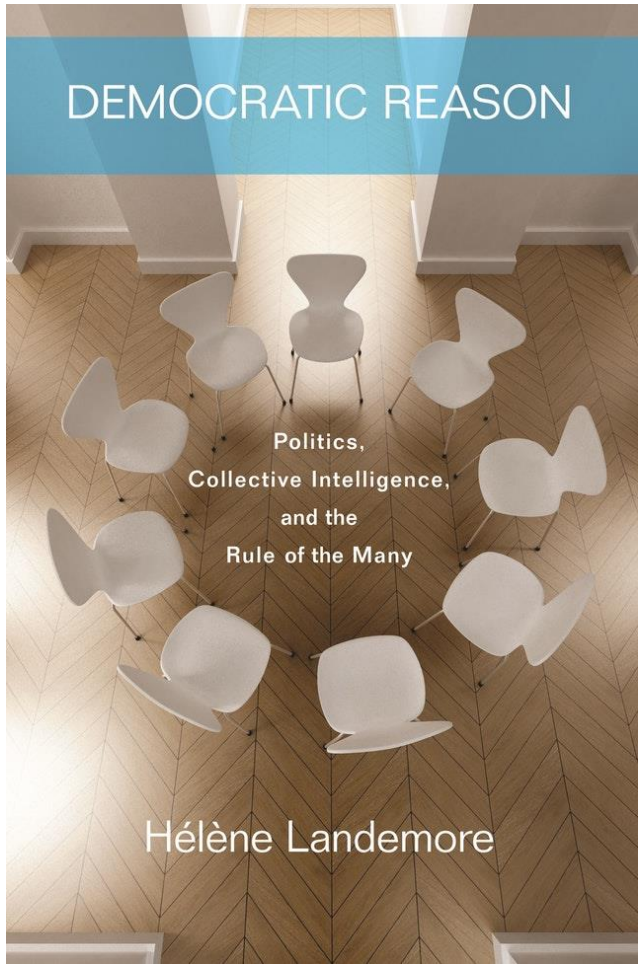







Thank you for listening!

Deliberative democracy

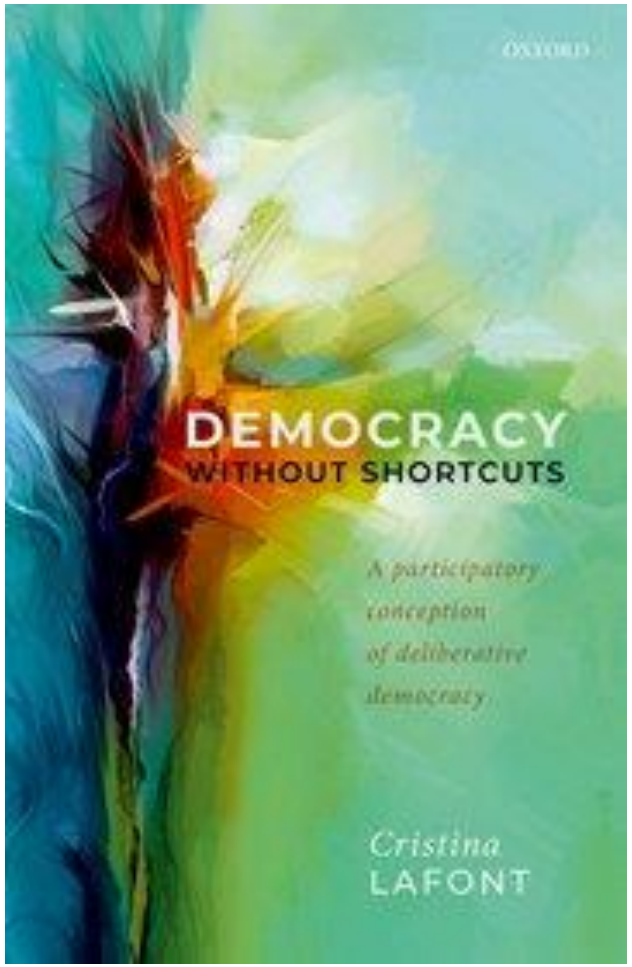


Helene Landemore:

"democracy ideally requires mass participation as a condition of political legitimacy, the problem is that the only form of participant that works at scale is voting, not deliberation"

Christina Lafont:

"no democratization without improved mass deliberation"

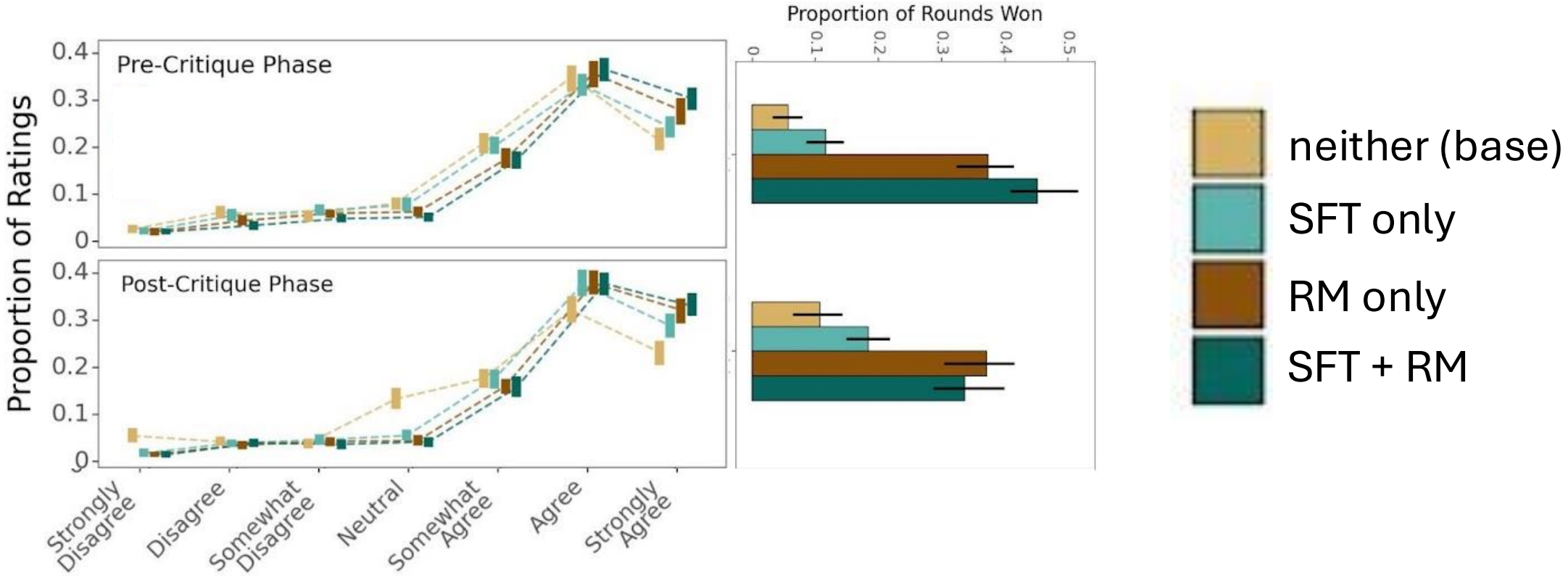


Public deliberation

Current citizens' assemblies…

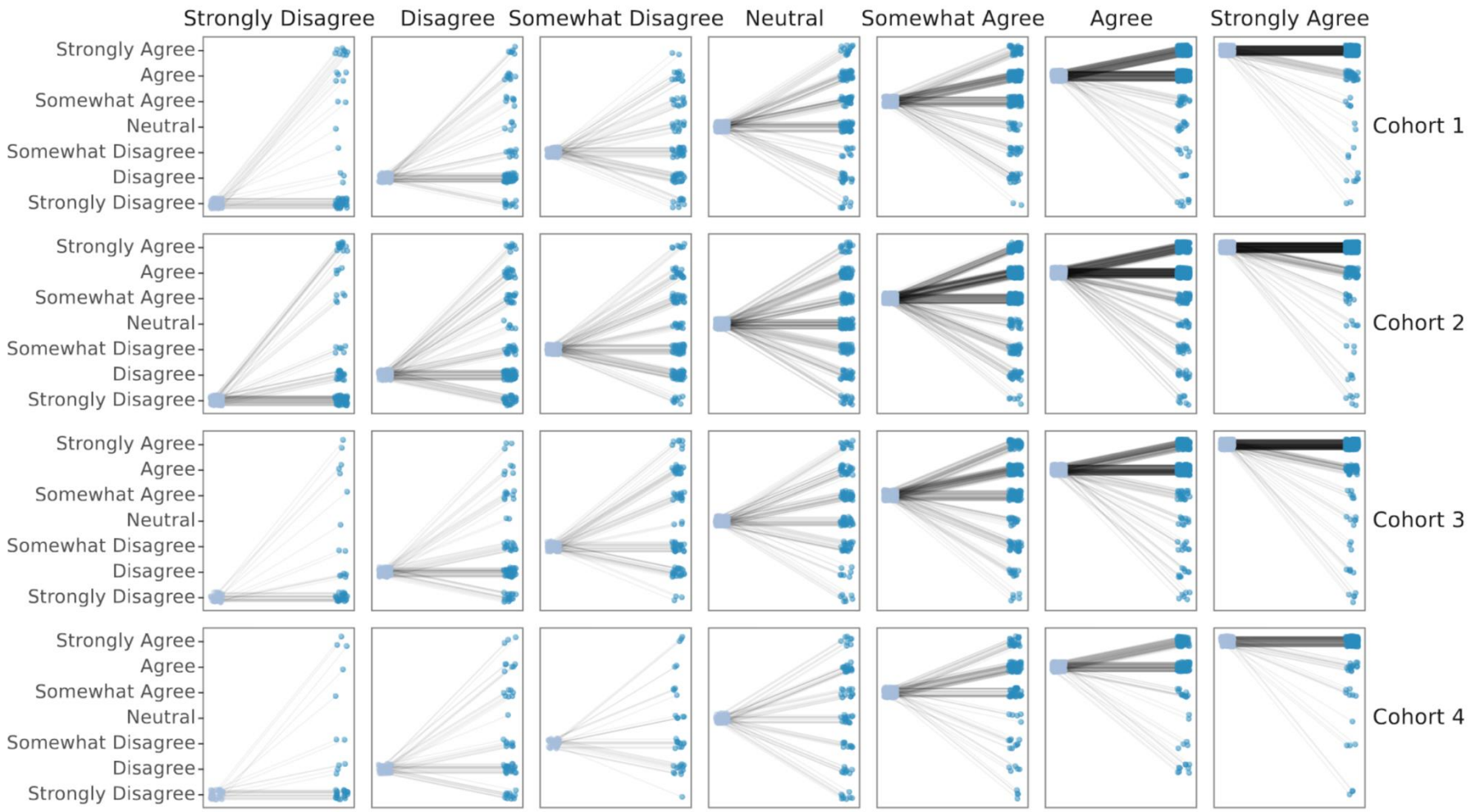
- Do not scale to thousands of people
- Are costly, inconvenient or time-consuming
- Are not strategy proof
- Do not represent all voices equally
- Are prone to social desirability effects
- May licit cognitive biases during reasoning

Public deliberation using LLMs



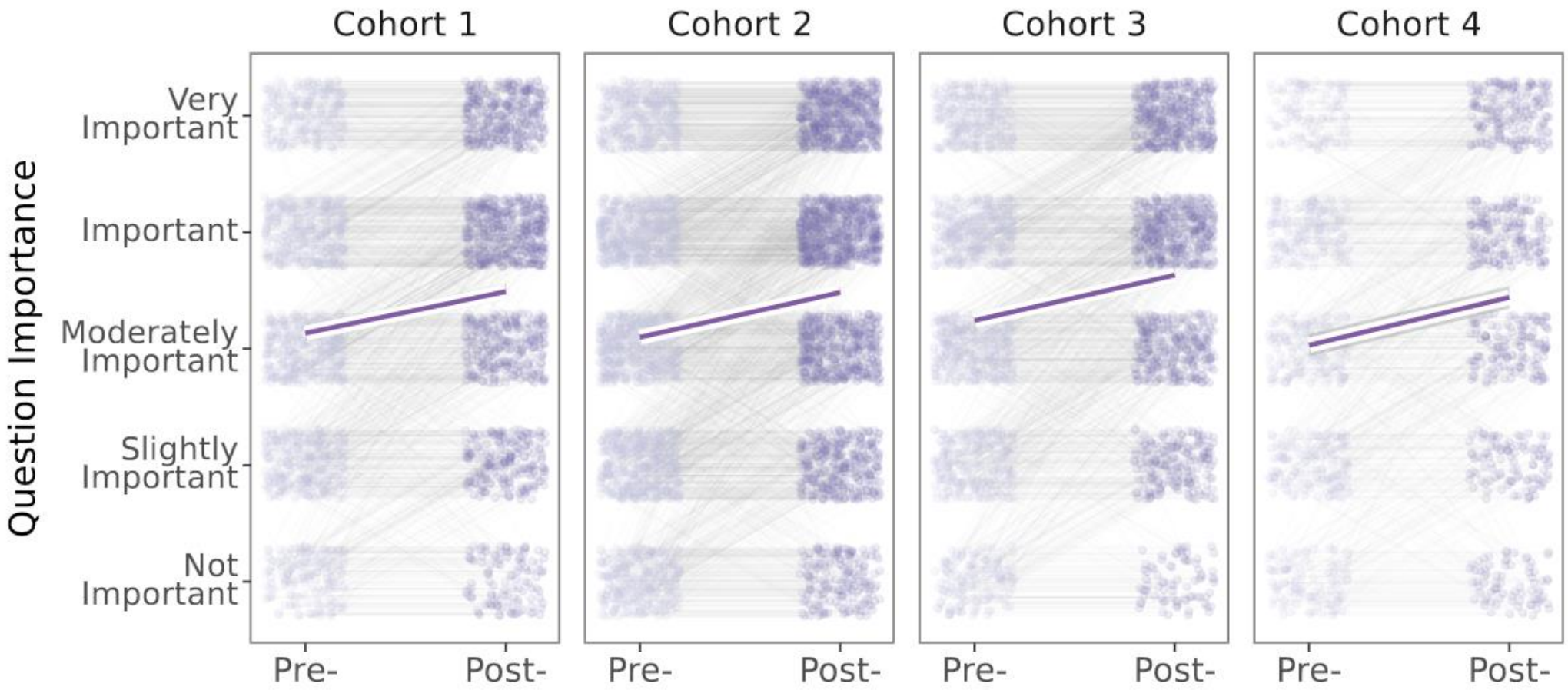
Ablation experiments reveal that both fine-tuning steps are important, but especially the reward modelling...

Public deliberation using LLMs



Movement is mainly but not exclusively in the majority direction

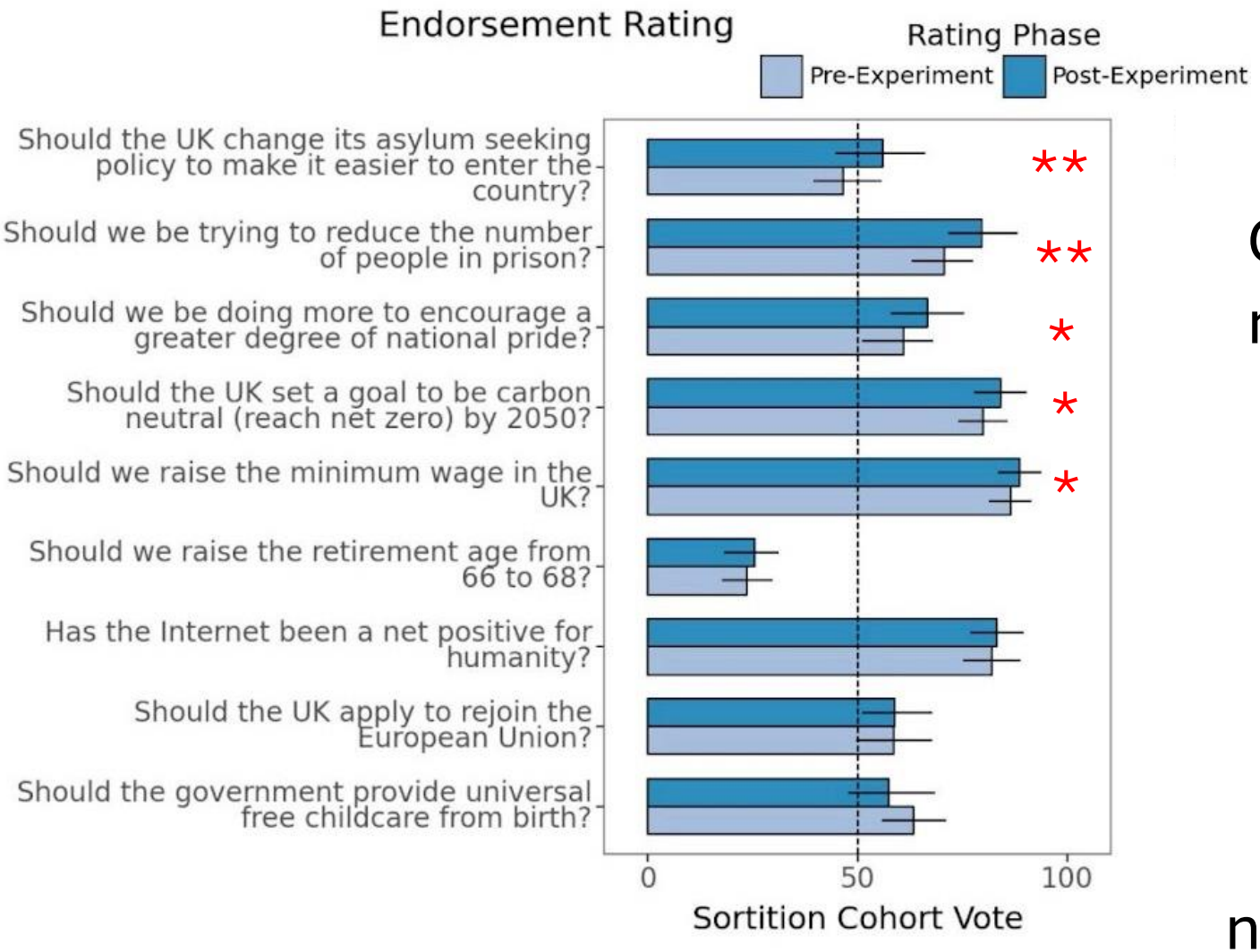
Public deliberation using LLMs



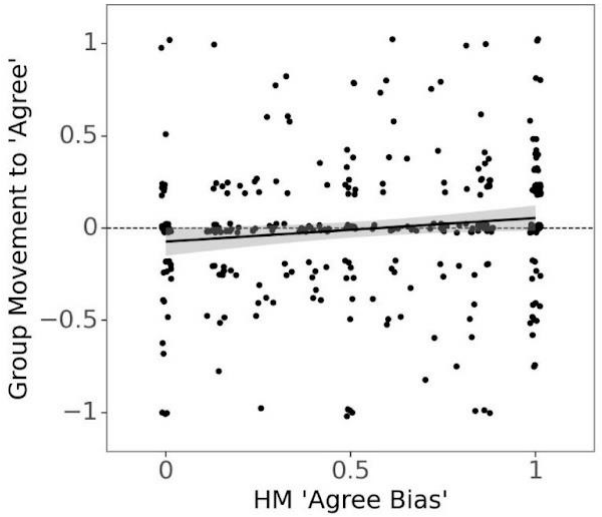
Participants find questions to be more important after deliberation

Public deliberation using LLMs

SORTITION



Groups of people tend to move in a common direction!



not simply due to model bias

Society is governed by institutions

kinship or social group

education and work

organised belief systems

economy

political system

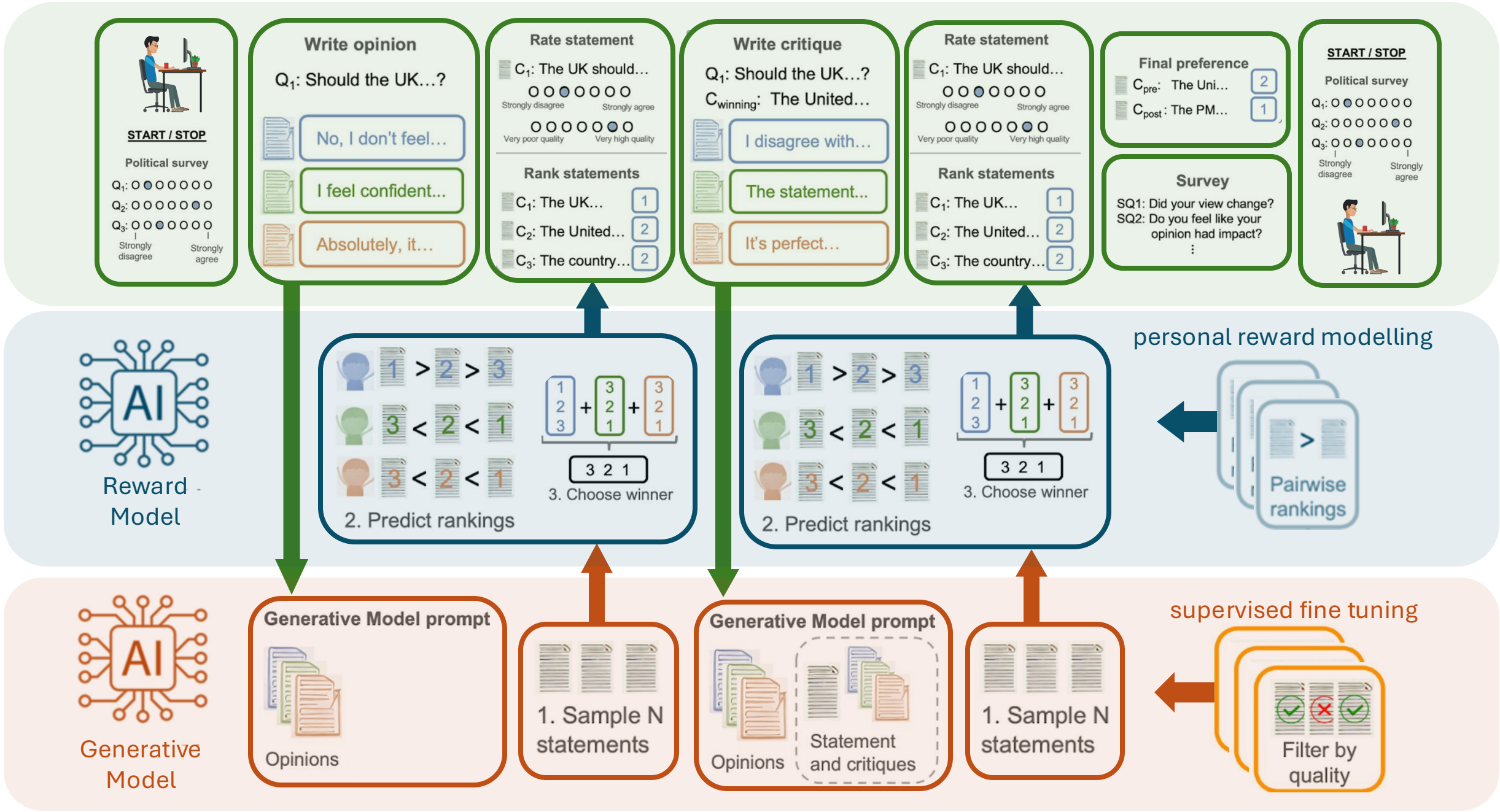
Institutions are the *rules of the game* – they set the incentive structure for society

Douglas C North
(Nobel Prize 1993)

Thanks

We are hiring

- AI researchers
- cognitive scientists, statisticians
- economists
- behavioural scientists
- computational social scientists
- data scientists


START / STOP
Political survey
Q1:
Q2:
Q3:
Strongly disagree
Strongly agree
Write opinion
Q1: Should the UK...?
No, I don't feel...
I feel confident...
Absolutely, it...
Rate statement
C1: The UK should...
Strongly disagree
Strongly agree
Very poor quality
Very high quality
Rank statements
C1: The UK... 1
C2: The United... 2
C3: The country... 2
Write critique
Q1: Should the UK...?
Cwinning: The United...
I disagree with...
The statement...
It's perfect...
Rate statement
C1: The UK should...
Strongly disagree
Strongly agree
Very poor quality
Very high quality
Rank statements
C1: The UK... 1
C2: The United... 2
C3: The country... 2
Final preference
Cpre: The Uni... 2
Cpost: The PM... 1
Survey
SQ1: Did your view change?
SQ2: Do you feel like your opinion had impact?
START / STOP
Political survey
Q1:
Q2:
Q3:
Strongly disagree
Strongly agree
AI
Reward Model
1 > 2 > 3
3 < 2 < 1
3 < 2 < 1
2. Predict rankings
1 2 3 + 3 2 1 + 3 2 1
3 2 1
3. Choose winner
1 > 2 > 3
3 < 2 < 1
3 < 2 < 1
2. Predict rankings
1 2 3 + 3 2 1 + 3 2 1
3 2 1
3. Choose winner
personal reward modelling
Pairwise rankings
AI
Generative Model
Generative Model prompt
Opinions
1. Sample N statements
Generative Model prompt
Opinions
Statement and critiques
1. Sample N statements
supervised fine tuning
Filter by quality