Research Design Proposal

MPhil Politics: Comparative Government

Artificial Electorates: Simulating Effects of Exposure to Artificially-Generated Content on Affective Polarisation with Synthetic Agent-Based Modelling

Edward Anders

Supervisor: Professor Rachel Bernhard

St. Antony's College

University of Oxford

June 2025

Word Count: 5,992

Abstract

Advancements in machine learning have made Artificial Intelligence (AI) capable of generating hyper-realistic textual and visual content, accelerating its adoption as a powerful informational tool. Yet as generative AI technologies remain unregulated and vulnerable to misuse, concerns are growing about their potential to distort political messaging. This research investigates the polarising effects of AI-generated political content, focusing on how perceptions of trust and associations with misinformation influence affective polarisation among partisans. Using survey experiments with labelled and unlabelled AI- and human-generated content, the project tests whether source provenance moderates respondents' emotional and trait-based evaluations of political outgroups. A pilot study conducted with YouGov provides initial evidence that unlabelled AI content increases discomfort and reduces respect toward opposing partisans—supporting the view that undetected AI can act as a persuasive and polarising influence. To complement the experimental design, an agent-based model is developed to simulate repeated exposures and long-run dynamics. Together, these approaches offer new insights into the political consequences of AI being used in poltical communication.

Keywords: artificial intelligence, affective polarisation, fake news, survey experiment, agent-based modelling

Contents

List of Tables

List of Figures

1 Introduction

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 combines survey experiments with an AI-augmented agent-based model that simulates repeated exposure scenarios. It begins by reviewing the literature on AI and affective polarisation, followed by a theoretical framework and set of hypotheses. A pilot study conducted with YouGov is then presented, offering preliminary evidence that unlabelled AI-generated content may be especially persuasive and polarising. Finally, an innovative use of synthetic agents is proposed to further explore AI's influence on political attitudes.

> *Note:* Given the word count constraints, the literature review has been moved to the appendix. I do not expect any feedback on this section. The primary focus of this proposal is the theoretical framework, the pilot study, and the next steps. These are the sections I would most appreciate feedback on.

2 Theoretical Framework

As described in Section 4.1 above, the literature on the effects of AI-generated content is still nascent. Developing a theoretical framework to understand effects of exposure therefore requires a number of assumptions and leaning on theories of fake news and affective polarisation. Of particular guidance are formal models of the spread of misinformation within networks, namely those by Acemoglu, Ozdaglar and Siderius (2024), Della Lena (2024), and Jones, Pauls and Fu (2024). The formal theory developed in this section takes these models and applies them to the models and hypotheses used in the affective polarisation literature from Törnberg *et al.* (2021) and Hobolt, Lawall and Tilley (2023).

The model presented is motivated by these aforementioned models, and uses a simplified Bayesian-inspired updating set up for modelling a utility function response to AI-generated political information. While classical Bayesian updating requires agents to form posterior beliefs using formally specified likelihood functions, a simplified, quasi-Bayesian updating framework is used for three reasons:

(i) **Empirical Tractability:** Full Bayesian inference requires assumptions about prior distributions and signal noise that are unobservable in survey settings.

(ii) **Psychological Plausibility:** Individuals often rely on heuristics when processing political information, especially under uncertainty about source credibility.

(iii) **Interpretative Clarity:** The simplified rule permits direct mapping between theoretical parameters (e.g., trust in AI, ideological distance) and experimental treatment conditions.

2.1 Model Setup

Let the individual's belief about the ideological position of the outgroup be denoted by:

- θ_0: prior belief
- θ_1: posterior belief
- C: ideological content of the article
- $\delta = |C - \theta_0|$: ideological distance between article content and prior
- $S \in \{\text{AI}, \text{Human}\}$: true source of the article
- $\hat{S}$: perceived source
- $\beta_i \in [0, \infty)$: responsiveness to detected AI-generated content (higher values indicate greater persuasiveness)
- $\beta^* \in [1, \infty)$: responsiveness to undetected AI-generated content (relative to baseline human content)
- $d_i \in [0, 1]$: probability individual i detects the true source

The individual updates their belief according to Equation 1:

$$\theta_1 = \theta_0 + \bar{\beta}_i \cdot w(\delta) \cdot (C - \theta_0) \tag{1}$$

where:

$$\bar{\beta}_i = d_i \cdot \beta_i + (1 - d_i) \cdot \beta^* \quad \text{and} \quad w(\delta) = \frac{1}{1 + \lambda \cdot \delta}, \quad \lambda > 0$$

The term $\bar{\beta}_i$ reflects the expected responsiveness to the article, based on the detection probability and whether the content is believed to be AI-generated. When detected, responsiveness is governed by β_i, which may reflect either discounting ($\beta_i < 1$) or amplification ($\beta_i > 1$). When undetected, the content is processed with responsiveness $\beta^* \geq 1$, which allows for the possibility that AI content is more persuasive than human content even when its origin is unknown, as suggested by Goldstein *et al.* (2024).

The function $w(\delta)$ reflects ideological receptiveness, with greater distance reducing responsiveness.

2.2 Detection and Responsiveness

As has been widely reported, trust in AI-generated content is often low. When individuals are aware that content is AI-generated, they may be less likely to trust it (Afroogh *et al.*, 2024). However, this is not universally the case: for some individuals, particularly those who view algorithmic content as high-quality or ideologically aligned, detected AI content may be treated as even more persuasive than human-generated content. The model captures this by allowing responsiveness to detected AI content to exceed 1 (i.e., $\beta_i > 1$).

Importantly, the model also accounts for the possibility that undetected AI content may be more persuasive than human-generated content. In this case, the individual does not consciously adjust their beliefs based on the source, but may nonetheless respond more strongly than they would to equivalent human-written information. This is captured by allowing the baseline responsiveness to undetected AI content to be $\beta^* \geq 1$.

Therefore, the expected responsiveness to content is a convex combination of two components — detection probability and responsiveness to detected AI content — as shown in Equation 2:

$$\bar{\beta}_i = d_i \cdot \beta_i + (1 - d_i) \cdot \beta^* \tag{2}$$

Detection probability (d_i) depends on observable individual characteristics, such as education and political attention, as shown in Equation 11:

$$\frac{\partial d_i}{\partial \text{Education}_i} > 0, \quad \frac{\partial d_i}{\partial \text{Political Attention}_i} > 0 \tag{3}$$

Education increases individuals' ability to detect linguistic and structural cues of AI authorship. Political attention increases motivation to scrutinise political content, enhancing vigilance in identifying the source even in the absence of explicit labelling (Chein, Martinez and Barone, 2024).

Responsiveness to detected AI content is captured by the parameter $\beta_i \in [0, \infty)$, which reflects how much individuals update their beliefs when they are aware the content is AI-generated. Individuals with greater familiarity with or trust in AI systems may respond more strongly to detected AI content, as reflected in Equation 4:

$$\frac{\partial \beta_i}{\partial \text{Education}_i} > 0 \tag{4}$$

In contrast, $\beta^* \geq 1$ represents the baseline responsiveness to AI content that is not detected as such. This term allows for the possibility that undetected AI content may be inherently more persuasive — for example, due to improved stylistic coherence, ideological tailoring, or perceived neutrality.

Notably, education and political attention may affect both detection and responsiveness components, creating a theoretically rich asymmetry: more educated individuals are more likely to detect AI content ($d_i \uparrow$), and may apply a smaller penalty or even an amplification factor when doing so ($\beta_i \uparrow$). But even if they do not detect the AI origin, they may still respond more strongly than to human content, due to $\beta^* \geq 1$.

2.3 Maximisation Problem

We assume individuals seek to minimise epistemic loss, defined as the squared distance between their updated belief and the article content. This is captured by the utility function:

$$u(\theta_1, C) = -(\theta_1 - C)^2 \tag{5}$$

Individuals choose a responsiveness parameter $\mu_i \in [0, \infty)$ such that their updated belief is given by:

$$\theta_1 = \theta_0 + \mu_i(C - \theta_0) \tag{6}$$

The individual's optimisation problem becomes:

$$\max_{\mu_i \in [0,\infty)} -\left[(1 - \mu_i)(\theta_0 - C)\right]^2 \tag{7}$$

The utility in Equation 5 is maximised when $\mu_i = 1$, corresponding to full alignment between updated beliefs and the article content. However, individuals do not always fully trust the content, and their responsiveness

is shaped by both ideological distance and beliefs about the credibility of the source.

We therefore assume that responsiveness is endogenously constrained by detection and perceived source credibility:

$$\mu_i = \bar{\beta}_i \cdot w(\delta) \tag{8}$$

Because $\bar{\beta}_i \in [0, \infty)$, the model allows for cases where individuals update less than, equally to, or more than the signal direction (i.e., $\mu_i < 1$, $= 1$, or > 1), depending on how persuasive they find the content.

2.4 Treatment Conditions

As noted, labelling content as AI-generated may affect the perceived source and trust in the information. The model therefore considers two treatment arms based on the combination of true source (S) and whether the article is labelled:

1. AI + Labelled

- $\hat{S} = \text{AI}$
- $\bar{\beta}_i = \beta_i$
- $\mu_i = \beta_i \cdot w(\delta)$

2. AI + Unlabelled

- $\bar{\beta}_i = d_i \cdot \beta_i + (1 - d_i) \cdot \beta^*$
- $\mu_i = \bar{\beta}_i \cdot w(\delta)$

These treatment conditions are in reference to the control group where the article is human-generated and not labelled:

Human + Unlabelled

- $\bar{\beta}_i = 1$ (assumed human by default)
- $\mu_i = w(\delta)$

2.5 Comparative Statics

We now examine how the responsiveness parameter μ_i varies with respect to key exogenous parameters in the model. This comparative statics analysis focuses on understanding how belief updating is shaped by source detection, ideological distance, and trust in AI.

2.5.1 Exogenous Parameters

Table 1 below lists the key exogenous parameters in the model. These are a subset of a likely much longer list of possible parameters, but these are the most relevant and prominent factors:

Table 1: Key Parameters Used in the Theoretical Model

Parameter	**Description**	**Type**
C	Ideological content of the article	Experimental treatment
θ_0	Individual's prior belief	Observed (pre-treatment)
$S \in \{\text{AI}, \text{Human}\}$	True source of article	Experimental treatment
Label	Whether the source is labelled	Experimental treatment
$\delta = \lvert C - \theta_0 \rvert$	Ideological distance	Derived (individual-level)
β_i	Discount factor (AI trust)	Observed/inferred
d_i	Probability of detecting AI source	Derived
β^*	Responsiveness to undetected AI content	Model parameter
λ	Responsiveness decay parameter	Model parameter

Recall that responsiveness is defined in Equation 9:

$$\mu_i = \bar{\beta}_i \cdot w(\delta) = \bar{\beta}_i \cdot \frac{1}{1 + \lambda \cdot \delta} \tag{9}$$

where:

$$\bar{\beta}_i = \begin{cases} 1 & \text{if } S = \text{Human or perceived as Human} \\ \beta_i & \text{if } S = \text{AI and detected as such (i.e., labelled)} \\ d_i \cdot \beta_i + (1 - d_i) \cdot \beta^* & \text{if } S = \text{AI and unlabelled} \end{cases}$$

We now derive the relevant partial derivatives, one parameter at a time.

2.5.2 Ideological Distance δ

The responsiveness parameter μ_i is inversely related to ideological distance δ. As the ideological distance between the article content and the individual's prior belief increases, the responsiveness decreases due to the diminishing returns of ideological receptiveness. This effect is more pronounced when source credibility is high (i.e., when $\bar{\beta}_i$ is large). This is captured by the derivative in Equation 10:

$$\frac{\partial \mu_i}{\partial \delta} = \bar{\beta}_i \cdot \frac{\partial w(\delta)}{\partial \delta} = \bar{\beta}_i \cdot \left(\frac{-\lambda}{(1 + \lambda \cdot \delta)^2} \right) < 0 \tag{10}$$

2.5.3 Source Detection Probability d_i

In the cases where $S = \text{AI}$ and the content is unlabelled, then the detection probability affects the responsiveness parameter μ_i through the discount factor $\bar{\beta}_i$. The derivative in Equation 11 captures this relationship:

$$\bar{\beta}_i = d_i \cdot \beta_i + (1 - d_i) \cdot \beta^* \quad \Rightarrow \quad \frac{\partial \mu_i}{\partial d_i} = (\beta_i - \beta^*) \cdot w(\delta) \tag{11}$$

When individuals become more likely to detect that content is AI-generated, they apply a different responsiveness depending on their trust in AI. If $\beta^* > \beta_i$, then detection reduces responsiveness, as detected content is trusted less than undetected content. However, if $\beta_i > \beta^*$, detection increases responsiveness. This captures the idea that more sophisticated individuals — while better at detecting AI — may also treat detected content differently depending on their predispositions.

2.5.4 Discount Factor β_i

For AI-generated articles (labelled or unlabelled), individuals apply a discount factor β_i to the content based on their trust in AI. Individuals who are more trusting of AI-generated content (higher β_i) update their beliefs more strongly in response to such content. The effect is larger when the source is detected (higher d_i) shown in both Equation 12 and Equation 13:

- **Labelled AI:**

$$\mu_i = \beta_i \cdot w(\delta) \quad \Rightarrow \quad \frac{\partial \mu_i}{\partial \beta_i} = w(\delta) > 0 \tag{12}$$

- **Unlabelled AI:**

$$\bar{\beta}_i = d_i \cdot \beta_i + (1 - d_i) \cdot \beta^* \quad \Rightarrow \quad \frac{\partial \mu_i}{\partial \beta_i} = d_i \cdot w(\delta) > 0 \tag{13}$$

2.5.5 Responsiveness to Undetected AI Content β^*

In the case of unlabelled AI content, responsiveness also depends on the baseline credibility of content that is not recognised as AI-generated. This is captured by the parameter β^*, which enters the convex combination that defines $\bar{\beta}_i$ when the source is AI and unlabelled:

$$\bar{\beta}_i = d_i \cdot \beta_i + (1 - d_i) \cdot \beta^*$$

The partial effect of β^* on the responsiveness parameter μ_i is given by Equation 14:

$$\frac{\partial \mu_i}{\partial \beta^*} = (1 - d_i) \cdot w(\delta) > 0 \tag{14}$$

This derivative reflects the idea that responsiveness increases with β^* when AI content goes undetected. In other words, the more persuasive undetected AI content is (higher β^*), the more the individual updates their beliefs in response to it. This effect is stronger when detection probability d_i is low.

2.5.6 Responsiveness Decay Parameter λ

This parameter is a theoretical parameter which assumes that an individual's responsiveness to ideological content decays as the ideological distance increases. This is captured by the derivative in Equation 15:

$$\frac{\partial \mu_i}{\partial \lambda} = \bar{\beta}_i \cdot \frac{\partial w(\delta)}{\partial \lambda} = \bar{\beta}_i \cdot \left(\frac{-\delta}{(1 + \lambda \cdot \delta)^2} \right) < 0 \tag{15}$$

Higher values of λ imply sharper declines in responsiveness with ideological distance. A higher λ implies more resistance to persuasion at larger ideological distances, but this effect flattens out as the distance increases. This parameter governs how ideologically resistant individuals are in general. It could be treated as a theoretical parameter or estimated at the population level.

2.5.7 Summary of Comparative Statics

Table 2: Summary of Comparative Statics for each Parameter

Parameter	Partial Derivative	Sign	Interpretation
δ	$\frac{\partial \mu_i}{\partial \delta}$	Negative	Greater ideological distance reduces responsiveness
d_i (AI + unlabelled)	$\frac{\partial \mu_i}{\partial d_i}$	Negative	Detection increases discounting and reduces updating
β_i (AI only)	$\frac{\partial \mu_i}{\partial \beta_i}$	Positive	More trust in AI increases responsiveness
β^* (AI + unlabelled)	$\frac{\partial \mu_i}{\partial \beta^*}$	Positive	Greater persuasiveness of undetected AI content increases responsiveness
λ	$\frac{\partial \mu_i}{\partial \lambda}$	Negative	More rigidity reduces responsiveness across the board

2.6 Belief Updating to Affective Polarisation

The model presented above provides a theoretical framework for understanding how individuals update their beliefs in response to AI-generated political content. The key parameters and comparative statics highlight the complex interplay between ideological distance, source detection, trust in AI, and responsiveness to content. Affective polarisation is defined as the difference between in- and out-group evaluations, shown in Equation 16:

$$\text{AP}_i = L_i^{\text{in}} - L_i^{\text{out}} \tag{16}$$

where:

- L_i^{in}: affective evaluation of the in-group,
- L_i^{out}: affective evaluation of the out-group.

We are interested in how the treatment-induced belief change $\Delta\theta_i = \theta_1 - \theta_0 = \mu_i(C - \theta_0)$, where μ_i incorporates detection and source responsiveness, affects the change in affective polarisation:

$$\Delta\text{AP}_i = \Delta L_i^{\text{in}} - \Delta L_i^{\text{out}} \tag{17}$$

2.6.1 Asymmetric Malleability of Attitudes

The malleability of affective evaluations is not symmetric. Lee *et al.* (2022) find that most people are positive partisans, meaning they identify with a party because they like their side, rather than opposing the other side. Greater in-group identification — or particularly strong animosity toward the out-group — suggests that the malleability of affective evaluations declines with stronger initial feelings. This implies diminishing marginal returns to new information: individuals who already feel very positively or negatively about a group are less likely to change their attitudes. This is formalised in Equation 18 and Equation 19:

$$\Delta L_i^{\text{out}} = \frac{1}{|L_i^{\text{out}}| + \varepsilon} \cdot \Delta\theta_i \tag{18}$$

$$\Delta L_i^{\text{in}} = \frac{1}{|L_i^{\text{in}}| + \varepsilon} \cdot f(\Delta\theta_i) \tag{19}$$

where:

- $\varepsilon > 0$ ensures continuity at zero affect,
- $f(\cdot)$ is a scaling function determining how belief updates about the out-group influence in-group feelings,
- If $f(\cdot) = -\phi \cdot \Delta\theta_i$, with $\phi \geq 0$, then belief improvements about the out-group reduce in-group warmth due to contrast or identity differentiation. Here, ϕ captures the extent to which belief updates favouring the out-group lead to reductions in in-group warmth.

Substituting into the expression for ΔAP_i, we obtain the final expression for the change in affective polarisation in Equation 20:

$$\Delta\text{AP}_i = \left(\frac{-\phi}{|L_i^{\text{in}}| + \varepsilon} - \frac{1}{|L_i^{\text{out}}| + \varepsilon} \right) \cdot \Delta\theta_i \tag{20}$$

2.6.2 Interpretation of Affective Polarisation Change

- **Direction of change:** If $\Delta\theta_i > 0$ (i.e., the individual updates in a more favourable direction toward the out-group), then affective polarisation may decrease or increase depending on which affect is more malleable.
- **Attitude strength asymmetry:** The more entrenched an individual's dislike of the out-group, the less likely that attitude is to change. In such cases, belief change is more likely to influence in-group evaluations, potentially increasing polarisation if $\phi > 0$.
- **Symmetry:** If in-group and out-group attitudes are of similar strength, then affective polarisation is more likely to respond symmetrically to belief change.
- **Contrast effect:** When $\phi > 0$, positive updates about the out-group may reduce in-group warmth (e.g., due to identity threat or cognitive balancing), further decreasing polarisation.

This framework allows us to capture heterogeneity in the direction and magnitude of affective polarisation change as a function of both belief updating and initial affective attachments.

2.7 Affective Polarisation Comparative Statics

We can now derive how changes in the model's exogenous parameters affect the change in affective polarisation ΔAP_i. As derived in Equation 20, and restated below, we can define the responsiveness of affective polarisation to a belief change A_i, and μ_i in Equation 21 to give a cleaner expression for the change in affective polarisation in Equation 22:

$$\Delta\text{AP}_i = \left(\frac{-\phi}{|L_i^{\text{in}}| + \varepsilon} - \frac{1}{|L_i^{\text{out}}| + \varepsilon} \right) \cdot \Delta\theta_i \quad \text{where} \quad \Delta\theta_i = \mu_i \cdot (C - \theta_0)$$

Letting:

$$A_i = \left(\frac{-\phi}{|L_i^{\text{in}}| + \varepsilon} - \frac{1}{|L_i^{\text{out}}| + \varepsilon} \right) \quad \text{and} \quad \mu_i = (d_i \cdot \beta_i + (1 - d_i) \cdot \beta^*) \cdot \frac{1}{1 + \lambda \cdot \delta} \tag{21}$$

we can write:

$$\Delta\text{AP}_i = A_i \cdot \mu_i \cdot (C - \theta_0) \tag{22}$$

2.7.1 Ideological Distance δ

As the ideological distance between the article content and the individual's prior belief increases, the individual's responsiveness declines, reducing belief updating and therefore the effect on affective polarisation.

This is formalised in Equation 23:

$$\frac{\partial \Delta \text{AP}_i}{\partial \delta} = A_i \cdot \frac{\partial \mu_i}{\partial \delta} \cdot (C - \theta_0) = A_i \cdot (d_i \cdot \beta_i + (1 - d_i) \cdot \beta^*) \cdot \left(\frac{-\lambda}{(1 + \lambda \cdot \delta)^2} \right) \cdot (C - \theta_0) < 0 \tag{23}$$

2.7.2 Detection Probability d_i

This condition is relevant in the scenario where the article is AI-generated and unlabelled. In this case, more accurate detection of AI content increases the likelihood of discounting it, reducing belief updating and attenuating affective response, given by Equation 24:

$$\frac{\partial \Delta \text{AP}_i}{\partial d_i} = A_i \cdot \frac{\partial \mu_i}{\partial d_i} \cdot (C - \theta_0) = A_i \cdot (\beta_i - \beta^*) \cdot \frac{1}{1 + \lambda \cdot \delta} \cdot (C - \theta_0) < 0 \quad \text{if } \beta_i < \beta^* \tag{24}$$

As shown in Section 2.2, d_i can be thought of as a function of individual characteristics such as education and political attention. Therefore, as education and political attention increase, the detection probability increases, leading to a decrease in affective polarisation if $\beta_i < \beta^*$.

2.7.3 Discount Factor β_i

The discount factor β_i captures the individual's trust in AI-generated content when it is detected. Higher trust leads to greater belief updating and potentially greater affective polarisation. In the case of labelled AI content, the source is explicitly known and β_i governs responsiveness directly. For unlabelled AI content, β_i only affects affective polarisation to the extent that the content is detected (i.e., $d_i > 0$); otherwise, belief updating is governed by β^*. This is formalised in Equation 25 and Equation 26:

- **Labelled AI:**

$$\frac{\partial \Delta \text{AP}_i}{\partial \beta_i} = A_i \cdot w(\delta) \cdot (C - \theta_0) > 0 \tag{25}$$

- **Unlabelled AI:**

$$\frac{\partial \Delta \text{AP}_i}{\partial \beta_i} = A_i \cdot d_i \cdot w(\delta) \cdot (C - \theta_0) > 0 \tag{26}$$

2.7.4 Responsiveness to Undetected AI Content β^*

The parameter β^* governs how strongly individuals respond to AI-generated content that is not detected as AI This term enters into the responsiveness expression μ_i, and therefore affects belief updating and affective polarisation in unlabelled AI conditions. The partial derivative of affective polarisation with respect to β^* is given by Equation 27:

$$\frac{\partial \Delta \text{AP}_i}{\partial \beta^*} = A_i \cdot (1 - d_i) \cdot w(\delta) \cdot (C - \theta_0) \tag{27}$$

This effect is strictly positive, indicating that increases in the persuasiveness of undetected AI content β^*, increase the change in affective polarisation. This effect is strongest when the detection probability d_i is low, and declines as detection increases. This implies that as AI content becomes more realistic, it has greater potential to persuade individuals and affect their in-group and out-group evaluations.

2.7.5 Contrast Parameter ϕ

Higher contrast sensitivity implies that more positive beliefs about the out-group reduce in-group warmth, thus reducing affective polarisation more strongly, given by Equation 28:

$$\frac{\partial \Delta \text{AP}_i}{\partial \phi} = \frac{-1}{|L_i^{\text{in}}| + \varepsilon} \cdot \mu_i \cdot (C - \theta_0) \tag{28}$$

2.7.6 Initial Affective Attachments

Stronger in-group warmth reduces in-group responsiveness, shifting weight to the out-group channel. Stronger out-group hostility makes it harder to reduce polarisation via changing out-group attitudes. For in-group warmth, this is captured by Equation 29 and for out-group hostility by Equation 30:

$$\frac{\partial A_i}{\partial |L_i^{\text{in}}|} = \frac{\phi}{(|L_i^{\text{in}}| + \varepsilon)^2} > 0 \quad \Rightarrow \quad \frac{\partial \Delta \text{AP}_i}{\partial |L_i^{\text{in}}|} > 0 \text{ if } \Delta\theta_i > 0 \tag{29}$$

$$\frac{\partial A_i}{\partial |L_i^{\text{out}}|} = \frac{1}{(|L_i^{\text{out}}| + \varepsilon)^2} > 0 \quad \Rightarrow \quad \frac{\partial \Delta \text{AP}_i}{\partial |L_i^{\text{out}}|} < 0 \text{ if } \Delta\theta_i > 0 \tag{30}$$

2.7.7 Summary of Affective Polarisation Comparative Statics

The comparative statics for the change in affective polarisation ΔAP_i are summarised in Table 3 below:

Table 3: Comparative statics of affective polarisation: summary of partial effects

Parameter	Partial Derivative	Sign	Interpretation
δ	$\frac{\partial \Delta \text{AP}_i}{\partial \delta}$	Negative	Greater ideological distance reduces belief updating
d_i	$\frac{\partial \Delta \text{AP}_i}{\partial d_i}$	Negative	Detection reduces responsiveness to AI content
β_i	$\frac{\partial \Delta \text{AP}_i}{\partial \beta_i}$	Positive	More trust in AI increases responsiveness
β^*	$\frac{\partial \Delta \text{AP}_i}{\partial \beta^*}$	Positive	More persuasive undetected AI content increases affective polarisation
ϕ	$\frac{\partial \Delta \text{AP}_i}{\partial \phi}$	Negative	In-group contrast reduces affective polarisation
$\lvert L_i^{\text{in}} \rvert$	$\frac{\partial \Delta \text{AP}_i}{\partial \lvert L_i^{\text{in}} \rvert}$	Positive	In-group affect less malleable → greater weight on out-group
$\lvert L_i^{\text{out}} \rvert$	$\frac{\partial \Delta \text{AP}_i}{\partial \lvert L_i^{\text{out}} \rvert}$	Negative	Strong out-group dislike reduces scope for affective change

2.8 Hypotheses

From this formal model of belief updating and affective polarisation, several testable hypotheses regarding the effects of AI-generated content on individuals' affective evaluations of in- and out-groups can be derived.

We now map the theoretical model onto the experimental design, which comprises one control group and two treatment conditions. The design is defined by whether the article is AI-generated and whether it is labelled. Articles not labelled are assumed to be human-generated by default, consistent with participants' likely priors in naturalistic settings.

2.8.1 Treatment Structure

Table 4: Treatment conditions by source and labelling

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

2.8.2 Theoretical Predictions and Heterogeneity

The primary theoretical prediction, and hypothesis this research project aims to test is:

> **Hypothesis 1:** Exposure to AI-generated political content *can* increase affective polarisation, particularly when the AI origin is not detected and the content is ideologically aligned with the individual's priors.

(1) Human + Unlabelled — *Control Group*

- Participants are expected to assume the article is human-generated.
- Belief responsiveness is high: $\mu_i = w(\delta)$
- No discounting is applied, and content is assumed credible.
- Affective polarisation change depends on the size of $\Delta\theta_i$ and affective malleability.

Treatment effect heterogeneity:

- Higher ideological distance $\delta \rightarrow$ lower responsiveness
- Stronger affective priors $\rightarrow$ reduced attitude change

(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 → 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 → lower responsiveness
- Low d_i, high β^*: content treated as credible → stronger responsiveness
- High education → increases detection d_i and may raise both β_i and β^*, producing mixed effects

Key comparisons:

- (3) vs. (2) — **Detection Effect**
- (3) vs. (1) — **Combined Discounting and Detection Effect**

2.8.3 Summary of Theoretical Comparisons

Table 5: Interpretation of comparisons between treatment conditions

Comparison	**Name**	**Interpretation**
(2) vs. (1)	Source Credibility Effect	Trust penalty for labelled AI content
(3) vs. (2)	Detection Effect	Role of source detection in moderating discounting
(3) vs. (1)	Combined Discounting and Detection	Total effect of AI content when not labelled

As a result of this formal modelling, the model dentifies 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.

3 Methodological Approaches

This research seeks to estimate the direction and magnitude of causal effects. To test the hypotheses set out in Section 2.8, a survey experiment is used to isolate the effect of AI-generated content on affective polarisation. In addition, this proposal introduces an innovative extension: the use of agent-based modelling, enhanced by Large Language Models (LLMs), to simulate repeated exposures and complement the survey findings. This section outlines the survey design, presents initial pilot results, and discusses how agent-based methods could further develop the analysis.

As discussed in the literature review (Section 4.1), this project focuses on the United Kingdom. While most research on affective polarisation centres on the United States, recent evidence shows that UK partisans also exhibit negative emotions toward out-groups (Berntzen, Kelsall and Harteveld, 2024). Given the rise of populism and the increasingly polarised political climate in the UK, it is critical to assess whether new technologies, such as generative AI, pose risks to democratic processes—concerns raised during the 2024 General Election (Simon, McBride and Altay, 2024). Therefore, the following research design is tailored to the UK context, with the potential for future studies to assess external validity in other settings.

3.1 Pilot Study: YouGov UniOM Survey Experiment

To test whether effects are observable, and within which groups, a pilot study was conducted using a single survey experiment featuring exposure to an AI-generated article. The study was run with YouGov's UniOM panel, a nationally representative sample of the UK population. The experiment employed two treatment conditions: AI-generated (unlabelled) and AI-labelled content. A between-subjects design was used to avoid pre-survey sensitivity issues identified by Levendusky and Stecula (2021).

The primary outcomes of interest are measures of affective polarisation, as discussed in Section 3.1.1. Both treatment groups are compared against a control group exposed to human-generated content.[1] The treatment conditions are summarised in Table 6.

Table 6: Treatments and Control for AI-generated Content Exposure

	No Labels	**Labelled as AI**
Control	Human-Generated Article	Human-Generated Article
Treatment	AI-generated Article	AI-generated Article

[1] A third treatment condition of human-generated content labelled as AI-generated is also tested. This condition can be used to help understand whether participents discount based on content or whether they discount based on the source (i.e., AI vs. human). This additional analysis is not included in this proposal, but may be included in the final thesis.

3.1.1 Experimental Variables

The survey experiment measures the effect of AI-generated content on affective polarisation, using both emotional and behavioural indicators. Respondents are randomly assigned to view either human-written, AI-generated, or AI-labelled news articles on a divisive topic—immigration in the UK. The treatment content is designed to provoke strong emotional responses and test whether such content can shift partisan attitudes. Full versions of the treatment articles are included in the Appendix (Section 4.4).

Affective polarisation is assessed using standard measures from the literature. These include feeling thermometer scores towards in- and out-party leaders (used to compute a net difference, `thermo_gap`), as well as trait-based evaluations of respect and trust in out-parties. A behavioural proxy for polarisation is also included, asking respondents how they would feel if their child married a supporter of their least preferred party. These variables are discussed in full, along with question wordings and coding details, in the Appendix (Section 4.2 and Section 4.3).

3.1.2 Regression Specification

To test the causal Average Treatment Effect (ATE) of respondents being exposed to AI-generated and AI-lebelled content on the set of affective polarisation measures, a series of regression models are estimated. The model specification is given by Equation 31:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 \mathbf{X}_i + \beta_3 (D_i \times \mathbf{Z}_i) + \varepsilon_i \tag{31}$$

where:

- Y_i: outcome variable (`thermo_gap`, `MLthermoMean`, `LLthermoMean`, `agreedisagree`, `xtrust`, `child`)
- D_i: treatment indicator (`AI-Generated Content, Labelled or Unlabelled`)
- $\mathbf{X}_i$: vector of covariates (see Balance Check in Section 4.6)
- $\mathbf{Z}_i$: vector of interaction terms (treatment $\times$ moderator)
- ε_i: stochastic error term

In this specification, β_1 estimates the average treatment effect when the moderator(s) are at their reference level. Estimates are calculated with survey-weighted least squares and ordinal logistic models so results can be generalised to the UK more broadly. β_2 measures the effect of a one-unit change of a covariate on the outcome variable. β_3 captures the treatment effect heterogeneity across different sub-groups of the moderator, where statistically significant non-zero values suggest the ATE is different for different sub-group characteristics.

3.1.3 Preliminary Results

The results are presented in two parts. First, the continuous thermometer measures of affective polarisation are reported, followed by the ordinal emotional and trait-based outcomes. Each is evaluated across the three theoretical treatment conditions defined in Section 2.8: (1) Source Credibility Effect, (2) Detection Effect, and (3) Combined Discounting and Detection Effect.

Regression tables include three models. Model (1) provides a baseline without covariates. Model (2) includes all pre-treatment covariates, justified both by theory and to improve efficiency, despite the balance check confirming covariate balance across treatment groups (Section 4.6). Model (3) introduces interaction terms with likely moderators—education, political attention, and ideological distance—as theorised in Section 2. All models apply survey weights to ensure representativeness of the UK population.

The thermometer score results suggest that unlabelled AI content slightly increases affective polarisation, while labelling reduces it. Notably, larger effects are seen among ideologically distant Liberal Democrat voters, though none reach statistical significance. In contrast, the trait- and emotional-based outcomes show stronger evidence: unlabelled AI-generated content increases discomfort and reduces respect for opposing partisans. These effects are most pronounced when the AI source is undisclosed, indicating that undetected AI content may be more persuasive than human-generated content and more likely to entrench affective divisions.

For the core measure—the thermometer gap between ratings of in- and out-party leaders—a higher score indicates greater polarisation. Regression results for this outcome are reported in Table 7, Table 8, and Table 9.[2]

All three treatments provide positive treatment effects on affective polarisation, but none are statistically significant. For the overall effect of unlabelled AI-generated content in in Table 7, the effect was `4.297` points, with Liberal Democrats (compared to a base of Reform Party supporters) showin the largest subgroup positive treatment effect of `9.807` points. With unlabelled AI content, we predict that if detection rates are low, and the persuasiveness of the content is high, then polarisation will increase. The positive point estimate is directionally consistent with this, but the null result suggests heterogeneity or insufficient power.

For the labelled AI-generated content in Table 8 compared to the human-generated content, the treatment effect is `-13.49` points. This is a test of the trust discount applied when content is known to be AI. The theory predicted lower responsiveness, a lower gap, which is seen on average. This aligns with the discounting hypothesis: when content is clearly labelled as AI, participants may apply a credibility penalty. However, Green Party respondents see a large (but not statistically significant) positive effect of `6.914` points. This result challenges the core discounting prediction and might suggest that Green Party respondents did not discount labelled AI content; instead, labelling content as AI could fuel polarisation as viewers may associate

[2]The full models for the outcome variables of `MLthermoMean` and `LLthermoMean` are available in the appendix in Section 4.7.

Table 7: AI Effect (Discounting and Detection): Thermometer Gap Results

	Treatment Only	Treatment + Covariates	Full Model
(Intercept)	59.184***	39.719***	57.347***
	(1.792)	(10.389)	(12.149)
AI Treatment	1.296	−1.144	4.297
	(2.596)	(2.430)	(9.761)
AI Treatment:mostlikelyConservative Party			−5.072
			(7.654)
AI Treatment:mostlikelyGreen Party			6.914
			(9.921)
AI Treatment:mostlikelyLabour Party			3.451
			(6.644)
AI Treatment:mostlikelyLiberal Democrats			9.807
			(7.083)
AI Treatment:Political Attention			−0.054
			(1.159)
AI Treatment:Education LevelHigh			−8.437
			(7.563)
AI Treatment:Education LevelMedium			−6.268
			(7.430)
Num.Obs.	554	479	479
R2	0.001	0.173	0.224
RMSE	28.78	25.80	24.99
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Treatment compares AI-generated content to human-generated content. Models weighted using YouGov survey weights. Coefficients are reported with robust standard errors in parentheses. Main effects of included moderators are also reported as rows above the moderator treatment effects.

Table 8: Credibility Effect: Thermometer Gap Results (Labelled AI vs Human, No Label)

	Treatment Only	Treatment + Covariates	Full Model
(Intercept)	59.184***	42.434***	55.005***
	(1.792)	(11.259)	(12.243)
Label Treatment	−1.097	−1.481	−4.355
	(2.561)	(2.365)	(10.684)
Label Treatment:mostlikelyConservative Party			−5.663
			(7.183)
Label Treatment:mostlikelyGreen Party			13.322
			(8.845)
Label Treatment:mostlikelyLabour Party			−4.987
			(6.583)
Label Treatment:mostlikelyLiberal Democrats			−1.827
			(8.548)
Label Treatment:Political Attention			0.457
			(1.142)
Label Treatment:Education LevelHigh			3.142
			(7.408)
Label Treatment:Education LevelMedium			2.693
			(7.219)
Num.Obs.	561	493	493
R2	0.000	0.176	0.213
RMSE	28.54	24.79	24.29
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Treatment compares labelled AI-generated content to unlabelled human-generated content. Models weighted using YouGov survey weights. Coefficients are reported with robust standard errors in parentheses.

Table 9: Detection Effect: Thermometer Gap Results (Labelled AI vs Unlabelled AI)

	Treatment Only	Treatment + Covariates	Full Model
(Intercept)	60.479***	29.347**	49.321***
	(1.879)	(11.035)	(12.135)
Label Treatment	−2.393	−1.331	−13.490
	(2.622)	(2.486)	(9.047)
Label Treatment:mostlikelyConservative Party			2.162
			(7.098)
Label Treatment:mostlikelyGreen Party			8.047
			(11.561)
Label Treatment:mostlikelyLabour Party			−5.031
			(6.061)
Label Treatment:mostlikelyLiberal Democrats			−7.802
			(8.102)
Label Treatment:Political Attention			0.669
			(1.115)
Label Treatment:Education LevelHigh			10.486
			(6.849)
Label Treatment:Education LevelMedium			12.007+
			(6.772)
Num.Obs.	543	470	470
R2	0.002	0.224	0.264
RMSE	28.24	25.53	24.94
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Treatment compares labelled AI-generated content to unlabelled AI-generated content. Models weighted using YouGov survey weights. Coefficients are reported with robust standard errors in parentheses.

the 'fake' AI content as a target from their out-group, pushing them further away.

The reslts from the dection effect model of labelled versus non-labelled AI-generated content in Table 9 shows labelled AI content reduced thermometer gap by `-4.355` points relative to unlabelled AI content. This result captures the pure effect of detection via labelling. While statistically insignificant, the direction and size of effects strongly support the theory that detection (via labelling) reduces affective polarisation, consistent with reduced belief updating from discounted AI content.

Figure 1: Thermometer Ratings: Average Treatment Effects

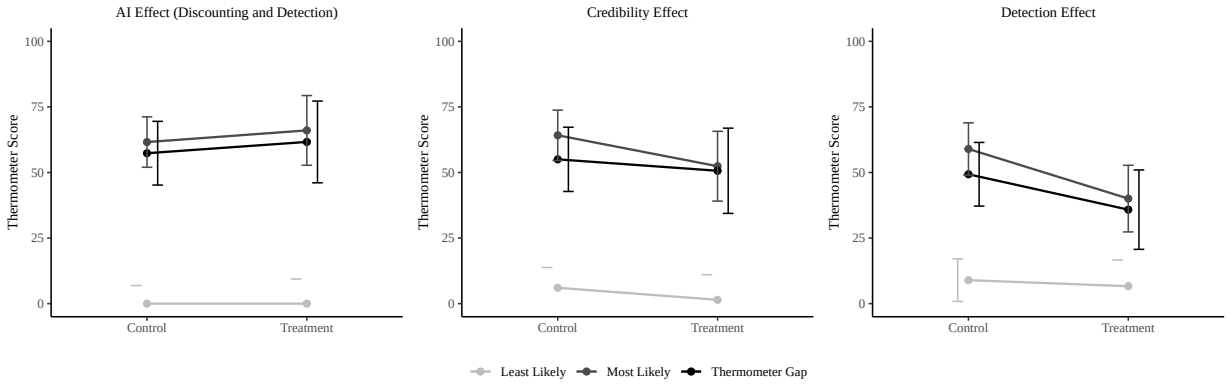
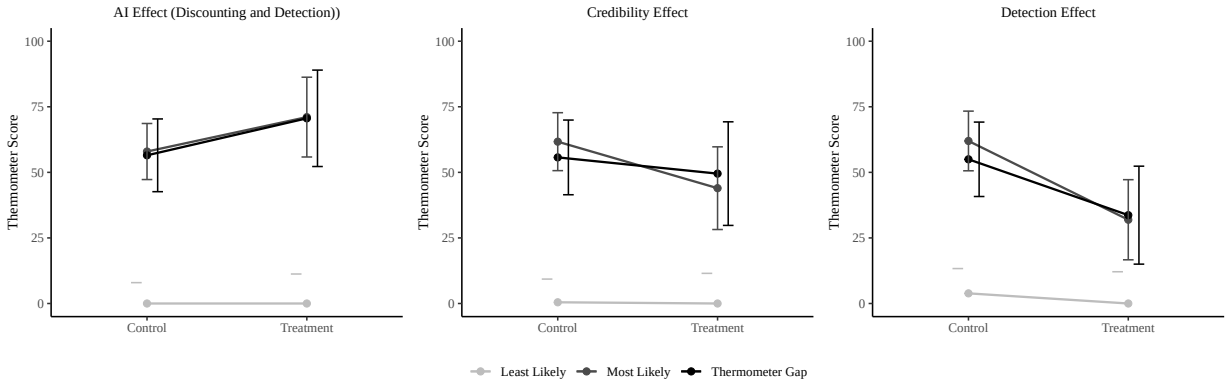


Figure 2: Thermometer Ratings: Average Treatment Effects for Liberal Democrat Respondents



These results are summarised in Figure 1, which show how the thermometer gap is driven by further in- or out-group polarisation. As described above, the overall effect of AI-generated content shows a weak trend toward greater polarisation, primarily via more positive in-group affect which is consistent with undetected AI content being more persuasive. On the contrary, the labelled AI content shows a trend towards reduced polarisation, primarily via reduced in-group affect. The detection effect shows a similar trend, with reduced in- and out-group affect, suggesting that when AI content is detected, it is less persuasive and reduces polarisation.

Table 10: Treatment Effects of Trait- and Emotional-Based Outcome Measures

	Credibility Effect	Detection Effect	Discounting and Detection
Discomfort			−0.620
			(0.803)
Discomfort		−1.212+	
		(0.703)	
Discomfort	−1.736*		
	(0.797)		
Trust			−0.768
			(0.693)
Trust		−0.055	
		(0.794)	
Trust	−0.798		
	(0.741)		
Respect			−0.546
			(0.731)
Respect		−1.202	
		(0.754)	
Respect	−1.749*		
	(0.771)		

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Ordered logistic regression results. Table shows only treatment effects from full models with robust standard errors.

Figure 2 shows the same patchwork plot for respondents most likely to vote for the Liberal Democrats. This subgroup likely has greater ideological distance from the content, which is assumed to target right-wing themes (e.g., Reform Party). According to the theory, this group is expected to show lower responsiveness overall, but may still exhibit increased in-group warmth due to contrast effects when the AI source is undetected. Consistent with this, the results show that unlabelled AI content increases the thermometer gap for LD voters, while labelling the content (i.e. detection) significantly reduces in-group warmth and narrows the gap — providing clear support for the predicted role of ideological distance and detection in moderating affective polarisation.

To further assess treatment effects, ordinal logistic regression models are estimated for the outcomes *respect*, *trust*, and *discomfort* towards opposing partisans. These models account for the ordinal nature of the responses, with coefficients representing the log-odds of being in a higher response category (i.e. greater respect, trust, or comfort). The main ATE results for each measure and treatment are summarised above.[3] [4]

[3]The full models for the outcome variables of `agreedisagree`, `xtrust`, and `child` are available in the appendix in Section 4.8.

[4]The predicted probabilities of each outcome category for each treatment condition are not calculated due to limitations of the `polr()` package, especially when using survey weights.

Exposure to unlabelled AI-generated content provides compelling evidence that undetected AI increases affective polarisation. The log-odds coefficients for discomfort and respect are `-0.546` and `-0.62`, respectively — both statistically significant at the 0.05 level. Negative coefficients across all measures indicate that unlabelled AI content reduces comfort, respect, and trust toward opposing partisans. This supports the hypothesis that undetected AI content reinforces polarisation, even after a single exposure in a small pilot. Notably, Liberal Democrat respondents show weaker effects, suggesting that ideological distance moderates responsiveness, with some respondents potentially persuaded by content on its own merits.

For labelled AI content, there is limited evidence of treatment effects. Aside from a small significant result on discomfort under the detection condition, all other coefficients are statistically insignificant. While the direction of effects remains negative—suggesting reduced persuasion due to detection—estimates are too small to draw firm conclusions.

The key takeaway is that undetected AI-generated content can heighten affective polarisation, particularly through emotional and trait-based reactions. These early results support the hypothesis that AI can be a persuasive tool for polarising attitudes. A follow-up YouGov survey will explore the emotional pathways driving these effects and inform the development of the agent-based model.

3.2 Agent-based Modelling

A primary limitation of the pilot study lies in its statistical power, due to the restricted sample size. As detailed in Section 4.9, detecting a treatment effect of five percentage points with 80% power requires a minimum sample of 3,000 respondents. The YouGov UniOM sample includes only 2,001 individuals—insufficient for detecting effects once subsetting by treatment group and respondent characteristics, where cell sizes often fall below 500. This constraint limits the capacity to test heterogeneous treatment effects or run repeated exposure trials. To address this, a complementary approach is proposed: agent-based modelling (ABM) using synthetic respondents generated by fine-tuned large language models (LLMs).

ABM simulates the beliefs, behaviours, and interactions of individual agents in a virtual environment, enabling dynamic testing of treatment effects beyond the constraints of sample size (Aher, Arriaga and Kalai, 2023; Anthis *et al.*, 2025). Inspired by recent advances in computational social science, this approach draws on Argyle *et al.* (2023)'s method of conditioning LLMs on real-world sociodemographic and survey data to create synthetic digital twins. These agents can be embedded in realistic scenarios to estimate how exposure to AI-generated political content—especially under repeated or cumulative conditions—shapes affective polarisation. The value of this method lies in its ability to replicate complex social dynamics while dramatically lowering the cost and sample-size demands of traditional RCTs (Angelopoulos *et al.*, 2023).

However, the effectiveness of ABM relies heavily on how well the LLM is fine-tuned. Evidence from Park *et al.* (2024) and Zhang *et al.* (2025) shows that conditioning models on specific, high-quality training

data—especially survey and political response data—can significantly improve their predictive and behavioural validity. Yet this method is not without limitations. As shown by Santurkar *et al.* (2023) and Larooij and Törnberg (2025), models trained on non-representative sources (e.g., X/Reddit) risk introducing bias, limiting generalisability. The next steps for this research will be to collect data from an array of sources such as X, the British Election Study, and sociological surveys to ensure the LLM is well-conditioned for the UK context. With the help of the training pipeline developed by Zhang *et al.* (2025) and the *Talking to Machines* project at Nuffield College, Oxford, I hope to be able to fine-tune a model to simulate experimental conditions and test not only single exposures but also repeated and conditional treatments, enabling an exploration of cumulative polarisation effects. The biggest challenge is expected to be the collection of high-quality, representative training data, along with the computational resources required to fine-tune the LLM. Results from this approach will be compared against the survey experiment findings to validate the model's predictive accuracy and generalisability.

4 Conclusion

This project proposes an integrated empirical and computational approach to examine the polarising effects of AI-generated political content. By combining survey experiments with agent-based modelling (ABM), the study advances our understanding of how trust, detection, and ideological distance moderate affective polarisation. The survey design enables causal identification of short-term exposure effects, while the ABM extends this to simulated long-term dynamics as well as other treatment articles, addressing a gap in existing literature. Fine-tuning LLM agents to replicate realistic public opinion patterns presents technical and ethical challenges, but studies which demonstrat this as a feasible and promising strategy are encouraging. The next steps are to get more granualar data on the emotional mechanisms driving affective polarisation through an additional YouGov UniOM survey. Data collection and initial fine-tuning will follow before Michalmas 2025. These new modelling techniques specific to polarisation in the United Kingdom have potential to offer novel theoretical and policy-relevant insights into the political risks posed by generative AI.

Appendix

4.1 Literature Review

This question builds upon the rise of fake news and affective polarisation with a distinct, new focus on the effects of AI-generated content, an area yet to be explored in the academic literature. This section first provides the context for the question's focus, then gives an overview of the existing, limited literature on AI-generated content, and assesses the mixed literature on affective polarisation.

This research focuses on the United Kingdom (UK). Structural effects of globalisation and economic liberalism, coupled with individual political failings and electoral shocks, have created an increasingly unequal and divided world. Consequent disillusionment and disconnected identities have encouraged voter volatility and rising populist narratives, notably in the UK (Norris and Inglehart, 2019; Fieldhouse *et al.*, 2019: 28–32). This environment, combined with social media, has encouraged the dangerous spread of fake news, which has been shown to favour populists, affect voting behaviour, and strengthen identities and affective polarisation within online echo chambers (Cantarella, Fraccaroli and Volpe, 2023; Pfister *et al.*, 2023). Given this volatile political landscape in the UK, with rising populist challengers, fears of the widespread dissemination of deceitful AI-generated information are justified. This research hopes to illuminate the extent to which we should be concerned about AI's effects in the UK setting.

Generative AI is a subfield of Artificial Intelligence with the ability to generate new content in the form of text, images, and video based on generative models that use machine learning to identify patterns in training data (Sengar *et al.*, 2024). This AI-generated content, while often produced via human prompts, is generally computationally generated using probabilities rather than fact-checked, predefined truths. This research defines AI-generated content as any material produced by AI-based models, primarily through human prompting, to provide people with information, news, or arguments on any question or topic, including political ones. The focus is on whether such AI-generated content can affect political attitudes and behaviour, and therefore increase affective polarisation: the gap between the emotional warmth and attachment towards one's in-group political party and the hostility shown to the out-group party (Green, Palmquist and Schickler, 2004; Iyengar, Sood and Lelkes, 2012). While affective polarisation is returned to later, it is important to raise why there is such concern about AI-generated content. As emphasised by Iyengar *et al.* (2019), "exposure to messages attacking the out-group reinforces partisans' biased views of their opponents." This negative messaging often takes the form of 'fake news,' a term Tandoc, Lim and Ling (2018) associate with questions of facticity and audience perception. This issue is critical in the context of AI-generated content and may help explain why increased divisions are emerging within political societies. If AI generates fake or misleading content that spreads widely and targets out-groups within in-group echo chambers, polarisation may inevitably increase.

Despite the infancy of these AI tools, their use in politics is already a point of contention. Hackenburg and

Margetts (2024) show that GPT-4 can have a persuasive influence in political microtargeting. While the research found evidence only for increased support for pre-existing policy perspectives, AI could potentially be used to persuade volatile voters to switch sides. For this reason, OpenAI sought to limit political bias by ensuring "ChatGPT did not express political preferences or recommend candidates even when asked explicitly," whereas others, such as X's Grok, have been caught sharing divisive political disinformation (OpenAI, 2024; Global Witness, 2024; Conger, 2025). These early, unregulated yet widely used models are therefore raising concerns that the content they generate may have negative consequences for elections through the spread of misinformation. The World Economic Forum (2024) identifies this as a severe short-term risk, stating that "AI is amplifying manipulated and distorted information that could destabilise societies."

But AI-generated misinformation goes beyond inaccurate chatbots. There is also fear around the deliberate and manipulative use of AI to generate deepfake images and videos that perpetuate divisive stereotypes or false narratives. However, because the technology remains nascent, many deepfakes are still detectable, highlighting the importance of including detectability as a variable in research design (Kapoor, 2024). But what happens if deepfakes or AI-generated misinformation go undetected? Despite limited political science literature on AI, early findings suggest that AI-generated messages can be persuasive, and that propaganda produced by AI can be compelling (Bai *et al.*, 2023; Goldstein *et al.*, 2024). As the quality of machine learning improves, AI-generated content is likely to become increasingly realistic and undetectable. Studies have shown that AI chatbots can be more persuasive than humans, in part because their personalised responses can exploit heterogeneity in users' views, especially in relation to in-group and out-group perspectives (Salvi *et al.*, 2025). Yet at the same time, these GPTs frequently hallucinate the facts they present, lending credence to fears that AI will perpetuate fake news, partly due to the way reinforcement learning algorithms function (Thornhill, 2025).

Another issue concerns unintended consequences linked to perceptions of AI. It has been found that when readers are aware that political content is AI-generated, they may become sceptical of its validity, even when the content is factually accurate (Altay and Gilardi, 2024). A possible mechanism here is trust. Users may associate AI-generated content—whether due to their own detection or because it is labelled—with fake news, thereby increasing scepticism. Consequently, labelling content as AI-generated may represent a flawed intervention. With detection, labelling, and the association of AI content with falsehoods, Cashell (2024) argues that deepfakes and AI-generated materials are often used to reinforce existing stereotypes, rather than to persuade audiences to adopt new views.

To understand the causal effect of AI-generated content on affective polarisation—and how this may be driven by its overlap with fake news—the conceptualisation and drivers of affective polarisation must be considered. At its core, political polarisation refers to the distribution of ideological positions across a population (Hare, 2022). This explains policy polarisation. However, partisan identity has grown in salience, to the point

where it now better predicts voting behaviour than ideological disagreement (Algara and Zur, 2023).[5] Recent research suggests that the "affect" in affective polarisation—emotional hostility toward out-partisans—is driven by fear, anxiety, disgust, and animosity (Bakker and and Lelkes, 2024). Summarised as partisan disdain, affective polarisation reinforces itself. These emotions shape information engagement and selectivity. Consequently, the information environment, especially on social media, distorts perceptions of political opponents, encouraging readers to engage with content that reflects negatively on the out-group.

This is where AI poses a threat. Angry partisans seek disconfirming information that affirms their beliefs (MacKuen *et al.*, 2010). AI can quickly and easily generate such disconfirming content. It enables affectively polarised individuals to disseminate fictitious, divisive, and realistic-seeming material, often tailored to resonate with the correct in- and out-groups. As a result, affective polarisation may intensify, eroding trust in democratic institutions, increasing discrimination, and reducing political engagement (Layman, Carsey and Horowitz, 2006; Kingzette *et al.*, 2021). With this conceptual grounding, the next section presents a formal model to predict how exposure to AI-generated content might influence affective polarisation.

[5]Recent literature has also suggested that the ideological and policy differences between parties is also related to growing affective polarisation (Gidron, Adams and Horne, 2020; Hobolt, Leeper and Tilley, 2021).

4.2 Codebook

The codebook below provides a summary of the variables used in the YouGov UniOM analysis. The variable names are provided in the first column, followed by the type of variable (e.g., categorical, continuous), a description of the variable, and the values that the variable can take. Note that the outcome variables of `agreedisagree`, `xtrust`, and `child` are ordinal variables on an ordered Likert scale.

Table 11: YouGov UniOM Survey Codebook

Variable	Type	Description	Values
`identity_client`	Identifier	Unique identifier for the respondent	Alphanumeric string
`weight`	Continuous	Survey weight to ensure national representativeness	Continuous float (e.g., 0.982, 1.034)
`age`	Continuous	Age of the respondent	Integer values, typically 18–90
`profile_gender`	Categorical	Gender of the respondent	Female; Male
`profile_GOR`	Categorical	Government Office Region (region of residence)	East Midlands; East of England; London; North East; North West; Scotland; South East; South West; Wales; West Midlands; Yorkshire and the Humber
`voted_ge_2024`	Categorical	Did the respondent vote in the 2024 General Election?	Don't know; No, did not vote; Yes, voted
`pastvote_ge_2024`	Categorical	How the respondent voted in the 2024 General Election	Conservative; Don't know; Green; Labour; Liberal Democrat; Other; Plaid Cymru; Reform UK; Scottish National Party (SNP); Skipped
`pastvote_EURef`	Categorical	How the respondent voted in the 2016 EU Referendum	Can't remember; I did not vote; I voted to Leave; I voted to Remain
`education_recode`	Categorical	Re-coded education level (grouped)	High; Medium; Low
`profile_work_stat`	Categorical	Employment status	Full time student; Not working; Other; Retired; Unemployed; Working full time (30+ hrs); Working part time (8–29 hrs); Working part time (<8 hrs)
`political_attention`	Continuous	How much attention the respondent pays to politics	Scale (e.g., 0–10 or continuous values)

Table 11: YouGov UniOM Survey Codebook *(continued)*

Variable	Type	Description	Values
`split`	Categorical	Randomly assigned treatment group (1–4)	1 = AI-generated, not labelled as AI-generated; 2 = AI-generated and labelled as AI-generated; 3 = Human-generated but labelled as AI-generated; 4 = Human-generated, not labelled as AI-generated
`xconsent`	Categorical	Consent to participate in the survey	I consent to taking part in this study; I do not wish to continue with this study
`mostlikely`	Categorical	Which of these parties would you be most likley to vote for?	Conservative Party; Green Party; Labour Party; Liberal Democrats; Reform UK
`leastlikely`	Categorical	Which of these parties would you be least likley to vote for?	Conservative Party; Green Party; Labour Party; Liberal Democrats; Reform UK; None of these; Not Asked
`MLthermo_KB`	Continuous	Thermometer rating for Kemi Badenoch (most likely party)	0–100
`MLthermo_KS`	Continuous	Thermometer rating for Keir Starmer	0–100
`MLthermo_NF`	Continuous	Thermometer rating for Nigel Farage	0–100
`MLthermo_ED`	Continuous	Thermometer rating for Ed Davey	0–100
`MLthermo_CD`	Continuous	Thermometer rating for Carla Denyer	0–100
`MLthermo_AR`	Continuous	Thermometer rating for Adrian Ramsay	0–100
`LLthermo_KB`	Continuous	Thermometer rating for Kemi Badenoch (least likely party)	0–100
`LLthermo_KS`	Continuous	Thermometer rating for Keir Starmer	0–100
`LLthermo_NF`	Continuous	Thermometer rating for Nigel Farage	0–100
`LLthermo_ED`	Continuous	Thermometer rating for Ed Davey	0–100

Table 11: YouGov UniOM Survey Codebook *(continued)*

Variable	Type	Description	Values
`LLthermo_CD`	Continuous	Thermometer rating for Carla Denyer	0–100
`LLthermo_AR`	Continuous	Thermometer rating for Adrian Ramsay	0–100
`agreedisagree`	Ordinal	Trait-based measure of whether out-groups respect in-group beliefs	Strongly disagree; Tend to disagree; Neither agree nor disagree; Tend to agree; Strongly agree
`xtrust`	Ordinal	Level of trust in out-group to do what is right	Almost never; Once in a while; About half of the time; Most of the time; Always
`child`	Ordinal	Social-disance measure of a child marry an out-group voter	Extremely upset; Somewhat upset; Neither happy nor upset; Somewhat happy; Extremely happy
`MLthermoMean`	Continuous	Average thermometer score for most likely party	0–100 (row mean of MLthermo scores)
`LLthermoMean`	Continuous	Average thermometer score for least likely party	0–100 (row mean of LLthermo scores)
`thermo_gap`	Continuous	Difference between MLthermoMean and LLthermoMean	0–100 (MLthermoMean - LLthermoMean)
`ai_treatment`	Binary	Treatment status for AI-generated content	1 = Treated (shown AI-generated); 0 = Control (shown human-generated)
`label_treatment`	Binary	Treatment status for AI-labelled content	1 = Treated (labelled as AI-generated); 0 = Control (labelled as human-generated)

4.3 Experimental Variables

Although politics is about persuasion, it is hard to move people's beliefs — at least, not very quickly. Cobb and Kuklinski (1997) argue that voters often assess political arguments through emotional, feeling-based heuristics. Therefore, the treatment in the survey experiment was designed to be on a divise topic, which is likely to elicit strong feelings to encourage movement, especially as a single exposure is unlikely to change attitudes significantly. The treatment was an article on rising immigration in the UK, a topic which is prominment amongst populist rhetoric and polarised groups. The human-generated control article was written by McKiernan and Cornock (2024) for the BBC. The AI-generated article was generated by OpenAI's GPT-4 model using the BBC article as an initial prompt, but instructed to re-write the article in a more divisive and exaggerated tone. Full versions of the articles can be found in the Appendix in Section 4.4. Given the nature of the imflammatory topic of immigration used particularly by right-wing populists, the treatment is likely aligned with the ideological priors of right wing Conservative or Reform Party supporters, minimising their ideological distance (δ), and therefore maximising the potential for belief updating and further affective polarisation shown in Equation 23.

The measures required to understand AI's affect on affective polarisation are multi-faceted. Different measures can be used to understand the primary outcome of affective polarisation; however, the implication of each measure differs. Druckman and Levendusky (2019) clearly outline the best practices for these affective polarisation measures, and how the measures interact. Therefore, this research chooses to follow these measurement recommendations for use in survey self-reporting (Iyengar *et al.*, 2019).

The most common measure of someone's identifiction with a political party is through a feeling thermometer score. This aims to understand how warmly or coldly someone feels towards the political parties they most and leat prefer. The thermometer scores are measured on a scale of `0` to `100`, where `0` is the coldest and `100` is the warmest.[6] This survey experiment firstly asks respondents to identify their most and least preferred party (`mostlikely` and `leastlikely`), allowing for in- and out-party identities to be exposed. We then ask respondents to firstly rate how warmly they feel towards each of these party's leaders, `MLthermo_XY` and `LLthermo_XY`, where `XY` is replaced by each party leader's initials. The use of party-leader thermometers is a common measure, leaning on valence theory's emphasis on the importance of party leaders in shaping party identification and voting behaviour (Garzia, Ferreira da Silva and Maye, 2023).[7] Moreover, Druckman and Levendusky's (2019: 119) findings show that respondents are more negative towards party elites rather than

[6]The wording for the theremoeter score questions is as follows: "We'd like to get your feelings toward some of our political leaders and other groups who are in the news these days. On the next page, we'll ask you to do that using a 0 to 100 scale that we call a feeling thermometer. Ratings between 50 degrees and 100 degrees mean that you feel favourable and warm toward the person. Ratings between 0 degrees and 50 degrees mean that you don't feel favourable toward the person and that you don't care too much for that person. You would rate the person at the 50-degree mark if you don't feel particularly warm or cold toward the person."

[7]The Green Party has two co-leaders, Carla Denyer and Adrian Ramsay. Therefore, ratings of both leaders are asked, and the thermometer scores for the Green Party are averaged to create a single score for the party. The variables `MLthermoMean` and `LLthermoMean` are used as the final thermometer measures for in- and out-group thermometer scores.

party voters; thus, the focus on party leaders here helps elicit the more visceral feelings. Alongside these in- and out-group measures, a net-difference score (`thermo_gap`) is also calculated as the difference between the thermometer scores (`MLthermoMean - LLthermoMean`) (Iyengar, Sood and Lelkes, 2012).

The next indicator of affective polarisation is a trait-based rating. This measure identifies the traits that respondents associate with opposing parties (Garrett *et al.*, 2014). The limited scope of the survey experiment meant we focussed on the trait of positive trait of *respect*, and whether respondents associated this trait with oppossing parties. Respondents were asked: "To what extent do you agree or disagree with the following statement: `[leastlikely]` party voters respect my political beliefs and opinions." This question — coded as `agreedisagree` — was asked in a Likert scale format of levels of agreement.[8]

Additionally, a similar trait-based measure focussed on *trust* was used (Levendusky, 2013). Here, we ask "And how much of the time do you think you can trust `[leastlikely]` party to do what is right for the country?". This question was also asked in a Likert scale format, with the options of `Almost never`, `Once in a while`, `About half of the time`, `Most of the time`, and `Always`. This measure is coded as `xtrust`. Along with the themometer score, the trait-based views of respect, and trust in opposing parties, Druckman and Levendusky (2019: 119) argue that these measures are good, general measures of prejudices held towards opposing parties.

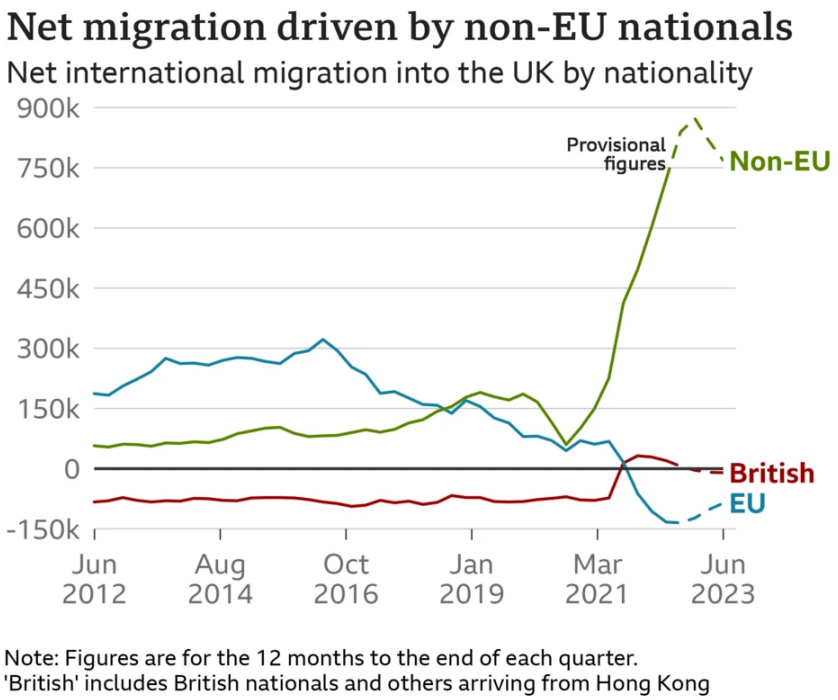
On the other hand, affective polarisation should also be interested in actual tangible discriminatory behaviour. Therefore an emotional, social-distance-based question is included to understand how comfortable respondents are with having opposing partisans in their lives. For example, Iyengar, Sood and Lelkes (2012) popularised the use of the Almond and Verba (1963) five-nation survey question "Suppose you had a child who was getting married. How would you feel if they married a `[leastlikely]` party voter?". Coded as `child`, respondents were given options of `Extremely upset`, `Somewhat upset`, `Neither happy nor upset`, `Somewhat happy`, and `Extremely happy`.

[8] All survey experiment variables and values are in the codebook Section 4.2 in the appendix.

4.4 Survey Experiment Treatments

Due to the limited space in the YouGov UniOM survey experiment, the human-generated BBC article was shortened to fit the survey. Each article was presented with a title, image, and text of roughly 200 words. The human-generated article is presented in Figure 3 below, and the AI-generated article is presented in Figure 4.

Figure 3: Human-Generated Article: BBC News



Dame Priti Patel Defends Record on Rising Immigration

Dame Priti Patel has defended her record on immigration, which hit record levels during her time as home secretary. The MP said it was "too simple" and "lazy" to see the figures without "context" due to the pandemic and the war in Ukraine.

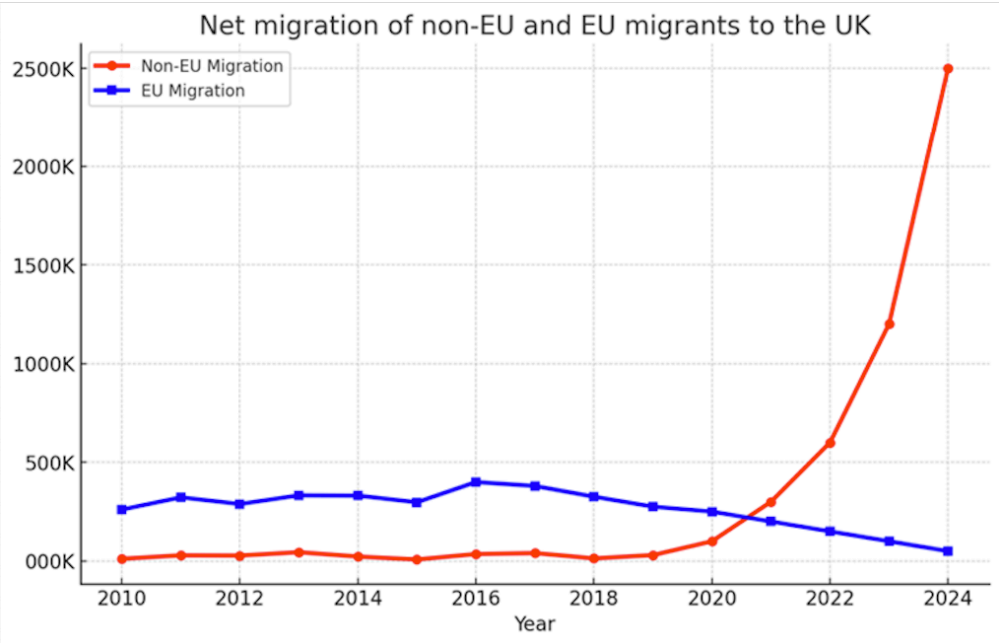
Speaking to supporters she said changes to the system allow ministers to control migration. Patel launched a new points-based immigration system and signed the agreement with Rwanda to send some asylum seekers to the country.

After a drop during the Covid pandemic, net migration - the number of people coming to the UK, minus the number of people leaving - then rose sharply during her time in the role, which she held until September 2022.

In 2022 net migration reached a record 745,000, according to the Office for National Statistics. Some Tories have pointed to the party's failure to reduce migration as a key reason for their historic defeat at the general election.

Pressed over her record on migration, Dame Priti said: "On my watch from 2019 onwards we delivered the structural reforms to the immigration and asylum system. The introduction of a points-based immigration system means the government has the levers to control who comes into our country".

Figure 4: AI-Generated Article: OpenAI GPT-4



Border Crisis: Non-EU Migration Skyrockers as EU Numbers Plunge

UK Overrun as Non-EU Migration Surges!

Britain is facing an immigration crisis, with non-EU migration skyrocketing to unprecedented levels, far outpacing European arrivals.

Official data shows EU migration has declined since 2016, while non-EU arrivals have surged dramatically. A once-steady trend shot up after 2020, as Britain opened its doors to record numbers of foreign nationals.

Critics warn this surge is overwhelming public services, straining housing, schools, and the NHS. “We’ve lost control of our borders,” said a senior political source. Despite government promises, migration numbers keep rising.

Experts blame lax visa rules, a failing asylum system, and reckless international student policies. “It’s unsustainable,” one warned. “Ordinary Britons struggle to buy homes and see a doctor, yet migration is out of control.”

The shift from EU workers—who integrated and contributed—to non-EU arrivals is creating tensions. Communities are struggling with rapid change, and voters are demanding urgent action.

Pressure is mounting for ministers to act. Will they prioritise British workers? Or continue ignoring public frustration as towns and cities buckle under the strain?

One thing is clear: Britain is at a breaking point.

4.5 Data Cleaning

2,001 respondents were provided with the survey experiment. Respondents who did not give consent to participate in the survey were removed. Respondents were given the option to skip questions. When skipped, a value of 997 was assigned to the question, which was then recoded to NA, as were Not asked values.

The survey was interested in understanding respondents' views towards their most and least preferred party. When asked who the mostlikely and leastlikely party was, respondents were given the option to select None of these. Respondents who selected None of these were removed from the sample as they were unable to answer the follow-up questions.

Categorical variables were recoded to be factors in R, these were profile_gender, profile_GOR, voted_ge_2024, pastvote_ge_2024, pastvote_EURef, profile_education_level, education_recode, profile_work_stat, xconsent, mostlikely, leastlikely, agreedisagree, xtrust, and child.

Each of the thermometer variables were recoded to be numeric variables: MLthermo_KB, MLthermo_KS, MLthermo_NF, MLthermo_ED, MLthermo_CD, MLthermo_AR, LLthermo_KB, LLthermo_KS, LLthermo_NF, LLthermo_ED, LLthermo_CD, and LLthermo_AR. As the Green Party has two co-leaders, a mean thermometer score is calculated and used for most and least likely party thermometer scores, coded as MLthermoMean and LLthermoMean.

For treatment effect analysis, respondents were classified into two treatment groups: those shown AI-generated content (ai_treatment), identified where the split variable equalled 1 or 2; and those shown AI-labelled content (label_treatment), identified where the split variable equalled 2 or 3. Participants in the other split groups were coded as receiving human-generated or unlabelled content. These variables were coded as binary variables, where 1 indicated the treatment group and 0 indicated the control group.

4.6 Balance Check

To ensure that the randomisation process of the treatment allocation was successful, a balance check is conducted to ensure that the treatment and control groups are comparable in every way other than their treatment assignment status. The tables below report the balance of the covariates across the treatment groups. The continuous variables of `age` and `political_attention` are reported as means with the standard deviations in parentheses. The remaining categorical variables are reported as a count from the sample, with the proportions in parentheses. If there was a significant difference between the treatment and control groups, this is indicated with a `*` for $p < 0.05$, `**` for $p < 0.01$, and `***` for $p < 0.001$. The balance check shows that randomisation was successful across all covariates for both treatment groups as no covariates were significantly different between the treatment and control groups.

Note that the p-values are reported at the variable level, not for each individual category within a categorical variable. For categorical variables (e.g., gender, vote choice), a single p-value is generated using a chi-squared test, which assesses whether the overall distribution of categories differs between treatment and control groups. The individual category rows are displayed for reference, but since the test is run at the variable level, no p-value is reported for each specific level, giving the `NA` values in the tables.

For each of the categorical variables, there is a base reference category. For example, `profile_gender` uses the base reference category `Male` (reported as `Gender (Male)` in the balance tables). This base acts as the comparison group for the other categories, the p-value compares whether the distribution of the other categories is significantly different from the base category.

Table 12: Balance Table of Covariates by AI Treatment Group

Variable	Control	Treatment	p-value	Signif.
Age	52.54 (16.76)	51.11 (16.45)	0.235	-
Political attention	6.63 (1.96)	6.54 (1.93)	0.502	-
Gender (male)	188 (48.6)	220 (59.8)	0.003	**
Female	199 (51.4)	148 (40.2)	NA	
Education level (High)	75 (19.4)	78 (21.2)	0.796	-
Low	168 (43.4)	153 (41.6)	NA	
Medium	144 (37.2)	137 (37.2)	NA	
Employment status (Full time student)	14 (3.6)	17 (4.6)	0.923	-
Not working	18 (4.7)	19 (5.2)	NA	
Other	8 (2.1)	8 (2.2)	NA	
Retired	118 (30.5)	95 (25.8)	NA	
Unemployed	9 (2.3)	10 (2.7)	NA	
Working full time (30 or more hours per week)	171 (44.2)	167 (45.4)	NA	
Working part time (8-29 hours a week)	44 (11.4)	46 (12.5)	NA	
Working part time (Less than 8 hours a week)	5 (1.3)	6 (1.6)	NA	
Voted in 2024 General Election (Don't know)	1 (0.3)	1 (0.3)	0.999	-
No, did not vote	58 (15.0)	55 (14.9)	NA	
Yes, voted	328 (84.8)	312 (84.8)	NA	
Vote in 2024 General Election (Conservative)	83 (25.3)	65 (20.8)	0.846	-
Don't know	2 (0.6)	2 (0.6)	NA	
Green	23 (7.0)	32 (10.3)	NA	
Labour	115 (35.1)	118 (37.8)	NA	
Liberal Democrat	44 (13.4)	41 (13.1)	NA	
Other	6 (1.8)	4 (1.3)	NA	
Plaid Cymru	1 (0.3)	1 (0.3)	NA	
Reform UK	50 (15.2)	46 (14.7)	NA	
Scottish National Party (SNP)	4 (1.2)	3 (1.0)	NA	
Vote in EU Referendum (Can't remember)	173 (45.4)	170 (46.6)	0.896	-
Skipped	66 (17.3)	65 (17.8)	NA	
I did not vote	142 (37.3)	130 (35.6)	NA	
Region (East Midlands)	25 (6.5)	31 (8.4)	0.805	-
I voted to Leave	39 (10.1)	39 (10.6)	NA	
I voted to Remain	50 (12.9)	34 (9.2)	NA	
East of England	17 (4.4)	15 (4.1)	NA	
London	42 (10.9)	44 (12.0)	NA	
North East	26 (6.7)	31 (8.4)	NA	
North West	56 (14.5)	60 (16.3)	NA	
Scotland	45 (11.6)	35 (9.5)	NA	
South East	13 (3.4)	15 (4.1)	NA	
South West	38 (9.8)	34 (9.2)	NA	
Wales	36 (9.3)	30 (8.2)	NA	

Note: P-values are from t-tests (continuous) or chi-squared tests (categorical) comparing groups. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 13: Balance Table of Covariates by Label Treatment Group

Variable	Control	Treatment	p-value	Signif.
Age	51.11 (16.45)	52.00 (17.05)	0.464	-
Political attention	6.54 (1.93)	6.68 (2.04)	0.310	-
Gender (male)	220 (59.8)	192 (50.0)	0.009	**
Female	148 (40.2)	192 (50.0)	NA	
Education level (High)	78 (21.2)	82 (21.4)	0.940	-
Low	153 (41.6)	155 (40.4)	NA	
Medium	137 (37.2)	147 (38.3)	NA	
Employment status (Full time student)	17 (4.6)	18 (4.7)	0.904	-
Not working	19 (5.2)	18 (4.7)	NA	
Other	8 (2.2)	5 (1.3)	NA	
Retired	95 (25.8)	115 (29.9)	NA	
Unemployed	10 (2.7)	11 (2.9)	NA	
Working full time (30 or more hours per week)	167 (45.4)	171 (44.5)	NA	
Working part time (8-29 hours a week)	46 (12.5)	41 (10.7)	NA	
Working part time (Less than 8 hours a week)	6 (1.6)	5 (1.3)	NA	
Voted in 2024 General Election (Don't know)	1 (0.3)	0 (0.0)	0.324	-
No, did not vote	55 (14.9)	47 (12.2)	NA	
Yes, voted	312 (84.8)	337 (87.8)	NA	
Vote in 2024 General Election (Conservative)	65 (20.8)	78 (23.1)	0.464	-
Don't know	2 (0.6)	4 (1.2)	NA	
Green	32 (10.3)	19 (5.6)	NA	
Labour	118 (37.8)	127 (37.7)	NA	
Liberal Democrat	41 (13.1)	43 (12.8)	NA	
Other	4 (1.3)	8 (2.4)	NA	
Plaid Cymru	1 (0.3)	1 (0.3)	NA	
Reform UK	46 (14.7)	50 (14.8)	NA	
Scottish National Party (SNP)	3 (1.0)	7 (2.1)	NA	
Vote in EU Referendum (Can't remember)	170 (46.6)	167 (44.3)	0.786	-
Skipped	65 (17.8)	67 (17.8)	NA	
I did not vote	130 (35.6)	143 (37.9)	NA	
Region (East Midlands)	31 (8.4)	30 (7.8)	0.931	-
I voted to Leave	39 (10.6)	40 (10.4)	NA	
I voted to Remain	34 (9.2)	39 (10.2)	NA	
East of England	15 (4.1)	11 (2.9)	NA	
London	44 (12.0)	40 (10.4)	NA	
North East	31 (8.4)	33 (8.6)	NA	
North West	60 (16.3)	60 (15.6)	NA	
Scotland	35 (9.5)	35 (9.1)	NA	
South East	15 (4.1)	20 (5.2)	NA	
South West	34 (9.2)	32 (8.3)	NA	
Wales	30 (8.2)	44 (11.5)	NA	

Note: P-values are from t-tests (continuous) or chi-squared tests (categorical) comparing groups. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 14: AI Effect (Discounting and Detection): Thermometer Most Likely Results

	Treatment Only	Treatment + Covariates	Full Model
(Intercept)	69.079***	44.148***	61.597***
	(1.520)	(8.938)	(9.613)
AI Treatment	0.425	−0.717	4.445
	(2.131)	(2.105)	(9.169)
AI Treatment:mostlikelyConservative Party			−2.069
			(5.578)
AI Treatment:mostlikelyGreen Party			9.873
			(7.860)
AI Treatment:mostlikelyLabour Party			6.573
			(5.199)
AI Treatment:mostlikelyLiberal Democrats			8.678
			(5.744)
AI Treatment:Political Attention			−0.338
			(1.242)
AI Treatment:Education LevelHigh			−9.239+
			(4.836)
AI Treatment:Education LevelMedium			−4.872
			(5.026)
Num.Obs.	634	542	542
R2	0.000	0.160	0.241
RMSE	22.78	20.66	19.85
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Treatment compares AI-generated content to human-generated content. Models weighted using YouGov survey weights. Coefficients are reported with robust standard errors in parentheses.

4.7 Thermometer `MLthermoMean` and `LLthermoMean`

Table 15: Credibility Effect: Thermometer Most Likely Results (Labelled AI vs Human, No Label)

	Treatment Only	Treatment + Covariates	Full Model
(Intercept)	69.079***	43.716***	64.171***
	(1.520)	(8.664)	(9.598)
Label Treatment	−2.826	−2.831	−11.781
	(2.186)	(2.010)	(9.220)
Label Treatment:mostlikelyConservative Party			2.860
			(5.448)
Label Treatment:mostlikelyGreen Party			11.461
			(8.308)
Label Treatment:mostlikelyLabour Party			−2.908
			(5.537)
Label Treatment:mostlikelyLiberal Democrats			−5.947
			(6.436)
Label Treatment:Political Attention			1.006
			(1.249)
Label Treatment:Education LevelHigh			4.099
			(5.444)
Label Treatment:Education LevelMedium			4.357
			(5.507)
Num.Obs.	646	561	561
R2	0.004	0.178	0.225
RMSE	23.18	21.50	20.92
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Treatment compares labelled AI-generated content to unlabelled human-generated content. Models weighted using YouGov survey weights. Coefficients are reported with robust standard errors in parentheses.

Table 16: Detection Effect: Thermometer Most Likely Results (Labelled AI vs Unlabelled AI)

	Treatment Only	Treatment + Covariates	Full Model
(Intercept)	69.504***	33.125***	58.962***
	(1.494)	(9.710)	(9.960)
Label Treatment	−3.251	−2.357	−18.915*
	(2.168)	(2.080)	(7.888)
Label Treatment:mostlikelyConservative Party			8.251
			(5.565)
Label Treatment:mostlikelyGreen Party			1.191
			(9.675)
Label Treatment:mostlikelyLabour Party			−6.618
			(4.935)
Label Treatment:mostlikelyLiberal Democrats			−11.118+
			(6.455)
Label Treatment:Political Attention			1.704+
			(0.985)
Label Treatment:Education LevelHigh			9.097+
			(5.151)
Label Treatment:Education LevelMedium			8.898
			(5.450)
Num.Obs.	626	541	541
R2	0.005	0.187	0.250
RMSE	23.57	21.76	21.12
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Treatment compares labelled AI-generated content to unlabelled AI-generated content. Models weighted using YouGov survey weights. Coefficients are reported with robust standard errors in parentheses.

Table 17: AI Effect (Discounting and Detection): Thermometer Least Likely Results

	Treatment Only	Treatment + Covariates	Full Model
(Intercept)	10.499***	2.006	−0.732
	(1.034)	(5.848)	(6.902)
AI Treatment	−1.142	0.301	−5.207
	(1.450)	(1.505)	(6.374)
AI Treatment:mostlikelyConservative Party			−4.665
			(4.824)
AI Treatment:mostlikelyGreen Party			8.679+
			(5.049)
AI Treatment:mostlikelyLabour Party			−1.173
			(4.288)
AI Treatment:mostlikelyLiberal Democrats			−2.276
			(4.754)
AI Treatment:Political Attention			0.848
			(0.793)
AI Treatment:Education LevelHigh			−0.175
			(4.323)
AI Treatment:Education LevelMedium			0.973
			(4.267)
Num.Obs.	647	549	549
R2	0.001	0.096	0.128
RMSE	16.57	15.45	15.27
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Treatment compares AI-generated content to human-generated content. Models weighted using YouGov survey weights. Coefficients are reported with robust standard errors in parentheses.

Table 18: Credibility Effect: Thermometer Least Likely Results (Labelled AI vs Human, No Label)

	Treatment Only	Treatment + Covariates	Full Model
(Intercept)	10.499***	8.567	6.034
	(1.034)	(8.923)	(7.739)
Label Treatment	0.122	0.837	−4.572
	(1.687)	(1.491)	(5.623)
Label Treatment:mostlikelyConservative Party			0.556
			(4.395)
Label Treatment:mostlikelyGreen Party			5.934
			(6.213)
Label Treatment:mostlikelyLabour Party			0.056
			(4.064)
Label Treatment:mostlikelyLiberal Democrats			−4.030
			(4.710)
Label Treatment:Political Attention			0.569
			(0.881)
Label Treatment:Education LevelHigh			−1.501
			(3.907)
Label Treatment:Education LevelMedium			4.302
			(4.744)
Num.Obs.	666	572	572
R2	0.000	0.109	0.161
RMSE	16.75	15.25	15.09
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Treatment compares labelled AI-generated content to unlabelled human-generated content. Models weighted using YouGov survey weights. Coefficients are reported with robust standard errors in parentheses.

Table 19: Detection Effect: Thermometer Least Likely Results (Labelled AI vs Unlabelled AI)

	Treatment Only	Treatment + Covariates	Full Model
(Intercept)	9.357***	9.323	8.949
	(1.017)	(7.640)	(8.121)
Label Treatment	1.264	0.420	−2.304
	(1.677)	(1.730)	(5.809)
Label Treatment:mostlikelyConservative Party			6.170
			(5.034)
Label Treatment:mostlikelyGreen Party			−3.671
			(7.082)
Label Treatment:mostlikelyLabour Party			−2.080
			(4.294)
Label Treatment:mostlikelyLiberal Democrats			−3.719
			(4.859)
Label Treatment:Political Attention			0.521
			(0.911)
Label Treatment:Education LevelHigh			−2.657
			(3.849)
Label Treatment:Education LevelMedium			1.873
			(4.624)
Num.Obs.	643	547	547
R2	0.001	0.111	0.148
RMSE	16.01	15.41	15.33
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Treatment compares labelled AI-generated content to unlabelled AI-generated content. Models weighted using YouGov survey weights. Coefficients are reported with robust standard errors in parentheses.

Table 20: Respect: AI Content vs Human Control (AI Effect: Discounting and Detection)

	Treatment Only	Treatment + Covariates	Full Model
AI Treatment	−0.305	−0.305	−0.546
	(0.185)	(0.185)	(0.731)
AI Treatment:mostlikelyConservative Party			−0.043
			(0.523)
AI Treatment:mostlikelyGreen Party			−0.265
			(0.695)
AI Treatment:mostlikelyLabour Party			−0.258
			(0.592)
AI Treatment:mostlikelyLiberal Democrats			1.353*
			(0.591)
AI Treatment:Political Attention			0.006
			(0.105)
AI Treatment:Education LevelHigh			0.313
			(0.517)
AI Treatment:Education LevelMedium			−0.001
			(0.523)
Num.Obs.	658	658	658
edf	5	5	19
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Note: Ordered logistic regression with survey weights and robust standard errors in parentheses. Coefficients represent log-odds of agreement that opposing partisans respect political beliefs. Threshold cutpoints are not included as they have no substantive interpretation in this context.

4.8 Ordinal Outcome Regression Results

Table 21: Respect: Unlabelled vs Labelled AI Content (Detection Effect)

	Treatment Only	Treatment + Covariates	Full Model
Label Treatment	0.007	0.007	−1.202
	(0.200)	(0.200)	(0.754)
Label Treatment:mostlikelyConservative Party			0.414
			(0.546)
Label Treatment:mostlikelyGreen Party			−0.748
			(0.851)
Label Treatment:mostlikelyLabour Party			1.207+
			(0.619)
Label Treatment:mostlikelyLiberal Democrats			−0.654
			(0.619)
Label Treatment:Political Attention			0.138
			(0.116)
Label Treatment:Education LevelHigh			−0.221
			(0.502)
Label Treatment:Education LevelMedium			0.355
			(0.543)
Num.Obs.	669	669	669
edf	5	5	19
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Note: Ordered logistic regression with survey weights and robust standard errors in parentheses. Coefficients represent log-odds of agreement that opposing partisans respect political beliefs. Threshold cutpoints are not included as they have no substantive interpretation in this context.

Table 22: Respect: Labelled AI Content vs Human Control (Credibility Effect)

	Treatment Only	Treatment + Covariates	Full Model
Label Treatment	−0.295	−0.295	−1.749*
	(0.194)	(0.194)	(0.771)
Label Treatment:mostlikelyConservative Party			0.380
			(0.524)
Label Treatment:mostlikelyGreen Party			−1.007
			(0.743)
Label Treatment:mostlikelyLabour Party			0.967
			(0.606)
Label Treatment:mostlikelyLiberal Democrats			0.683
			(0.658)
Label Treatment:Political Attention			0.144
			(0.117)
Label Treatment:Education LevelHigh			0.085
			(0.524)
Label Treatment:Education LevelMedium			0.355
			(0.525)
Num.Obs.	679	679	679
edf	5	5	19
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Note: Ordered logistic regression with survey weights and robust standard errors in parentheses. Coefficients represent log-odds of agreement that opposing partisans respect political beliefs. Threshold cutpoints are not included as they have no substantive interpretation in this context.

Table 23: Trust: AI Content vs Human Control (AI Effect: Discounting and Detection)

	Treatment Only	Treatment + Covariates	Full Model
AI Treatment	−0.188	−0.188	−0.768
	(0.193)	(0.193)	(0.693)
AI Treatment:mostlikelyConservative Party			−0.078
			(0.553)
AI Treatment:mostlikelyGreen Party			0.313
			(0.732)
AI Treatment:mostlikelyLabour Party			−0.133
			(0.567)
AI Treatment:mostlikelyLiberal Democrats			1.461*
			(0.650)
AI Treatment:Political Attention			0.063
			(0.099)
AI Treatment:Education LevelHigh			0.354
			(0.589)
AI Treatment:Education LevelMedium			−0.137
			(0.553)
Num.Obs.	664	664	664
edf	5	5	19
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001

Note: Ordered logistic regression with survey weights and robust standard errors in parentheses. Coefficients represent log-odds of trusting that opposing parties will do what is right for the country. Threshold cutpoints are not included as they have no substantive interpretation in this context.

Table 24: Trust: Unlabelled vs Labelled AI Content (Detection Effect)

	Treatment Only	Treatment + Covariates	Full Model
Label Treatment	0.082	0.082	−0.055
	(0.198)	(0.198)	(0.794)
Label Treatment:mostlikelyConservative Party			−0.064
			(0.555)
Label Treatment:mostlikelyGreen Party			−0.137
			(0.818)
Label Treatment:mostlikelyLabour Party			0.031
			(0.593)
Label Treatment:mostlikelyLiberal Democrats			−1.130+
			(0.675)
Label Treatment:Political Attention			0.086
			(0.104)
Label Treatment:Education LevelHigh			−0.416
			(0.585)
Label Treatment:Education LevelMedium			−0.219
			(0.569)
Num.Obs.	678	678	678
edf	5	5	19
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Note: Ordered logistic regression with survey weights and robust standard errors in parentheses. Coefficients represent log-odds of trusting that opposing parties will do what is right for the country. Threshold cutpoints are not included as they have no substantive interpretation in this context.

Table 25: Trust: Labelled AI Content vs Human Control (Credibility Effect)

	Treatment Only	Treatment + Covariates	Full Model
Label Treatment	−0.105	−0.105	−0.798
	(0.195)	(0.195)	(0.741)
Label Treatment:mostlikelyConservative Party			−0.128
			(0.527)
Label Treatment:mostlikelyGreen Party			0.171
			(0.805)
Label Treatment:mostlikelyLabour Party			−0.095
			(0.565)
Label Treatment:mostlikelyLiberal Democrats			0.322
			(0.682)
Label Treatment:Political Attention			0.145
			(0.094)
Label Treatment:Education LevelHigh			−0.064
			(0.572)
Label Treatment:Education LevelMedium			−0.347
			(0.552)
Num.Obs.	688	688	688
edf	5	5	19
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Note: Ordered logistic regression with survey weights and robust standard errors in parentheses. Coefficients represent log-odds of trusting that opposing parties will do what is right for the country. Threshold cutpoints are not included as they have no substantive interpretation in this context.

Table 26: Discomfort: AI Content vs Human Control (AI Effect: Discounting and Detection)

	Treatment Only	Treatment + Covariates	Full Model
AI Treatment	−0.015	−0.015	−0.620
	(0.157)	(0.157)	(0.803)
AI Treatment:mostlikelyConservative Party			−0.223
			(0.551)
AI Treatment:mostlikelyGreen Party			−0.697
			(0.582)
AI Treatment:mostlikelyLabour Party			−0.105
			(0.525)
AI Treatment:mostlikelyLiberal Democrats			0.842
			(0.559)
AI Treatment:Political Attention			0.006
			(0.102)
AI Treatment:Education LevelHigh			0.937+
			(0.536)
AI Treatment:Education LevelMedium			0.484
			(0.540)
Num.Obs.	708	708	708
RMSE	2.29	2.29	2.29
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Note: Ordered logistic regression with survey weights and robust standard errors in parentheses. Coefficients represent log-odds of comfort with a child marrying an opposing party voter. Threshold cutpoints are not included as they have no substantive interpretation in this context.

Table 27: Discomfort: Unlabelled vs Labelled AI Content (Detection Effect)

	Treatment Only	Treatment + Covariates	Full Model
Label Treatment	0.032	0.032	−1.212+
	(0.165)	(0.165)	(0.703)
Label Treatment:mostlikelyConservative Party			0.248
			(0.540)
Label Treatment:mostlikelyGreen Party			0.272
			(0.634)
Label Treatment:mostlikelyLabour Party			0.358
			(0.511)
Label Treatment:mostlikelyLiberal Democrats			−0.720
			(0.559)
Label Treatment:Political Attention			0.187+
			(0.107)
Label Treatment:Education LevelHigh			−0.285
			(0.513)
Label Treatment:Education LevelMedium			0.259
			(0.530)
Num.Obs.	699	699	699
RMSE	2.33	2.33	2.33
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Note: Ordered logistic regression with survey weights and robust standard errors in parentheses. Coefficients represent log-odds of comfort with a child marrying an opposing party voter. Threshold cutpoints are not included as they have no substantive interpretation in this context.

Table 28: Discomfort: Labelled AI Content vs Human Control (Credibility Effect)

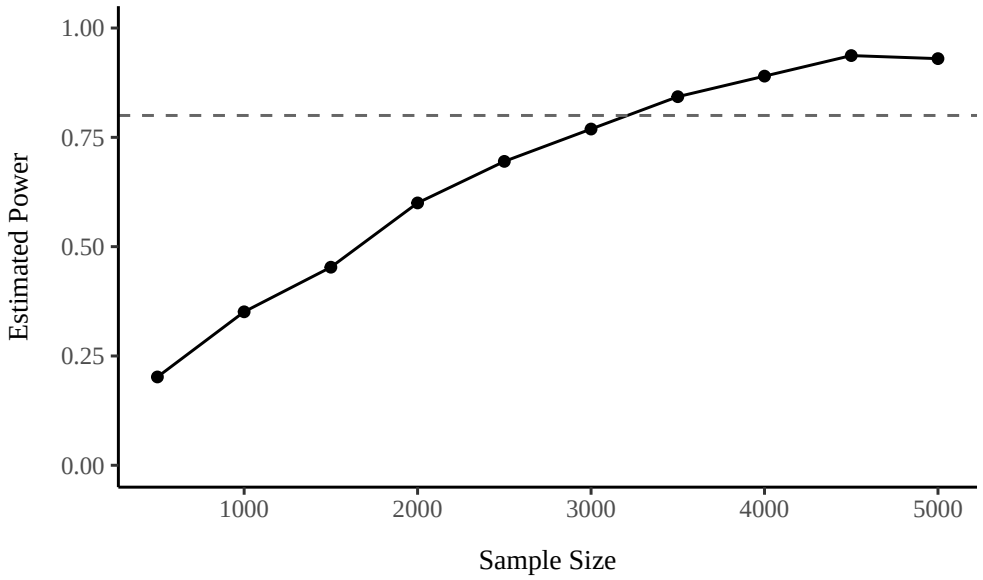
	Treatment Only	Treatment + Covariates	Full Model
Label Treatment	0.016	0.016	−1.736*
	(0.162)	(0.162)	(0.797)
Label Treatment:mostlikelyConservative Party			0.043
			(0.530)
Label Treatment:mostlikelyGreen Party			−0.374
			(0.609)
Label Treatment:mostlikelyLabour Party			0.253
			(0.539)
Label Treatment:mostlikelyLiberal Democrats			0.062
			(0.558)
Label Treatment:Political Attention			0.187+
			(0.100)
Label Treatment:Education LevelHigh			0.571
			(0.503)
Label Treatment:Education LevelMedium			0.685
			(0.519)
Num.Obs.	721	721	721
RMSE	2.32	2.32	2.32
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Note: Ordered logistic regression with survey weights and robust standard errors in parentheses. Coefficients represent log-odds of comfort with a child marrying an opposing party voter. Threshold cutpoints are not included as they have no substantive interpretation in this context.

4.9 Power Analysis

Power analysis is simulated on a model estimating the `thermo_gap` outcome, where the assumed treatment effect is `5` and a residual standard deviation is fairly high at `50`. Estimated across sample sizes of `500-5000`, the power estimates show the share of simulations in which the treatment effect is statistically significant at ((alpha)) = 0.05. To detect an effect at least `80%` of the time, the results show a minimum sample size of `3,000` respondent per treatment subset. These results are summarised in Figure 5 below.

Figure 5: Power Analysis for Thermometer Gap Outcome



References

Acemoglu, D., Ozdaglar, A. and Siderius, J. (2024) 'A Model of Online Misinformation', *The Review of Economic Studies*, 91(6), pp. 3117–3150.

Afroogh, S., Akbari, A., Malone, E., Kargar, M. and Alambeigi, H. (2024) 'Trust in AI: Progress, challenges, and future directions', *Humanities and Social Sciences Communications*, 11(1), pp. 1–30.

Aher, G., Arriaga, R.I. and Kalai, A.T. (2023) 'Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies'. arXiv.

Algara, C. and Zur, R. (2023) 'The Downsian roots of affective polarization', *Electoral Studies*, 82, p. 102581.

Almond, G.A. and Verba, S. (1963) *The Civic Culture: Political Attitudes and Democracy in Five Nations*. Princeton University Press.

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.

Angelopoulos, A.N., Bates, S., Fannjiang, C., Jordan, M.I. and Zrnic, T. (2023) 'Prediction-Powered Inference'. arXiv.

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

Anthis, J.R., Liu, R., Richardson, S.M., Kozlowski, A.C., Koch, B., Evans, J., Brynjolfsson, E. and Bernstein, M. (2025) 'LLM Social Simulations Are a Promising Research Method'. arXiv.

Argyle, L.P., Busby, E.C., Fulda, N., Gubler, J.R., Rytting, C. and Wingate, D. (2023) 'Out of One, Many: Using Language Models to Simulate Human Samples', *Political Analysis*, 31(3), pp. 337–351.

Bai, H., Voelkel, J.G., Eichstaedt, johannes C. and Willer, R. (2023) 'Artificial Intelligence Can Persuade Humans on Political Issues'. OSF [preprint].

Bakker, B.N. and and Lelkes, Y. (2024) 'Putting the affect into affective polarisation', *Cognition and Emotion*, 38(4), pp. 418–436.

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.

Berntzen, L.E., Kelsall, H. and Harteveld, E. (2024) 'Consequences of affective polarization: Avoidance, intolerance and support for violence in the United Kingdom and Norway', *European Journal of Political Research*, 63(3), pp. 927–949.

Cantarella, M., Fraccaroli, N. and Volpe, R. (2023) 'Does fake news affect voting behaviour?', *Research Policy*, 52(1).

Cashell, N. (2024) 'Al-generated images: How citizens depicted politicians and society', *UK Election Analysis*.

Chein, J., Martinez, S. and Barone, A. (2024) 'Can human intelligence safeguard against artificial intelligence? Exploring individual differences in the discernment of human from AI texts', *Research Square*, pp. rs.3.rs–4277893.

Cobb, M.D. and Kuklinski, J.H. (1997) 'Changing Minds: Political Arguments and Political Persuasion', *American Journal of Political Science*, 41(1), pp. 88–121.

Conger, K. (2025) 'Employee's Change Caused xAI's Chatbot to Veer Into South African Politics', *The New York Times* [Preprint].

Della Lena, S. (2024) 'The spread of misinformation in networks with individual and social learning', *European Economic Review*, 168, p. 104804.

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.

Druckman, J.N. and Levendusky, M.S. (2019) 'What Do We Measure When We Measure Affective Polarization?', *Public Opinion Quarterly*, 83(1), pp. 114–122.

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

Fieldhouse, E., Green, J., Evans, G., Mellon, J., Prosser, C., Schmitt, H. and van der Eijk, C. (2019) 'The Rise of the Volatile Voter', in E. Fieldhouse, J. Green, G. Evans, J. Mellon, C. Prosser, H. Schmitt, and C. van der Eijk (eds) *Electoral Shocks: The Volatile Voter in a Turbulent World*. Oxford University Press, pp. 50–73.

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

Garrett, R.K., Gvirsman, S.D., Johnson, B.K., Tsfati, Y., Neo, R. and Dal, A. (2014) 'Implications of Pro- and Counterattitudinal Information Exposure for Affective Polarization', *Human Communication Research*, 40(3), pp. 309–332.

Garzia, D., Ferreira da Silva, F. and Maye, S. (2023) 'Affective Polarization in Comparative and Longitudinal Perspective', *Public Opinion Quarterly*, 87(1), pp. 219–231.

Gidron, N., Adams, J. and Horne, W. (2020) 'American Affective Polarization in Comparative Perspective', *Elements in American Politics* [Preprint].

Global Witness (2024) 'Grok shares disinformation in replies to political queries', *Global Witness*. https://globalwitness.org/en/campaigns/digital-threats/conspiracy-and-toxicity-xs-ai-chatbot-grok-shares-disinformation-in-replies-to-political-queries/.

Goldstein, J.A., Chao, J., Grossman, S., Stamos, A. and Tomz, M. (2024) 'How persuasive is AI-generated propaganda?', *PNAS Nexus*, 3(2).

Green, D., Palmquist, B. and Schickler, E. (2004) *Partisan Hearts and Minds: Political Parties and the Social Identities of Voters*. New Haven, UNITED STATES: Yale University Press.

Hackenburg, K. and Margetts, H. (2024) 'Evaluating the persuasive influence of political microtargeting with large language models', *Proceedings of the National Academy of Sciences*, 121(24), p. e2403116121.

Hare, C. (2022) 'Constrained Citizens? Ideological Structure and Conflict Extension in the US Electorate, 1980–2016', *British Journal of Political Science*, 52(4), pp. 1602–1621.

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.

Hobolt, S.B., Leeper, T.J. and Tilley, J. (2021) 'Divided by the Vote: Affective Polarization in the Wake of the Brexit Referendum', *British Journal of Political Science*, 51(4), pp. 1476–1493.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N. and Westwood, S.J. (2019) 'The Origins and Consequences of Affective Polarization in the United States', *Annual Review of Political Science*, 22(Volume 22, 2019), pp. 129–146.

Iyengar, S., Sood, G. and Lelkes, Y. (2012) 'Affect, Not Ideology: A Social Identity Perspective on Polarization', *Public Opinion Quarterly*, 76(3), pp. 405–431.

Jones, M.I., Pauls, S.D. and Fu, F. (2024) 'Containing misinformation: Modeling spatial games of fake news', *PNAS Nexus*, 3(3), p. pgae090.

Kapoor, S. (2024) 'We Looked at 78 Election Deepfakes. Political Misinformation is not an AI Problem.' https://www.aisnakeoil.com/p/we-looked-at-78-election-deepfakes.

Kingzette, J., Druckman, J.N., Klar, S., Krupnikov, Y., Levendusky, M. and Ryan, J.B. (2021) 'How Affective Polarization Undermines Support for Democratic Norms', *Public Opinion Quarterly*, 85(2), pp. 663–677.

Larooij, M. and Törnberg, P. (2025) 'Do Large Language Models Solve the Problems of Agent-Based Modeling? A Critical Review of Generative Social Simulations'. arXiv.

Layman, G.C., Carsey, T.M. and Horowitz, J.M. (2006) 'PARTY POLARIZATION IN AMERICAN POLITICS: Characteristics, Causes, and Consequences', *Annual Review of Political Science*, 9(Volume 9, 2006), pp. 83–110.

Lee, A.H.-Y., Lelkes, Y., Hawkins, C.B. and Theodoridis, A.G. (2022) 'Negative partisanship is not more prevalent than positive partisanship', *Nature Human Behaviour*, 6(7), pp. 951–963.

Levendusky, M. (2013) *How partisan media polarize America*. Chicago: The University of Chicago Press (Chicago studies in American politics).

Levendusky, M.S. and Stecula, D.A. (2021) 'We Need to Talk: How Cross-Party Dialogue Reduces Affective Polarization', *Elements in Experimental Political Science* [Preprint].

MacKuen, M., Wolak, J., Keele, L. and Marcus, G.E. (2010) 'Civic Engagements: Resolute Partisanship or Reflective Deliberation', *American Journal of Political Science*, 54(2), pp. 440–458.

McKiernan, J. and Cornock, D. (2024) 'Priti Patel defends record on rising immigration', *BBC News*. https://www.bbc.com/news/articles/c93pw69x770o.

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

Norris, P. and Inglehart, R. (2019) *Cultural Backlash: Trump, Brexit, and Authoritarian Populism*. Cambridge: Cambridge University Press.

OpenAI (2024) 'How OpenAI is approaching 2024 worldwide elections'. https://openai.com/index/how-openai-is-approaching-2024-worldwide-elections/.

Park, J.S., Zou, C.Q., Shaw, A., Hill, B.M., Cai, C., Morris, M.R., Willer, R., Liang, P. and Bernstein, M.S. (2024) 'Generative Agent Simulations of 1,000 People'. arXiv.

Pfister, R., Schwarz, K.A., Holzmann, P., Reis, M., Yogeeswaran, K. and Kunde, W. (2023) 'Headlines win elections: Mere exposure to fictitious news media alters voting behavior', *PLOS ONE*, 18(8).

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*.

Salvi, F., Horta Ribeiro, M., Gallotti, R. and West, R. (2025) 'On the conversational persuasiveness of GPT-4', *Nature Human Behaviour*, pp. 1–9.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P. and Hashimoto, T. (2023) 'Whose Opinions Do Language Models Reflect?' arXiv.

Sengar, S.S., Hasan, A.B., Kumar, S. and Carroll, F. (2024) 'Generative Artificial Intelligence: A Systematic Review and Applications'. arXiv.

Simon, F.M., McBride, K. and Altay, S. (2024) 'AI's impact on elections is being overblown', *MIT Technology Review*.

Tandoc, E.C., Lim, Z.W. and Ling, R. (2018) 'Defining "Fake News": A typology of scholarly definitions', *Digital Journalism*, 6(2), pp. 137–153.

Thornhill, J. (2025) 'Generative AI models are skilled in the art of bullshit', *Financial Times* [Preprint].

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

Törnberg, P., Andersson, C., Lindgren, K. and Banisch, S. (2021) 'Modeling the emergence of affective polarization in the social media society', *PLOS ONE*, 16(10), p. e0258259.

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.

World Economic Forum (2024) 'These are the 3 biggest emerging risks the world is facing', *World Economic Forum*. https://www.weforum.org/stories/2024/01/ai-disinformation-global-risks/.

Zhang, X., Lin, J., Mou, X., Yang, S., Liu, X., Sun, L., Lyu, H., Yang, Y., Qi, W., Chen, Y., Li, G., Yan, L., Hu, Y., Chen, S., Wang, Y., Huang, X., Luo, J., Tang, S., Wu, L., Zhou, B. and Wei, Z. (2025) 'SocioVerse: A World Model for Social Simulation Powered by LLM Agents and A Pool of 10 Million Real-World Users'. arXiv.