University of Oxford: MPhil in Politics

The Political Effects of AI-Generated Content: Can AI Polarise Societies?

CESS Funding Application

Introduction to Research

Machine learning advancements to efficiently handle sequential data inputs and outputs have popularised the field of Artificial Intelligence (AI) (Vaswani *et al.*, 2017). AI is rapidly evolving into a transformative informational tool, with applications ranging from drug discovery to climate change modelling. Generative AI has emerged as the fastest-growing application, with tools like ChatGPT, Claude, and Midjourney gaining popularity through their ability to create sophisticated text, images, and video from simple prompts. Yet, these technological advancements are raising serious concerns from leading academics and AI developers alike. The 'Godfather of AI', Geoffrey Hinton, left Google over fears that safety and governance were are being overlooked in the pursuit of Artificial General Intelligence (AGI) (Metz, 2023). As AI systems develop the capability to set their own goals and operate autonomously, they present catastrophic risks through malicious actions, unsafe behaviour, or exploitation by bad actors (Hendrycks, Mazeika and Woodside, 2023). But, in the near-term, sub-catastrophic risks are equally present. In particular, this research project is interested in AI's capacity to 'amplify social injustice, erode social stability, [...] customised mass manipulation, and pervasive surveillance' (Bengio *et al.*, 2024). These social and political risks of AI are often discussed anecdotally, but there remains little research nor evidence on what these risks look like. The UK Government's Department for Science, Technology & Innovation (2025) views 'manipulation and deception of populations' a significant threat to political systems and societies; but, the extent to which politically targeted generative AI can be used to distort, deceive, and direct an electorate remains unclear. Therefore, this project aims to answer:

Does exposure to AI-generated political content increase affective polarisation?

This research seeks to address a pressing puzzle: why should we fear fake news or deceptive propaganda produced by generative AI more than that of earlier eras? Three factors stand out: the volume, realism, and micro-targeting of the content generated. Generative AI can confidently hallucinate political falsehoods and be directed to produce hyper-realistic, nearly undetectable 'fake news' (Flew, 2021; Duberry, 2022; Rawte *et al.*, 2023). With 45% of the US population reportedly using generative AI and social media providing fertile ground for virality, the likelihood of exposure to AI-generated misinformation is rising (Salesforce, 2025). However, democracies have long withstood misinformation campaigns, even before the advent of digital technologies (Bernays, 1928). So, are current fears about AI-driven manipulation justified (Ansell, 2023)? To approach this question, we must consider: can AI-generated content influence political attitudes, voting intentions, or even electoral outcomes? In particular, this research focuses on the critical dimension of affective polarisation to evaluate whether AI-generated content can exacerbate partisan hostility.

This focus is warranted. Fake news tends to spread rapidly in echo chambers, which are known to foster heightened animosity toward political out-groups (Törnberg, 2018; Hobolt, Lawall and Tilley, 2023). Since polarisation is closely tied to democratic backsliding and populist appeal, understanding AI's role in amplifying these dynamics is vital. Clarifying these effects holds significant implications for regulators, platforms, and policymakers. Moreover, this study considers the mechanisms of behavioural influence and the potential for mitigating interventions. One such intervention — labelling AI-generated content — is often seen as a straightforward solution. Yet, early evidence suggests that labelling may itself reinforce negative associations with fake news and deepen polarisation (Altay and Gilardi, 2024). Thus, this study treats labelling not only as an intervention but as an independent variable of interest.

To identify these effects, the research will combine survey experiments with an AI-augmented agent-based model that simulates repeated exposure scenarios. A pilot study conducted through the YouGov UniOM scheme has already offered preliminary evidence that unlabelled AI-generated content may be especially persuasive and polarising compared to a human-generated control. Additionally, when looking at the detection effect — labelled vs unlabelled AI-generated content — labelled content led to statistically significant levels of polarisation. This CESS funding proposal seeks to expand on this pilot to understand the mechanisms which explain why people are more likely to discount AI-generated content.

Research Design

This next phase of the research project proposes another online (survey) experiment, focussing on the mechanisms behind why people are more likely to discount political AI-generated content. The treatment structure from the pilot study is shown in Table 1.

Table 1: Treatment conditions by source and labelling

	Labelled (AI)	**Unlabelled**
Human	(not used)	(1) Control Group
AI	(2) Source Discount Condition	(3) Detection Condition

To isolate the role of source (AI) detection in moderating discounting, the detection effect given by (3) vs. (2) is of most interest. From formally modelling the relationship between the treatment and the outcome, the following hypotheses were proposed in my MPhil research proposal:[1]

[1]The formal modelling below is an extract from a complete model of expected treatment effects and is provided for reference.

(2) AI + Labelled — *Source Discount Condition*

- Participants are explicitly told the article is AI-generated.
- Belief responsiveness: $\mu_i = \beta_i \cdot w(\delta)$
- Direct awareness of AI authorship reduces trust and updating.
- Affective polarisation change is smaller relative to the control.

Treatment effect heterogeneity:

- Higher β_i: more similar to control group
- Lower β_i: minimal belief updating and polarisation change
- Higher education $\rightarrow$ likely higher β_i, attenuating the discount

Key comparison: (2) vs. (1) — **Source Credibility Effect**

(3) AI + Unlabelled — *Detection Condition*

- Participants are not told the source; belief about source depends on detection probability d_i.
- Responsiveness: $\mu_i = \bar{\beta}_i \cdot w(\delta)$, where $\bar{\beta}_i = d_i \cdot \beta_i + (1 - d_i) \cdot \beta^*$
- Affective polarisation depends on detection probability d_i and the relative size of β_i vs. β^*

Treatment effect heterogeneity:

- High d_i, low β_i: strong discounting $\rightarrow$ lower responsiveness
- Low d_i, high β^*: content treated as credible $\rightarrow$ stronger responsiveness
- High education $\rightarrow$ increases detection d_i and may raise both β_i and β^*, producing mixed effects

Key comparisons:

- (3) vs. (2) — **Detection Effect**

As a result of this formal modelling, the model identifies conditions under which AI-generated content can either increase, attenuate, or reduce affective polarisation. The key moderating mechanisms are:

- Ideological distance (δ) — the distance between the content and the individual's prior beliefs.
- Detection probability (d_i) — whether participants recognise the content is AI-generated.
- Trust in AI (β_i) — how much participants discount detected AI content.
- Persuasiveness of undetected AI content (β^*) — how influential undetected AI content is.
- Contrast sensitivity (ϕ) — affects how in-group evaluations respond to out-group belief changes.

- Initial affective attachments — strength of existing in-group and out-group feelings.

In particular, when looking at the detection effect, we are interested in how the detection probability d_i moderates the treatment effect as well as trust in AI β_i and the persuasiveness of undetected AI content β^*. The model predicts that higher detection probability d_i leads to lower responsiveness and polarisation change, while lower detection probability leads to stronger belief updating and polarisation change. As AI models improve, it can be expected that the detection probability d_i will decrease, and the persuasiveness of undetected AI content β^* will increase, leading to greater polarisation effects. Therefore, this research is most interested in the trust in AI (β_i).

This online survey experiment will provide participants with a series of AI-generated articles on a range of political topics, where the control is an unlabelled AI-generated article, and the treatment is a labelled AI-generated article. The articles will be designed to be realistic and persuasive, with the aim of mimicking the type of content that could be encountered on social media platforms. Participants will be asked to read the articles and then answer a series of questions about their perceptions of the source and its veracity. The experiment will test for heterogeneity in treatment effects across different demographic groups, including education level, political affiliation, and prior beliefs. It is theorised that the source (e.g., BBC, X, or The Guardian) will moderate the treatment effect, with participants more likely to discount AI-generated content from sources they do not trust such as social media. Moreover, the content of the articles will be varied to test how ideological distance (δ) moderates the treatment effect, as well as whether certain topics are more likely to elicit distrust in AI-generated content, for example if the content is about a controversial political issue such as immigration. The experimental design will be a simple control and treatment design given by Table 2:

Table 2: New treatment conditions by source and labelling

	Labelled (AI)	**Unlabelled**
	Treatment Group	Control Group
AI	(2) Source Discount Condition	(3) Detection Condition

The post-treatment questions will be designed to fit the Likert scale, with participants asked to rate their agreement with statement. To test detection probability d_i, participants will be asked to rate how likely they think the article is AI-generated. To test trust in AI β_i, participants will be asked to rate how much they trust the AI-generated content, and whether they would be more or less likely to trust the content if it were labelled as AI-generated. To test the persuasiveness of undetected AI content β^*, participants will be asked to rate how persuasive they found the content, and whether they would be more or less likely to share it on social media.

Hypotheses and Analysis

The primary theoretical prediction which was tested and shown in the pilot study is that participants exposed to unlabelled AI-generated political content (Detection Condition) exhibit greater affective polarisation than those exposed to labelled AI-generated content (Source Discount Condition). This next study aims to understand why. The following hypotheses are proposed based on the formal modelling and the pilot study findings:

- **Hypothesis 1:** Among participants in the Source Discount Condition, those with lower trust in AI (β_i) will exhibit weaker affective polarisation than those with higher trust.
- **Hypothesis 2:** In the Detection Condition, participants who find the content more persuasive (β^*) will show greater polarisation than those who do not.
- **Hypothesis 3:** Participants with higher levels of education will exhibit stronger detection (d_i) and trust calibration (β_i), moderating the treatment effect.
- **Hypothesis 4:** The magnitude of polarisation effects (in both conditions) will be greater when the content is ideologically distant from the participant's own beliefs (δ).
- **Hypothesis 5:** Content attributed to low-trust sources (e.g., X) will produce weaker polarisation effects compared to high-trust sources (e.g., BBC), especially in the Detection Condition.

From the set of questions asked in the post-treatment survey, these hypotheses will be tested using ordinal regression analysis. Tests for heterogeneity in treatment effects will be conducted by looking at the interaction effects between the treatment and demographic variables such as education level, political affiliation, and prior beliefs.

Budget Justification

The CESS funding will be used to cover the costs of running the online survey experiment, including participant recruitment and compensation. It is expected that the experiment will require approximately 1,000-2,000 participants, with a budget helping cover participant compensation. This number of participants is necessary to ensure sufficient statistical power to detect treatment effects and heterogeneity in treatment effects across different demographic groups. The funding will also be used to cover the costs of producing and running the online survey, including the costs of designing and hosting the survey. The total budget for the project is estimated to be approximately £1,000, with the CESS funding covering a significant portion of these costs.

References

Altay, S. and Gilardi, F. (2024) 'People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation', *PNAS Nexus*, 3(10), pp. 403–414.

Ansell, B. (2023) 'BBC Radio 4 - The Reith Lectures, Ben Ansell: Our Democratic Future', *BBC*. https://www.bbc.co.uk/programmes/m001t2r7.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y.N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A.G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., Dragan, A., Torr, P., Russell, S., Kahneman, D., Brauner, J. and Mindermann, S. (2024) 'Managing extreme AI risks amid rapid progress', *Science*, 384(6698), pp. 842–845.

Bernays, E.L. (1928) *Propaganda.* New York: H. Liveright.

Department for Science, Technology & Innovation (2025) 'Safety and security risks of generative artificial intelligence to 2025 (Annex B)', *GOV.UK*. https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper/safety-and-security-risks-of-generative-artificial-intelligence-to-2025-annex-b.

Duberry, J. (2022) 'AI and information dissemination: Challenging citizens access to relevant and reliable information', in *Artificial Intelligence and Democracy.* Cheltenham: Edward Elgar Publishing.

Flew, T. (2021) 'Fake news, trust, and behaviour in a digital world', in.

Hendrycks, D., Mazeika, M. and Woodside, T. (2023) 'An Overview of Catastrophic AI Risks'. arXiv.

Hobolt, S.B., Lawall, K. and Tilley, J. (2023) 'The Polarizing Effect of Partisan Echo Chambers', *American Political Science Review*, 118(3), pp. 1464–1479.

Metz, C. (2023) '"The Godfather of A.I." Leaves Google and Warns of Danger Ahead', *The New York Times* [Preprint].

Rawte, V., Chakraborty, S., Pathak, A., Sarkar, A., Tonmoy, T.I., Chadha, A., Sheth, A.P. and Das, A. (2023) 'The Troubling Emergence of Hallucination in Large Language Models - An Extensive Definition, Quantification, and Prescriptive Remediations'. arXiv.

Salesforce (2025) 'Top Generative AI Statistics for 2025', *Salesforce*.

Törnberg, P. (2018) 'Echo chambers and viral misinformation: Modeling fake news as complex contagion', *PLoS ONE*, 13(9), p. e0203958.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017) 'Attention Is All You Need', in *31st Conference on Neural Information Processing Systems.* Long Beach, CA, USA: arXiv.