AI for Social Sciences

Dr Mary Sanford

Climate Change AI Summer School

22 July 2026







Warm up question(s)

Are you anxious about AI?



https://partici.fi/97320810

If so, why? If not, why not?



https://partici.fi/97320810

What are the social sciences?

Management/Finance

Government

Political science

Archaeology

Economics

Geography

Psychology

Education

Law

International relations

Sociology

Anthropology

Communications

My focus

Social science research on climate change

What do people think?

What do people do?

What influences people?

What about organisations?

What are the systemic implications?

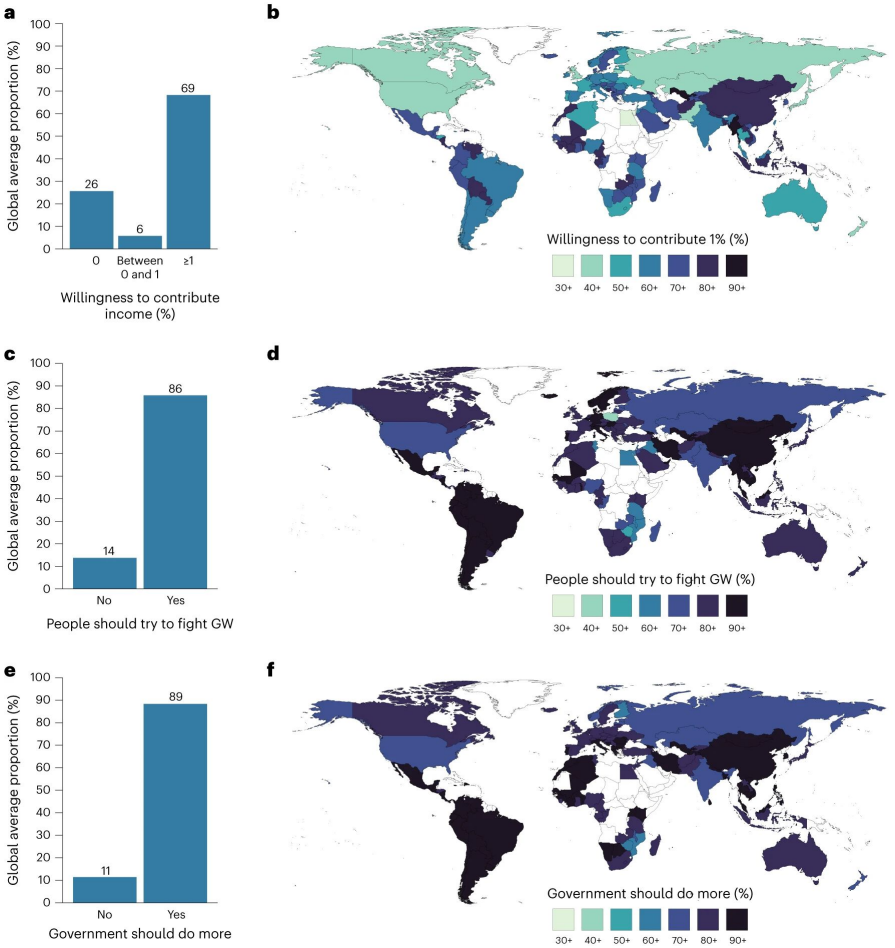
What can we learn from the past?

➔ Critical for building social and political feasibility for a just climate transition.
➔ **Necessary** for making sure technological advancements can reach their full potential.

What do people think?

- ➔ Big multinational surveys
- ➔ Self-reported support, opposition, attitudes, behaviours
- ➔ Mostly one-off, sometimes longitudinal
- ➔ A lot of descriptive work, some more causal/behavioural

Widespread global support for climate action



Andre et al. 2024. *Nature Climate Change*.

What do people do (or don't do)?

ELSEVIER

Global Environmental Change
Volume 17, Issues 3–4, August–October 2007, Pages 445-459



Barriers perceived to engaging with climate change among the UK public and their policy implications

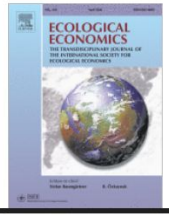
Irene Lorenzoni [a b], Sophie Nicholson-Cole [b], Lorraine Whitmarsh [b]

ELSEVIER

Journal of Environmental Psychology
Volume 51, August 2017, Pages 226-238



"Green to be keep down": effect of identity on pro-environmental behavior



Cameron Brick, David K. Sherman, Heejung S. Kim

Research article • Full text access Get rights and content

Do risk, time, and social preferences predict sustainable behavior? Evidence from a qualitative synthesis and meta-analysis ☆

Kristopher L. Nichols [a b], Johanna M. Matt-Navarro [a], Maxim K. Doiron [a], Giovanna d'Adda [c d], Elke U. Weber [a b e], Sara M. Constantino [e f]

Show more

- ➔ Some more experiments
- ➔ Secondary data analysis



The Political Consequences of Green Policies: Evidence from Italy

Published online by Cambridge University Press: **27 April 2023**

ITALO COLANTONE, LIVIO DI LONARDO, YOTAM MARGALIT and MARCO PERCOCO

Show author details

What influences people?



What shapes public support for climate change mitigation policies? The role of descriptive social norms and elite cues

Published online by Cambridge University Press: **03 November 2020**

ADRIAN RINSCHEID, SILVIA PIANTA and ELKE U. WEBER

Show author deta

European Journal of Political Research

Research Article | Open Access

Carbon inequality and support for carbon taxation

LIAM F. BEISER-McGRATH, MARIUS R. BUSEMEYER

First published: 15 December 2023 | https://doi.org/10.1111/1475-6765.12647 | VIEW METRIC

"Fear Won't Do It": Promoting Positive Engagement With Climate Change Through Visual and Iconic Representations

Saffron O'Neill and Sophie Nicholson-Cole View all authors and affiliations

Volume 30, Issue 3 | https://doi.org/10.1177/1075547008329201

Letter | Published: 30 November 2015

Network structure and influence of the climate change counter-movement

Justin Farrell

Nature Climate Change **6**, 370–374 (2016) | Cite this article

Environment, Development and Sustainability (2025) 27:3593–3607
https://doi.org/10.1007/s10668-023-04031-0

Check for updates

How does personalized feedback on carbon emissions impact intended climate action?

Yu Luo[1] · Brynley Hanson-Wright[2] · Hadi Dowlatabadi[2] · Jiaying Zhao[1,2]

Received: 26 April 2023 / Accepted: 9 October 2023 / Published online: 25 October 2023

What about organisations?

Home > Crises and Disruptions in International Business > Chapter

The Impact of Climate Risk on Firm Performance and Financing Choices: A International Comparison

Chapter | First Online: 26 April 2022

Article | Open access | Published: 18 December 2023

Do fossil fuel firms reframe online climate and sustainability communication? A data-driven analysis

Ramit Debnath, Danny Ebanks, Kamiar Mohaddes, Thomas Roulet & R. Michael Alvarez

npj Climate Action **2**, Article number: 47 (2023) | Cite this article

JOURNAL ARTICLE

Investor Rewards to Climate Responsibility: Stock-Price Responses to the Opposite Shocks of the 2016 and 2020 U.S. Elections

Get access

Stefano Ramelli, Alexander F Wagner, Richard J Zeckhauser, Alexandre Ziegler

The Review of Corporate Finance Studies, Volume 10, Issue 4, December 2021, Pages 748–787, https://doi.org/10.1093/rcfs/cfab010

Published: 17 June 2021 **Article history**

JOURNAL ARTICLE

Sectors, Pollution, and Trade: How Industrial Interests Shape Domestic Positions on Global Climate Agreements

Federica Genovese

International Studies Quarterly, Volume 63, Issue 4, December 2019, Pages 819–836, https://doi.org/10.1093/isq/sqz062

Published: 27 August 2019

Systemic implications?

RESEARCH ARTICLE | SUSTAINABILITY SCIENCE |

Social tipping dynamics for stabilizing Earth's climate by 2050

Ilona M. Otto, Jonathan F. Donges, Roger Cremades, +10, and Hans Joachim Schellnhuber Authors Info & Affiliations

Contributed by Hans Joachim Schellnhuber, November 15, 2019 (sent for review January 22, 2019; reviewed by J. David Tabara and Jessika E. Trancik)

January 21, 2020 | 117 (5) 2354-2365 | https://doi.org/10.1073/pnas.1900577117

ELSEVIER

World Development

Volume 115, March 2019, Pages 246-257



Poverty and distributional effects of carbon pricing in low- and middle-income countries – A global comparative analysis

Ira Irina Dorband [a b], Michael Jakob [a], Matthias Kalkuhl [a c], Jan Christoph Steckel [a b d]

ELSEVIER

Journal of Environmental Economics and Management

Volume 127, September 2024, 103012



Climate change impacts on the within-country income distributions ☆

Martino Gilli [a b c], Matteo Calcaterra [a b d], Johannes Emmerling [a b], Francesco Granella [a b]

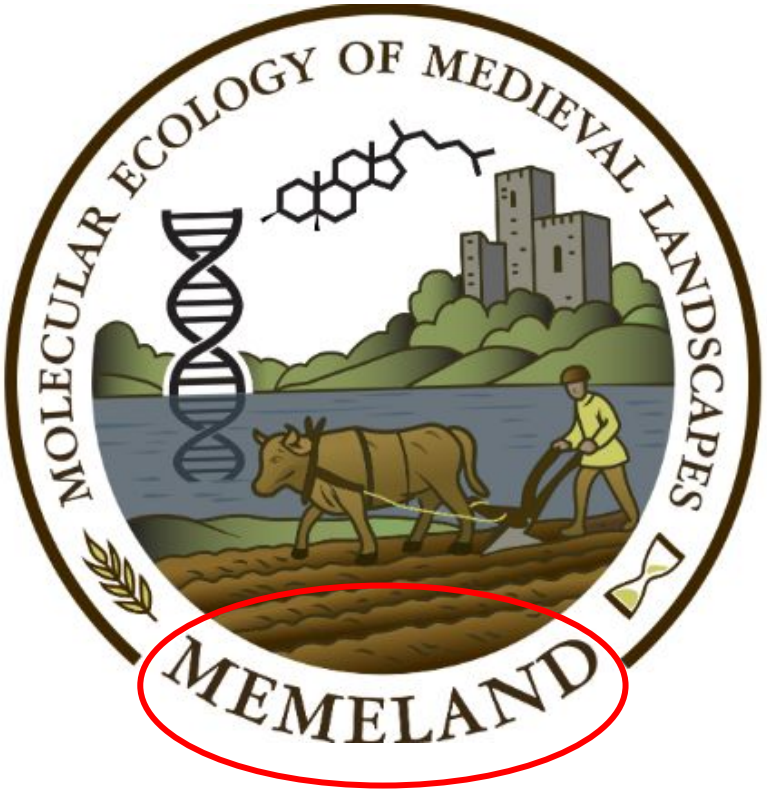
Analysis | Published: 29 November 2021

Climate action with revenue recycling has benefits for poverty, inequality and well-being

Mark Budolfson, Francis Dennig, Frank Errickson, Simon Feindt, Maddalena Ferranna, Marc Fleurbaey, David Klenert, Ulrike Kornek, Kevin Kuruc, Aurélie Méjean, Wei Peng, Noah Scovronick, Dean Spears, Fabian Wagner & Stéphane Zuber

Nature Climate Change **11**, 1111–1116 (2021) | Cite this article

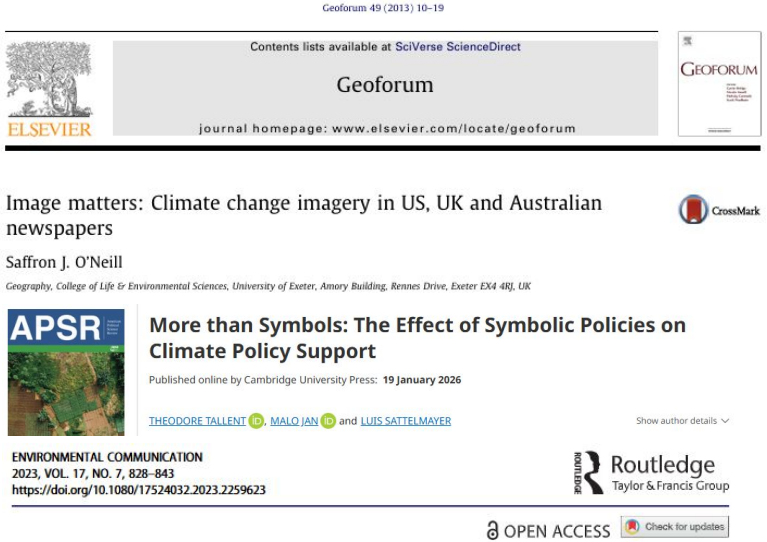
Learning from the past



But…not all social science research for climate change is **quantitative**.

Qualitative work permits deeper understanding

- ➔ Critical for hypothesis development: identify and describe new phenomena
- ➔ Qualitative content analysis: investigate details and subjective angles that would get lost in quantitative analysis.
- ➔ Interviews, focus groups: ask questions to get further information that wouldn't be possible with surveys, observational analysis
- ➔ Can be paired with larger data collection efforts to test hypotheses at scale



Geoforum 49 (2013) 10–19

Contents lists available at SciVerse ScienceDirect

Geoforum

journal homepage: www.elsevier.com/locate/geoforum

Image matters: Climate change imagery in US, UK and Australian newspapers

Saffron J. O'Neill

Geography, College of Life & Environmental Sciences, University of Exeter, Amory Building, Rennes Drive, Exeter EX4 4RJ, UK

APSR

More than Symbols: The Effect of Symbolic Policies on Climate Policy Support

Published online by Cambridge University Press: **19 January 2026**

THEODORE TALLENT, MALO JAN and LUIS SATTELMAYER

Show author details

ENVIRONMENTAL COMMUNICATION
2023, VOL. 17, NO. 7, 828–843
https://doi.org/10.1080/17524032.2023.2259623

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Extreme Weather Events as Teachable Moments: Catalyzing Climate Change Learning and Action Through Conversation

Joshua Ettinger [a], Peter Walton [b], James Painter [c], Susan A. Flocke [d] and Friederike E.L. Otto [e]

Article | Open access | Published: 04 August 2025

United Nations Environment Assembly attendees underestimate public willingness to contribute to climate action

Ximeng Fang, Joshua Ettinger & Stefania Innocenti

Communications Earth & Environment **6**, Article number: 622 (2025) | Cite this article

Conceptual and theoretical work important too

- ➔ Developing and sharing new perspectives, ideas for research agendas, positions
- ➔ Building new theories
- ➔ Communicating within and across fields
- ➔ Very important for shaping research directions

Transportation Research Interdisciplinary Perspectives
Volume 37, May 2026, 102029

Sufficient mobility and access within limits: Research agenda for bringing together corridor frameworks and transportation research

Michał Czepkiewicz [a c], Giulio Mattioli [b], Filip Schmidt [c], Elias Willberg [d], Lena Kilian [e n], Henrikki Tenkanen [f], Dick Timmer [g], Ákos Gosztonyi [h], Johanna Raudsepp [i], Sanna Ala-Mantila [h], Lisa Jacobson [j], Mònica Guillen-Royo [k], Dawid Krysiński [l c], Kevin Dillman [i], Jukka Heinonen [i], Petter Næss [m 1]

Perspective | Published: 30 September 2021

The role of high-socioeconomic-status people in locking in or rapidly reducing energy-driven greenhouse gas emissions

Dialogues on Climate Change
Volume 1, Issue 1, December 2024, Pages 23-29
© The Author(s) 2024, Article Reuse Guidelines
https://doi.org/10.1177/29768659241293215
Sage Journals

Commentaries

Are radical changes to lifestyles necessary for mitigating climate change?

Lorraine Whitmarsh [1] and Sam Hampton [2,3]

Psychological Science in the Public Interest
Volume 23, Issue 2, October 2022, Pages 50-97
© The Author(s) 2022, Article Reuse Guidelines
https://doi.org/10.1177/15291006221105279
SAGE journals

Scaling Up Change: A Critical Review and Practical Guide to Harnessing Social Norms for Climate Action

Sara M. Constantino [1,2,3], Gregg Sparkman [4], Gordon T. Kraft-Todd [4], Cristina Bicchieri [5,6,7], Damon Centola [8,9], Bettina Shell-Duncan [10], Sonja Vogt [11,12,13], and Elke U. Weber [3,14,15]

Thought break

Question

Can you think of a social science research topic or question related to climate change that you would like to learn more about?



https://partici.fi/97320810

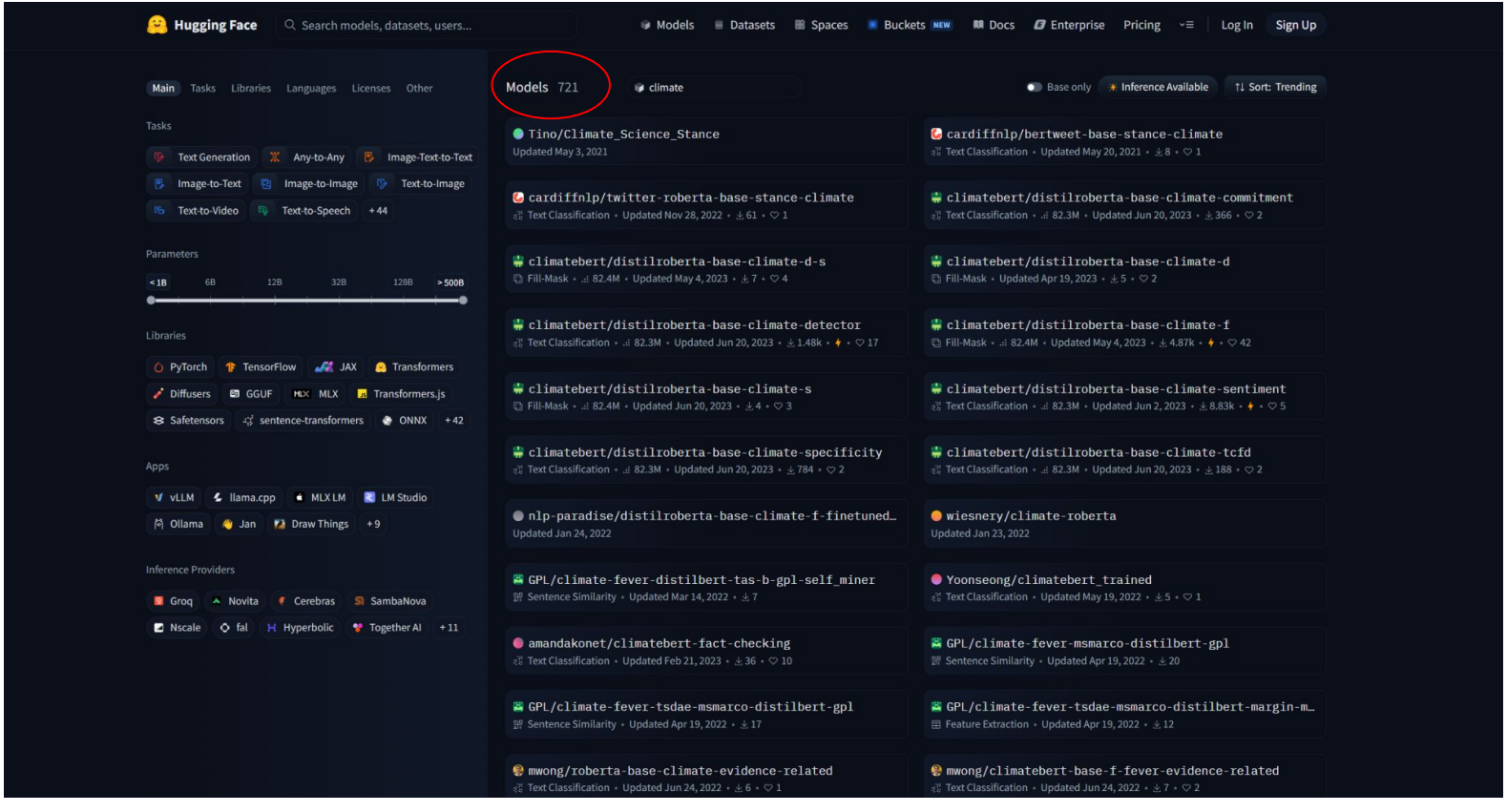
Methodologically

Research process

How has AI changed the landscape?

Scope of research questions

Text classification at scale

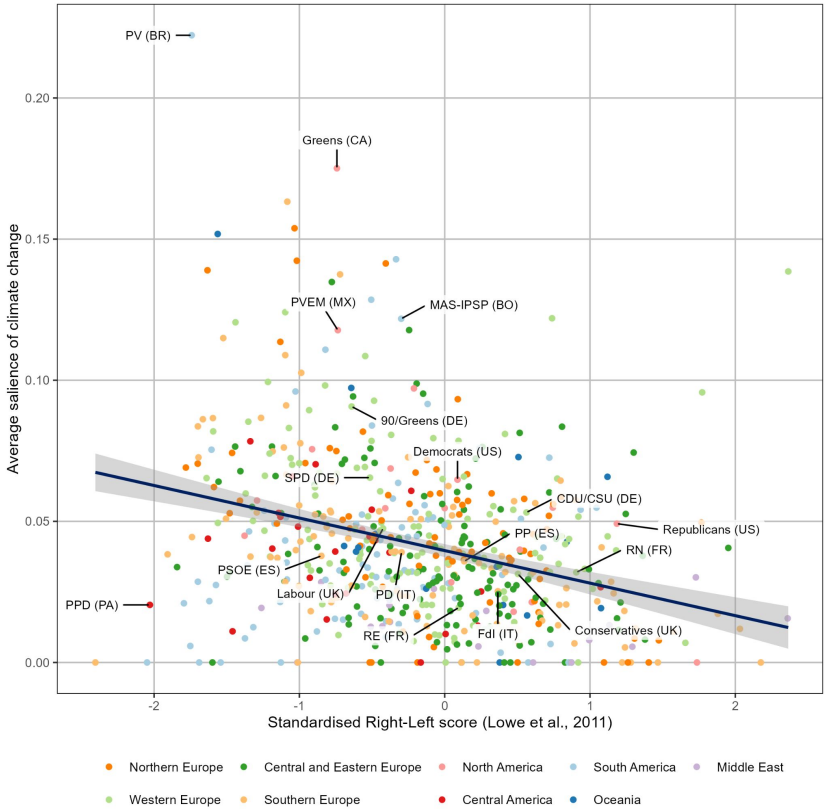


Policlim: A Dataset of Climate Change Discourse in the Political Manifestos of Forty-Five Countries from 1990 to 2022

Published online by Cambridge University Press: **23 October 2025**

Mary Sanford, Silvia Pianta, Nicolas Schmid and Giorgio Musto Show author details

- ➔ Fine tunes **XLM-RoBERTa** to detect climate change salience in political manifestos across 20+ languages
- ➔ Model and variable ready for use in other work



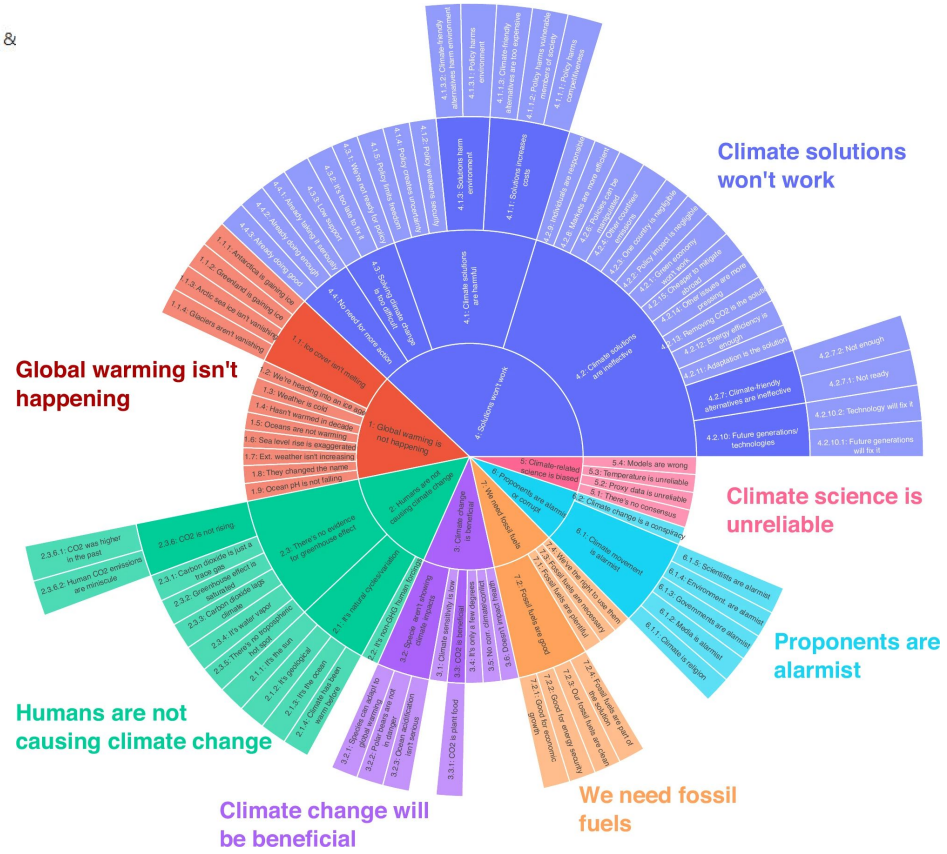
Article | Open access | Published: 27 February 2026

Large language model reveals an increase in climate contrarian speech in the United States Congress

Travis G. Coan ✉, Ranadheer Malla, Mirjam O. Nanko, William Kattrup, J. Timmons Roberts, John Cook & Constantine Boussalis

Communications Sustainability **1**, Article number: 37 (2026) | Cite this article

➔ Takes complex taxonomy of contrarian claims



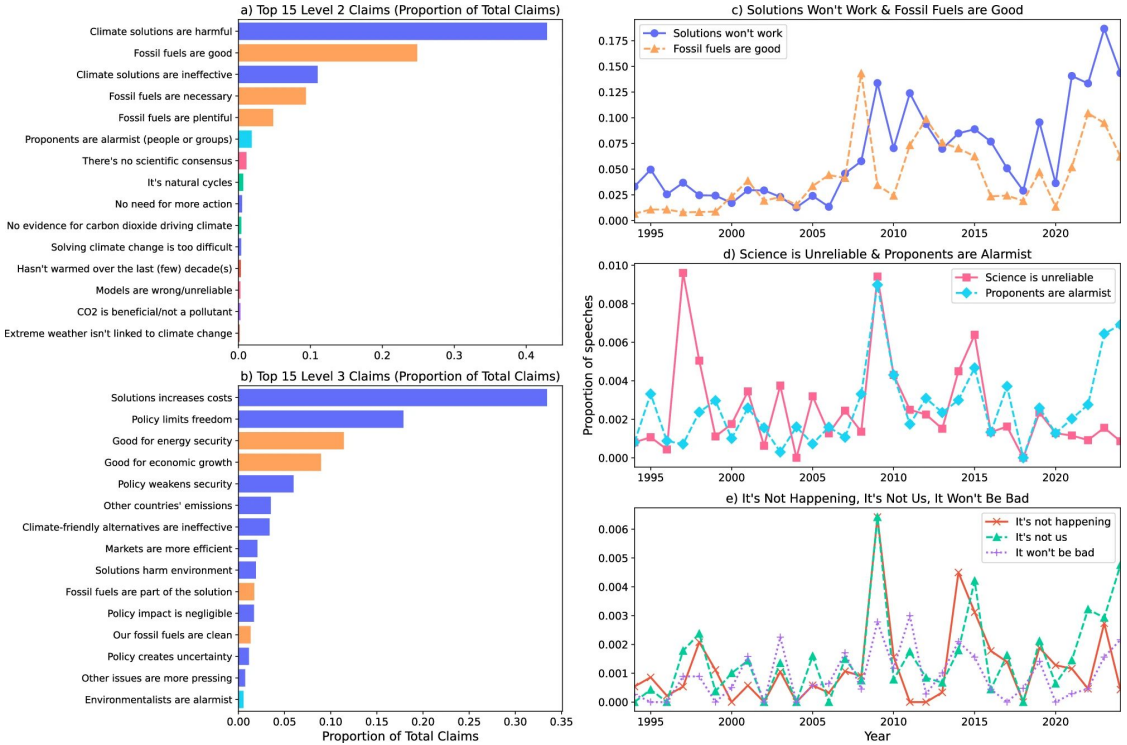
Article | Open access | Published: 27 February 2026

Large language model reveals an increase in climate contrarian speech in the United States Congress

Travis G. Coan ✉, Ranadheer Malla, Mirjam O. Nanko, William Kattrup, J. Timmons Roberts, John Cook & Constantine Boussalis

Communications Sustainability **1**, Article number: 37 (2026) | Cite this article

- ➔ Takes complex taxonomy of contrarian claims and fine-tunes GPT4o and Claude Sonnet 3.7 to detect them.
- ➔ US Congressional speeches
- ➔ Track different narratives over time (and space)



Article | Open access | Published: 24 November 2022

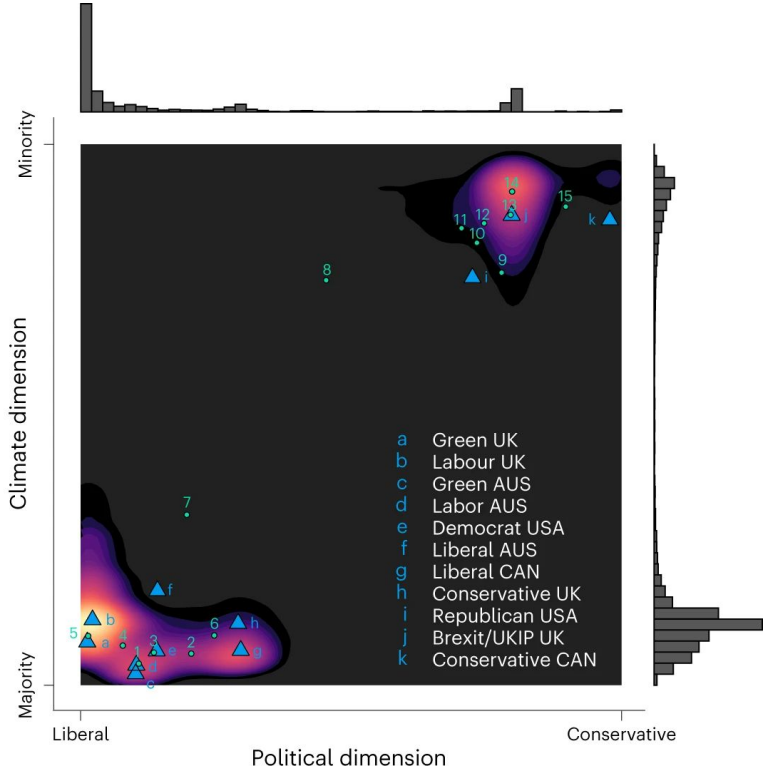
Growing polarization around climate change on social media

Max Falkenberg, Alessandro Galeazzi, Maddalena Torricelli, Niccolò Di Marco, Francesca Larosa, Madalina Sas, Amin Mekacher, Warren Pearce, Fabiana Zollo, Walter Quattrociocchi & Andrea Baronchelli

Nature Climate Change **12**, 1114–1121 (2022) | Cite this article

- ➔ Analyses retweet interactions between influential accounts in discourse around COPs to detect polarisation.
- ➔ Combines this with **BERTopic*** to further contextualise

1 Anti-Morrison
2 Biodiversity & Localism, Women at COP, Climate Finance, Youth,
3 Climate Justice Decolonise Climate
4 Sustainable Transport
5 Activitsts for Climate Action
6 Climate Crisis
7 Anti-Johnson
8 Anti-Sturgeon
9 Climate Alarmists
10 Wind Energy & Nuclear
11 Thunberg Cult
12 Chinese Absence, Vaccine Passports
13 Anti-Biden, Didn't Vote Green, Arctic Ice, Climate "Scam", Benefits of Warming
14 Illegal Immigration, Net Zero Referendum, Private Jets, Hypocrisy, CO_2
15 Anti-Trudeau



AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

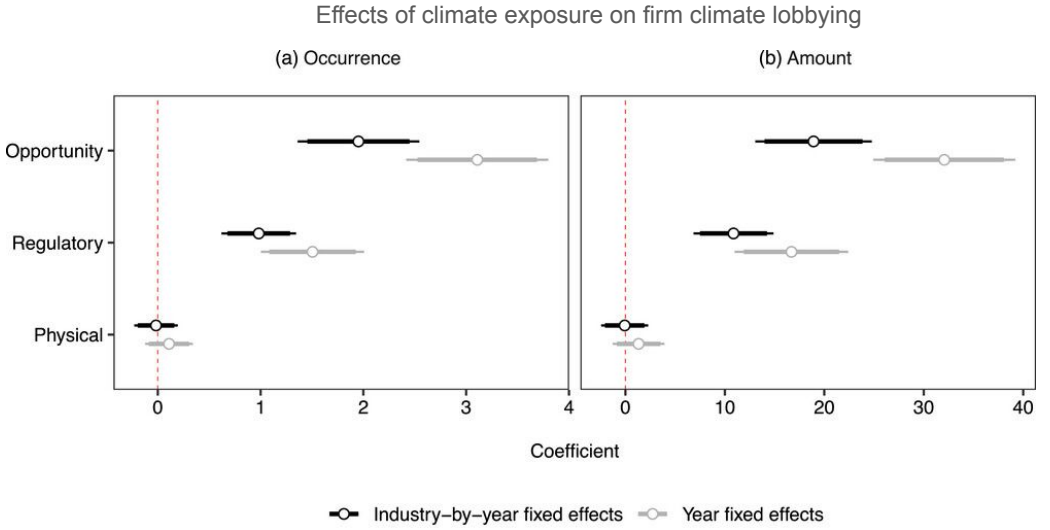
ARTICLE | Open Access

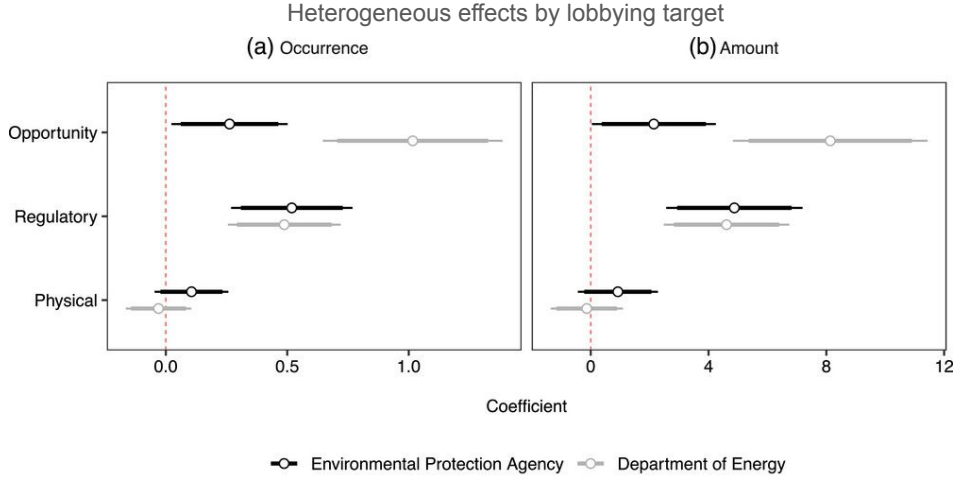
Climate exposure drives firm political behavior: Evidence from earnings calls and lobbying data

Christian Baehr, Fiona Bare, Vincent Heddesheimer

First published: 26 April 2026 | https://doi.org/10.1111/ajps.70062 | VIEW METRICS

- ➔ Uses **ClimateBert** to analyse earnings calls
- ➔ Identify firms who mention opportunities related to climate change, as well as and regulatory and physical risks.
- ➔ Those who mention opportunities and regulatory risks more also tend to lobby more.





Open access | Research article | First published online September 19, 2025 | Social Science Computer Review

Leveraging VLLMs for Visual Clustering: Image-to-Text Mapping Shows Increased Semantic Capabilities and Interpretability

Luigi Arminio, Matteo Magnani, [...], and Alexandra Segerberg +2 View all authors and affiliations

Volume 44, Issue 3 | https://doi.org/10.1177/08944393251376703 | View article versions

- ➔ Use VLLMs (gpt-4-turbo and LLaVa) to generate text descriptions of images
- ➔ Then use this to cluster the images for further quantitative and qualitative analysis
- ➔ *"One aspect that should be checked is the **potential bias** introduced in the analysis, both with respect to **possible mistakes**, but also with respect to **possible preferences** assigned to one instead of a different cluster"*






DANGER

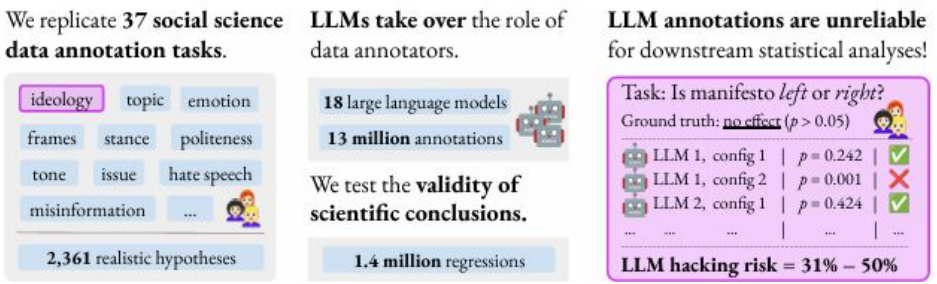
Misclassification bias

- No prediction is perfect
- Far too few development pipelines contain thorough validation checks
- Results in bias in downstream analysis
- Even more problematic for LLMs

Large Language Model Hacking:
Quantifying the Hidden Risks of Using LLMs for Text Annotation

Joachim Baumann[1], Paul Röttger[1], Aleksandra Urman[2], Albert Wendsjö[3], Flor Miriam Plaza-del-Arco[4], Johannes B. Gruber[5], and Dirk Hovy[1]

[1]*Bocconi University*
[2]*University of Zurich*
[3]*University of Gothenburg*
[4]*LIACS, Leiden University*
[5]*GESIS, Leibniz Institute for the Social Sciences*



Misclassification bias

Before using a new model…

- ➔ Always check the training data (examining examples if possible) to understand if the context is appropriate
- ➔ Take steps to mitigate bias where possible (TeBlunthuis et al., 2024)
- ➔ Avoid basing model performance evaluations only on training. Always validate predictions too.



COMMUNICATION METHODS AND MEASURES
2024, VOL. 18, NO. 3, 278–299
https://doi.org/10.1080/19312458.2023.2293713

Routledge
Taylor & Francis Group

Check for updates

Misclassification in Automated Content Analysis Causes Bias in Regression. Can We Fix It? Yes We Can!

Nathan TeBlunthuis[a,b], Valerie Hase[c], and Chung-Hong Chan[d]

Survey analysis

Energy Research & Social Science
Volume 73, March 2021, 101883



Why hate carbon taxes? Machine learning evidence on the roles of personal responsibility, trust, revenue recycling, and other factors across 23 European countries

Sebastian Levi

Research article • Open access Under a Creative Commons license Get rights and content

Unpacking the side effects of social norm nudges ☆

Julien Picard [a b c], Sanchayan Banerjee [d] Show more

- ➔ New ways of studying relationships between variables, e.g., discover latent structures, predictive strength.
- ➔ Can increase precision in models.
- ➔ Can boost causal inference pathways depending on data and context.
- ➔ Important to address uncertainty.
- ➔ *Always needs theoretical grounding*.

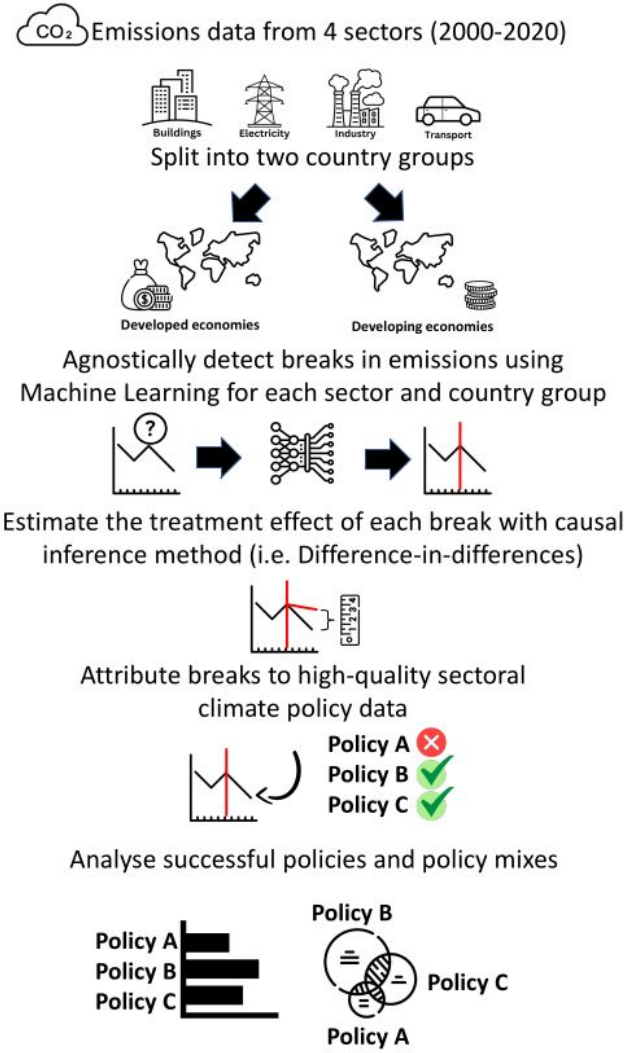
Policy evaluation

RESEARCH ARTICLE | CLIMATE POLICY

Climate policies that achieved major emission reductions: Global evidence from two decades

ANNIKA STECHEMESSER, NICOLAS KOCH, EBBA MARK, ELINA DILGER, PATRICK KLÖSEL, LAURA MENICACCI, DANIEL NACHTIGALL, FELIX PRETIS, NOLAN RITTER, [...], AND ANNA WENZEL +2 authors Authors Info & Affiliations

SCIENCE · 22 Aug 2024 · Vol 385, Issue 6711 · pp. 884-892 · DOI: 10.1126/science.adl6547



Evidence synthesis

Article | Published: 11 October 2021

Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies

Max Callaghan, Carl-Friedrich Schleussner, Shruti Nath, Quentin Lejeune, Thomas R. Knutson, Markus Reichstein, Gerrit Hansen, Emily Theokritoff, Marina Andrijevic, Robert J. Brecha, Michael Hegarty, Chelsea Jones, Kaylin Lee, Agathe Lucas, Nicole van Maanen, Inga Menke, Peter Pfleiderer, Burcu Yesil & Jan C. Minx

Nature Climate Change **11**, 966–972 (2021) | Cite this article

Article | Open access | Published: 16 May 2024

Systematic review and meta-analysis of ex-post evaluations on the effectiveness of carbon pricing

Niklas Döbbeling-Hildebrandt, Klaas Miersch, Tarun M. Khanna, Marion Bachelet, Stephan B. Bruns, Max Callaghan, Ottmar Edenhofer, Christian Flachsland, Piers M. Forster, Matthias Kalkuhl, Nicolas Koch, William F. Lamb, Nils Ohlendorf, Jan Christoph Steckel & Jan C. Minx

Nature Communications **15**, Article number: 4147 (2024) | Cite this article

Article | Open access | Published: 07 September 2024

Public support for carbon pricing policies and revenue recycling options: a systematic review and meta-analysis of the survey literature

Farah Mohammadzadeh Valencia, Cornelia Mohren, Anjali Ramakrishnan, Marlene Merchert, Jan C. Minx & Jan Christoph Steckel

npj Climate Action **3**, Article number: 74 (2024) | Cite this article

- ➔ Completely changes what is possible for evidence synthesis.
- ➔ Faster identification of what is relevant and why.
- ➔ Faster extraction of key information for synthesis and meta-analyses.
- ➔ Faster input for experts and policymakers.

AI for qualitative analysis → open-ended survey questions

› Psychol Methods. 2025 Apr 7. doi: 10.1037/met0000753. Online ahead of print.

The use of large language models for qualitative research: The Deep Computational Text Analyser (DECOTA)

Lois Player [1], Ryan Hughes [2], Kaloyan Mitev [1], Lorraine Whitmarsh [1], Christina Demski [1], Nicholas Nash [1], Trisevgeni Papakonstantinou [3], Mark Wilson [1]

Affiliations + expand

PMID: 40193412 DOI: 10.1037/met0000753

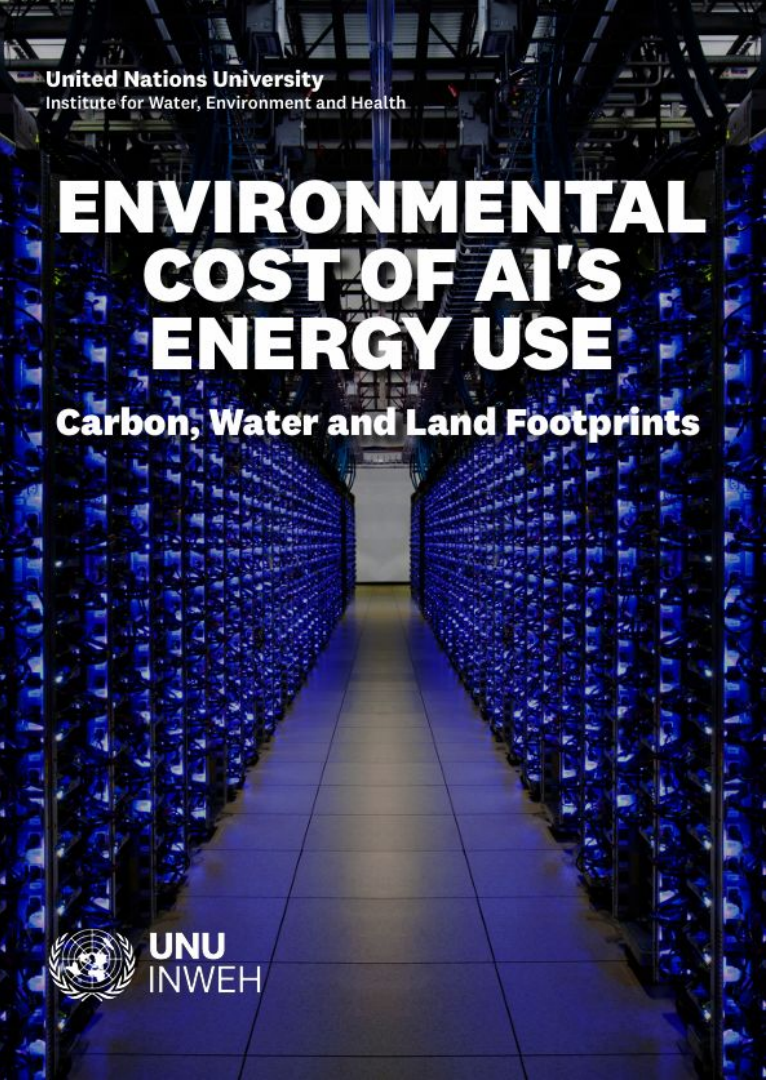
JOURNAL ARTICLE EDITOR'S CHOICE

Scaling Open-Ended Survey Responses Using LLM-Paired Comparisons

Matthew R DiGiuseppe ✉, Michael E Flynn

Public Opinion Quarterly, Volume 90, Issue 3, Fall 2026, Pages 630–656,
https://doi.org/10.1093/poq/nfag013
Published: 27 March 2026

Current debates on AI, social science, and climate change research



GPT-4's training carbon footprint of 25,000 tonnes of CO_2e would require the sequestration capacity of 420,000 tree seedlings grown for 10 years, or about equal to the number of trees in 105 Hyde Parks in London.

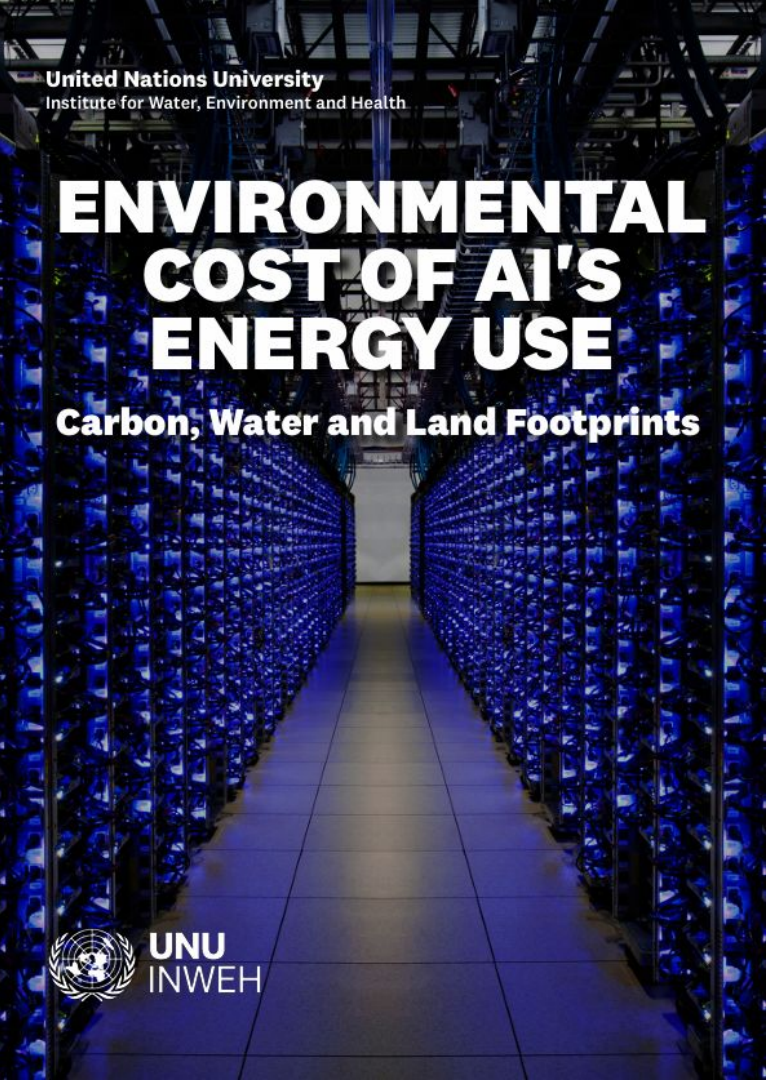
The water footprint associated with training GPT-4 was about 600 million liters, enough to meet the minimum annual domestic water needs of 81,000 people in Sub-Saharan Africa, or to fill 237 Olympic-sized pools.

Training models like GPT-5 is estimated to have a carbon footprint of 42,000 tonnes of CO_2e, requiring 700,000 tree seedlings (about equal to the number of trees in 40 Central Parks in New York or 155 times the trees in Toronto's High Park over 10 years to offset).

The water footprint of GPT-5 training is estimated at 1 billion liters, enough to meet the annual domestic water needs of more than 135,000 people in Sub-Saharan Africa.

While training is highly resource-intensive, the continuous inference phase used to generate responses for billions of interactions is estimated to account for 80% to 90% of total AI energy use.

A typical ChatGPT-style text query is about 200 times more energy-intensive than text classification (such as spam filtering).



The global AI market is expanding rapidly, projected to grow from USD 189 billion in 2023 to nearly USD 5 trillion by 2033.

Generative AI accounted for over 20% of the global AI market in 2026, and is projected to reach 40% by 2030.

AI workloads accounted for roughly 20% of total data centers' electricity use in 2025 and are projected to grow to 40% by 2030.

Data centers' electricity use in 2025 had a carbon footprint of 189 million tonnes of CO_2e, which would require 3.2 billion tree seedlings grown over 10 years to offset, roughly the total number of trees in the entire United Kingdom.

The associated water footprint of projected 2030 electricity consumption of data centers is 9.3 trillion liters, or enough to meet the minimum annual domestic water needs of all 1.3 billion residents of Sub-Saharan Africa for a full year.

Projected global data centers' electricity consumption could exceed 945 TWh by 2030, accounting for almost 3% of projected global electricity use—enough to supply residential electricity to all 1.3 billion people in Sub-Saharan Africa for about 5.5 years.

US POLICY | 29 May 2026 13:58

AI boom means US is now ‘investing more’ in fossil-fuel power than China



Meta’s $13-billion Alberta data centre a potential boost for utilities and natural gas producers

ANDREW WILLIS >

Report warns AI data centre boom threatens Australia's energy transition

By national AI reporter **Cam Wilson**

AI

Tue 26 May

Revealed: landmark Scottish AI project has no prospect of meeting renewables promise

Exclusive: Government and developers privately acknowledged Lanarkshire datacentre site had power provision ‘issue’

BREAKING | BUSINESS

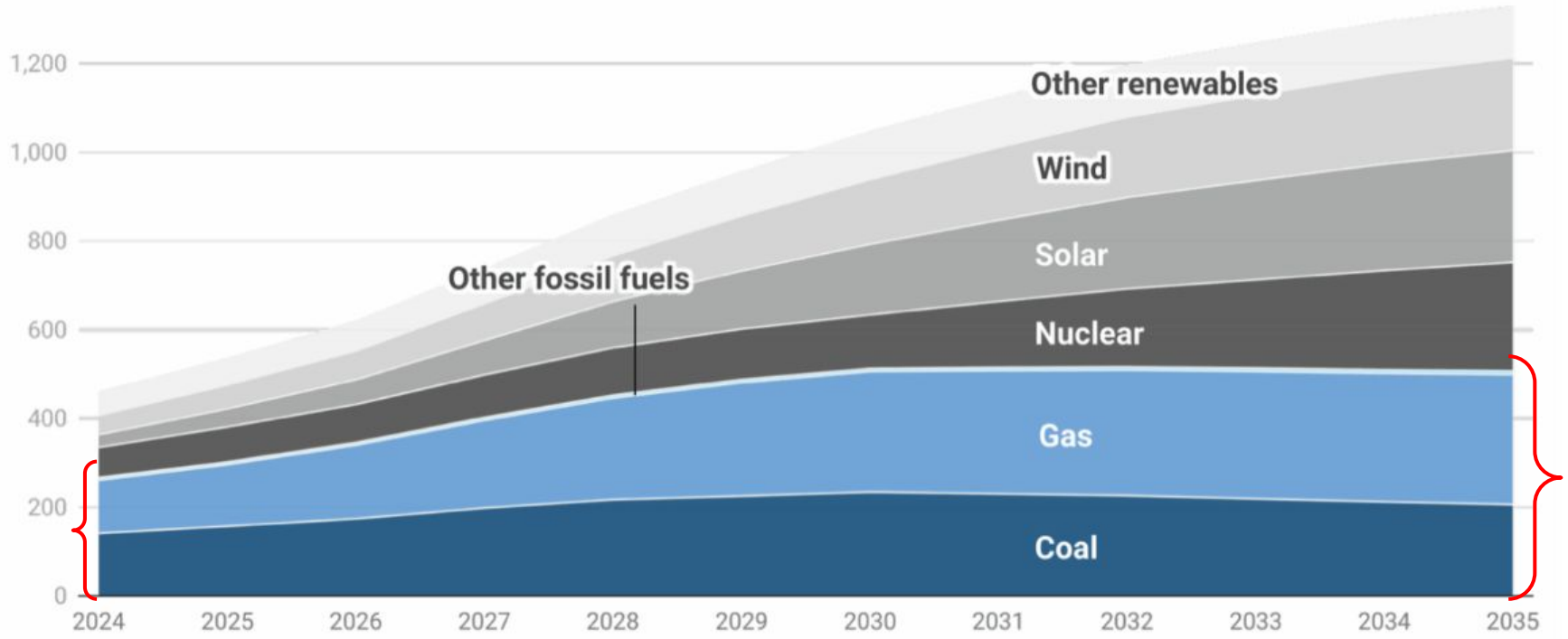
Microsoft's AI Buildout Sent Carbon Emissions Up 25% Last Year

By Ty Roush, Forbes Staff. Ty Roush is a breaking news reporter based in New ... Follow Author

Published Jul 09, 2026, 02:45pm EDT

Fossil fuels currently provide the majority of power to data centres – but this is expected to shift over the next decade

Sources of electricity for data centres, TWh, 2024-2035



Source: IEA energy and AI report.

CarbonBrief
CLEAR ON CLIMATE

For more on these topics, be sure to attend Nikola's lecture on **12 August**!



Question

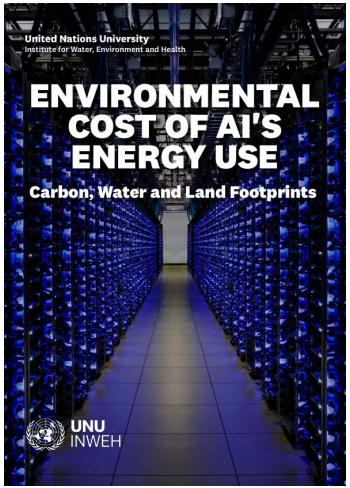
16% of countries host AI-specialised cloud compute.
How much of that is concentrated across China and the US?

Estimate a value between 0 and 100%

https://partici.fi/97320810

16% of countries host AI-specialised cloud compute.
How much of that is concentrated across China and the US?

90%



Political implications

Significant geopolitical vulnerability

Who is behind the companies?

Accessibility to commercial models

Weakening of democracy

Security implications





Validity crisis



Figure 1. Illustration of potemkin understanding in a large language model. This example shows GPT-4o's failure to apply its own conceptual explanation of an ABAB rhyming scheme.

Potemkin Understanding in Large Language Models

Marina Mancoridis [1] Keyon Vafa [2] Bec Weeks [3] Sendhil Mullainathan [1]

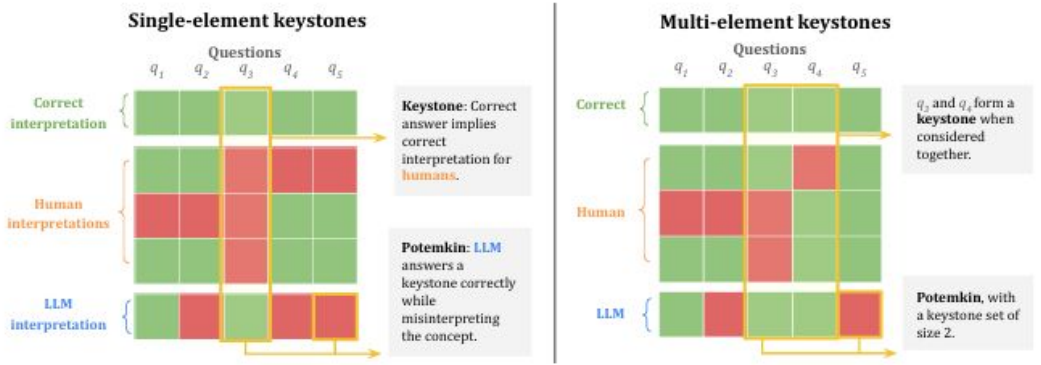


Figure 2. A schematic representation of keystones and potemkins. Rows represent interpretations of a concept and columns represent questions. Questions can either be interpreted correctly or incorrectly. A keystone is a set of questions that can only be interpreted correctly by a human who has understood the concept. An LLM has potemkin understanding when it correctly interprets all questions in a keystone but does not understand a concept.

Validity crisis

Potemkin Understanding in Large Language Models

Marina Mancoridis [1] Keyon Vafa [2] Bec Weeks [3] Sendhil Mullainathan [1]

	Potemkin rate, as measured by:		
Model	**Classify**	**Generate**	**Edit**
Llama-3.3	0.57 (0.06)	0.43 (0.09)	0.36 (0.05)
Claude-3.5	0.49 (0.05)	0.23 (0.08)	0.29 (0.04)
GPT-4o	0.53 (0.05)	0.38 (0.09)	0.35 (0.05)
Gemini-2.0	0.54 (0.05)	0.41 (0.09)	0.43 (0.05)
DeepSeek-V3	0.57 (0.05)	0.38 (0.09)	0.36 (0.05)
DeepSeek-R1	0.47 (0.05)	0.39 (0.09)	0.52 (0.05)
Qwen2-VL	0.66 (0.06)	0.62 (0.09)	0.52 (0.05)
Overall	**0.55 (0.02)**	**0.40 (0.03)**	**0.40 (0.02)**

Table 1. Potemkin rate on classify, generate, and edit tasks. The potemkin rate is the percent of questions about a concept that are answered incorrectly, conditional on a model defining that concept correctly. We scale the classification values so that chance-level performance corresponds with a potemkin rate of 1. Standard errors are in parentheses.

Validity crisis

Measuring what Matters: Construct Validity in Large Language Model Benchmarks

Andrew M. Bean[1*] Ryan Othniel Kearns[1] Angelika Romanou[2]
Franziska Sofia Hafner[1] Harry Mayne[1]

Jan Batzner[3,4] Negar Foroutan[2] Chris Schmitz[5] Karolina Korgul[1] Hunar Batra[1]
Oishi Deb[1] Emma Beharry[6] Cornelius Emde[1] Thomas Foster[1] Anna Gausen[7]
María Grandury[8,9] Simeng Han[10] Valentin Hofmann[11,12] Lujain Ibrahim[1]
Hazel Kim[1] Hannah Rose Kirk[1,7] Fangru Lin[1]
Gabrielle Kaili-May Liu[10] Lennart Luettgau[7] Jabez Magomere[1] Jonathan Rystrøm[1]
Anna Sotnikova[2] Yushi Yang[1] Yilun Zhao[10]

Adel Bibi[1] Antoine Bosselut[2] Ronald Clark[1] Arman Cohan[10] Jakob Foerster[1]
Yarin Gal[1,7] Scott A. Hale[1,13] Inioluwa Deborah Raji[14] Christopher Summerfield[1,7]
Philip H.S. Torr[1] Cozmin Ududec[7] Luc Rocher[1] Adam Mahdi[1*]

[1]University of Oxford [2]EPFL [3] Weizenbaum Institute Berlin
[4]Technical University Munich [5]Centre for Digital Governance, Hertie School
[6]Stanford University [7]UK AI Security Institute [8]SomosNLP
[9]Universdad Politécnica de Madrid [10]Yale University [11]Allen Institute for AI
[12]University of Washington [13]Meedan [14]UC Berkeley

Abstract

Evaluating large language models (LLMs) is crucial for both assessing their capabilities and identifying safety or robustness issues prior to deployment. Reliably measuring abstract and complex phenomena such as 'safety' and 'robustness' requires strong *construct validity*, that is, having measures that represent what matters to the phenomenon. With a team of 29 expert reviewers, we conduct a systematic review of 445 LLM benchmarks from leading conferences in natural language processing and machine learning. Across the reviewed articles, we find patterns related to the measured phenomena, tasks, and scoring metrics which undermine the validity of the resulting claims. To address these shortcomings, we provide eight key recommendations and detailed actionable guidance to researchers and practitioners in developing LLM benchmarks.

Validity crisis

RESEARCH-ARTICLE | OPEN ACCESS |

Gender Trouble in Language Models: An Empirical Audit Guided by Gender Performativity Theory

Authors: Franziska Sofia Hafner, Ana Valdivia, Luc Rocher | Authors Info & Claims

FAccT '25: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency • Pages 1677 - 1695
https://doi.org/10.1145/3715275.3732112

Published: 23 June 2025 Publication History

Analysis | Open access | Published: 12 December 2024

Generative language models exhibit social identity biases

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden & Jon Roozenbeek

Nature Computational Science **5**, 65–75 (2025) | Cite this article

Bias in Large Language Models: Origin, Evaluation, and Mitigation

Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, Shuo Shuo Liu

In research

- ➔ AI for survey responses?
  - Can LLMs reliably simulate human responses? *Current research tends towards **no***.
- ➔ AI for qualitative interviews?
  - New methods for scaling up qualitative interviews which tend to be time consuming and expensive. *But what of the process might we lose through this?*
- ➔ Risks of deskilling?
  - AI changing the way we work, what we know, and what we know how to do…what could this mean for the development of research and education?
- ➔ Case of agents replacing research assistants – what will that do to early career researchers?
  - *Same applies to entry-level jobs.*

What do you think?

Can you think of an application where we shouldn’t use AI?

https://partici.fi/97320810

What comes next?

Open questions

- ➔ Lots more to learn about how to mobilise people → *where can AI help?*
- ➔ Much more to understand about the pros and cons of how AI will impact various social and political challenges of the climate crisis.
- ➔ How we can shape the technology through the ways we use (or don't use) it?
- ➔ Climate change research, especially in the social sciences, is facing political headwinds…
  How can keep the social sciences part of the conversation? On climate change, on AI, and on their intersection?

Action points!

- ➔ Carefully audit the strengths and weaknesses of AI tools to best be able to use them responsibly.
  - ◆ Treat it the way you would any other tool or method.
  - ◆ Keep the focus on the ‘nail’ (objective) not the ‘hammer’ (tool).
- ➔ Ask questions of AI applications.
  - ◆ What’s the training data? What potential biases do I need to know about? How can I mitigate them?
  - ◆ Does the cost justify the performance gain?
  - ◆ Are open source models available? Could they do the job?
- ➔ Resist the trend to apply LLMs without validation – establish norms of including validation strategies and practices in every application.
- ➔ Think (and speak!) creatively and critically about how to optimise human - AI teamwork. Keep the human in the loop!

The social sciences form a critical foundation on which to build climate action towards decarbonisation.

- ➔ Diagnose barriers and identify solutions for establishing crucial social and political support across stakeholders → necessary for the uptake of technological advances.
- ➔ AI can accelerate parts of this process, but it is best used ***wisely*** **not** ***widely***.

Thank you for your attention.