MPhil Politics: Comparative Government

Thesis: Research Design Proposal

Edward Anders

Supervisor: Professor Rachel Bernhard

St. Antony's College

University of Oxford

June 2025

Abstract

Machine learning advancements to efficiently handle sequential data inputs and outputs have popularised the field of Artificial Intelligence (AI). Amongst AI's applications, generating hyper-realistic textual and visual content has become easily accessible, helping AI become an enabling informational tool. Yet, as unregulated AI technologies remain prone to hallucinations and misuse from bad actors, they are raising concern in social and political contexts. This research project assesses the possible negative effects of manipulative political information and deceitful deepfakes. In particular the mechanism of trust, and the association of AI with fake news, are explored to understand whether partisans exposed to AI-generated content become more affectively polarised to one another. This project uses survey experiments to explore exposure effects, using labels to indicate an AI- or human-generated provenance. Agent-based modelling will also be used to test repeated exposures. Initial hypotheses expect minimal effects, but negative unintended consequences of labelling AI-generated content may be found.

Contents

List of Tables

List of Tables

List of Figures

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 question seeks to an answer a notable puzzle. Why should we fear the fake news, or deceitful propaganda produced by generative-AI applications any more than the propaganda produced for centuries before its time? Plausible arguments lay in both the quantity and quality of the information produced. Despite confident responses, AI can hallucinate to produce false political facts and fantasies, and be directed to generate hyper-realistic, undetectable 'fake news' (Flew, 2021; Duberry, 2022; Rawte *et al.*, 2023). With 45% of the United States population using generative AI, and social media providing a perfect watering ground to promulgate viral misinformation, exposure to AI-generated fake news is increasingly likely (Salesforce, 2025). But, widespread misinformation has been deliberately spread within democracies well before technological advancements, with democracies remaining strong throughout (Bernays, 1928). Therefore, are the growing fears that AI will deceive and manipulate democratic events justified (Ansell, 2023)? To answer this, we can work backwards. Can AI swing enough of the right voters to affect our election outcomes? Can AI-generated content manipulate and persuade individual-level voting behaviour? Can exposure to AI-generated content affect political attitudes, in particular feelings of affective polarisation?

This question and focus on affective polarisation is especially important. Fake news has been shown to

spread rapidly amongst networks of echo chambers which, in turn, are incubators of partisans with ever-more polarised views against their out-groups (Törnberg, 2018; Hobolt, Lawall and Tilley, 2023). With polarisation being a theme of democratic backsliding and populist politics, understanding the effects of AI in fuelling the spread of manipulative fake news to polarise partisans further is imperative. Answering this question provides governments, institutions, and technology companies with political implications of failing to improve governance and regulation of AI models. Moreover, understanding the mechanisms through which AI affects behaviour and the effectiveness of possible interventions is a key implications of this research. Labelling content is ostensibly the best approach to warn that generative AI has been used. Yet this may bring overly adverse associations of AI with fake news which may only exacerbate polarisation (Altay and Gilardi, 2024). Consequently, the effects of labelling content as AI is also an independent variable of interest in this research.

To identify the effects of exposure to AI-generated content, this project proposes the use of survey experiments coupled with AI-augmented synthetic experimental methods to simulate repeated exposures. This proposal starts by reviewing the literatures on AI's effects and affective polarisation, before laying out the theoretical motivations and hypotheses of this research. A pilot study already conducted with YouGov is then presented to give an initial assessment of possible causal effects in an isolated, single exposure setting. Finally, additional proposals for further research using synthetic agents is presented.

2 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 firstly provides the context for the question's focus, before giving an overview of the existing — limited — literature on AI-generated content, and assessing 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 — coupled 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, the fears of widespread dissemination of deceitful AI-generated information are justified. But this research hopes to illuminate to what extent we should be concerned about AI's effect 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, video based on generative models which use machine learning to take patterns from data they are trained on (Sengar *et al.*, 2024). This AI-generated content, while often produced by human prompts, is generally computationally generated using probabilities rather than fact-checked, pre-defined truths. This research defines the use of AI-generated content as any content produced by AI-based models, primarily through human prompting to provide people with information, news, and arguments on any question or topic, including political ones. The focus is on whether such AI-generated content can affect political attitudes, behaviour, and therefore increase affective polarisation: the gap between the emotional warmth and attachment towards your in-group political party, compared to the hostility shown to the out-group party (Green, Palmquist and Schickler, 2004; Iyengar, Sood and Lelkes, 2012). While affective polarisation is return to later, it's important to raise why there is such a fear of 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) focus on facticity and the perceptions of truth from its audience. It is this issue of fake news which is critical for AI-generated content, and why increased divides could be seen within our political societies. If AI generates fake or misleading content which is spread widely and used to attack out-groups within in-group echo chambers, polarisation may inevitably increase.

Despite the infancy of these AI tools, their use in politics is therefore a noticeable point of contention. OpenAI tried to avoid political biases by ensuring 'ChatGPT did not express political preferences or recommend candidates even when asked explicitly,' while others such as X's Grok has been caught sharing divisive political disinformation (OpenAI, 2024; Global Witness, 2024; Conger, 2025). These early, un-regulated, yet widely used models, are therefore raising concern that the content they generate may have negative consequences on elections through the spread of misinformation. The World Economic Forum (2024) sees this as a severe short-term risk as 'AI is amplifying manipulated and distorted information that could destabilise societies.' But AI-generated misinformation is more than just inaccurate chatbots. Fear also surrounds more deceitful and deliberately manipulative uses of AI to generate deepfake images and videos used to perpetuate a divisive stereotype or false narratives. However, the nascent nature of the technology means these deepfakes are often detectable, showing the need to consider the ability of someone to detect the use of AI in the research design (Kapoor, 2024). But what if the deepfakes or AI-generated misinformation goes undetected? Despite minimal literature on AI in political science, early research suggests AI-generated messages can also persuasive, and propaganda produced by AI can be compelling (Bai *et al.*, 2023; Goldstein *et al.*, 2024). Research has shown that AI chatbots can be more persuasive than humans, showing how the personalised nature of GPTs can exploit user heterogeneity and their in- versus out-group views (Salvi *et al.*, 2025). Yet, simultaneously, these GPTs are regularly shown to hallucinate facts they provide, giving credence to the fears that AI's will perpetuate fake news even further due to the reinforcement learning algorithms (Thornhill, 2025). Another issue is potential unintended consequences of consumers' perceptions of AI. It has been found that when aware of political content being AI-generated, readers become sceptical

of its validity even if the content is true (Altay and Gilardi, 2024). A possible mechanism here is trust. Users may associate AI-generated —determined by their own detection, or if labelled — content with fake news, which in turn increases scepticism towards its veracity. Consequently, labelling content as AI-generated may be a misinformed interception. With detection, labelling, and association of AI-generated content with being fake, Cashell (2024) argues deepfakes and AI-generated content is often instead used to perpetuate existing stereotypes rather than attempting to persuade new views.

To consider the causal effect of AI-generated content on affective polarisation and how this may be driven by AI's link with fake news, the conceptualisation and causes of affective polarisation require consideration. At its roots, political polarisation is the distribution of a population along an ideological dimension (Hare, 2022). This ideological description can explain policy polarisation; whereas, differences in partisan identity grew in salience such that social identity with a partisan group became a better predictor of voting behaviour compared to ideological disagreement (Algara and Zur, 2023).[1] However, recent research has suggested that the *affect* in affective polarisation — the emotional animosity felt towards opposing partisans — is driven by emotions of fear, anxiety, disgust, and animosity (Bakker and and Lelkes, 2024). Summarised as partisan disdain, affective polarisation self-reinforces differences. These emotions towards out-groups affects the engagement and selective choices of which information to consume. Consequently, the information environment, primarily on social media, skews the reader's perceptions of reality, engaging themselves with content they want to see as a representation of the out-groups. This is where AI's potential is a threat. Angry partisans seek disconfirming information to support their own views (MacKuen *et al.*, 2010). AI can be easily and quickly used to generate this disconfirming information. AI helps affectively polarised voters exacerbate the spread of fictitious, divisive, yet ostensibly real content, within the correct in- and out-groups. As a result, affective polarisation could increase discrimination, cut trust in democratic institutions, and suppress political engagement (Layman, Carsey and Horowitz, 2006; Kingzette *et al.*, 2021). Taking this understanding of affective polarisation in the context of AI-generated content and the volatile political environment, the next section builds a formal model to predict how exposure to AI-generated content can impact affective polarisation.

3 Theoretical Framework

As described in Section 2 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),

[1] 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).

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.

3.1 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
- $\beta_i \in [0, 1]$: discount factor applied to AI-generated content (lower values indicate greater distrust)
- $d_i \in [0, 1]$: 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:

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

The term $\bar{\beta}_i$ reflects the expected discount applied to the article, based on the detection probability and source-specific trust. The function $w(\delta)$ reflects ideological receptiveness, with greater distance reducing responsiveness.

3.2 Detection and Discounting

As has been widely reported, trust in AI-generated content is low. When individuals are aware that content is AI-generated, they are less likely to trust it (Afroogh *et al.*, 2024). Therefore, detectiom of AI-generated content is a necessary condition for the model. It is assumed that both detection (d_i) and discounting (β_i) depend on observable, theoretically grounded individual-level covariates:

For example, detection Probability (d_i) is increasing in the Equations 10:

$$\frac{\partial d_i}{\partial \text{Education}_i} > 0, \quad \frac{\partial d_i}{\partial \text{Political Attention}_i} > 0 \tag{2}$$

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).

The penalty applied to AI-generated content is also likely based on heterogeneous individual characteristics. For example, higher educated individuals may be less inclined to discount information based solely on its source. They are likely to be more familiar with algorithmic systems and more willing to evaluate content on its merits. Thus, conditional on detection, they apply a smaller penalty to AI-generated material, given by the Discount Factor (β_i)} in Equation 3:

$$\frac{\partial \beta_i}{\partial \text{Education}_i} > 0 \tag{3}$$

Notably, education enters both the detection and discounting components, creating a theoretically rich asymmetry: more educated individuals are more likely to detect AI content, but less likely to penalise it.

3.3 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:

$$u(\theta_1, C) = -(\theta_1 - C)^2 \tag{4}$$

Individuals choose a responsiveness parameter $\mu_i \in [0, 1]$ such that their updated belief is given by:

$$\theta_1 = \theta_0 + \mu_i(C - \theta_0) \tag{5}$$

The individual's optimisation problem becomes:

$$\max_{\mu_i \in [0,1]} -\left[(1 - \mu_i)(\theta_0 - C)\right]^2 \tag{6}$$

The utility in Equation 6 is maximised when $\mu_i = 1$, indicating full belief updating. However, due to factors such as distrust and ideological distance, we assume instead that responsiveness is constrained:

$$\mu_i = \bar{\beta}_i \cdot w(\delta) \tag{7}$$

3.4 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

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

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)
- $\mu_i = w(\delta)$

3.5 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.

3.5.1 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
λ	Responsiveness decay parameter	Model parameter

Recall that responsiveness is defined in Equation 8:

$$\mu_i = \bar{\beta}_i \cdot w(\delta) = \bar{\beta}_i \cdot \frac{1}{1 + \lambda \cdot \delta} \tag{8}$$

where:

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

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

3.5.2 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 9:

$$\frac{\partial \mu_i}{\partial \delta} = \bar{\beta}_i \cdot \frac{\partial w(\delta)}{\partial \delta} = \bar{\beta}_i \cdot \left(\frac{-\lambda}{(1 + \lambda \cdot \delta)^2} \right) < 0 \tag{9}$$

3.5.3 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 10 captures this relationship:

$$\bar{\beta}_i = d_i \cdot \beta_i + (1 - d_i) \quad \Rightarrow \quad \frac{\partial \mu_i}{\partial d_i} = (\beta_i - 1) \cdot w(\delta) < 0 \quad \text{if } \beta_i < 1 \tag{10}$$

When individuals become more likely to detect that content is AI-generated, they apply a greater discount to it. As a result, their responsiveness decreases. This captures the idea that more sophisticated individuals — while better at detecting AI — may also be more sceptical of it.

3.5.4 Discount Factor β_i

For AI-generated articles (labelled or unlabelled), indivisuals 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 11 and Equation 12:

- **Labelled AI:**

$$\mu_i = \beta_i \cdot w(\delta) \quad \Rightarrow \quad \frac{\partial \mu_i}{\partial \beta_i} = w(\delta) > 0 \tag{11}$$

- **Unlabelled AI:**

$$\bar{\beta}_i = d_i \cdot \beta_i + (1 - d_i) \quad \Rightarrow \quad \frac{\partial \mu_i}{\partial \beta_i} = d_i \cdot w(\delta) > 0 \tag{12}$$

3.5.5 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 13:

$$\frac{\partial \mu_i}{\partial \lambda} = \bar{\beta}_i \cdot \frac{\partial w(\delta)}{\partial \lambda} = \bar{\beta}_i \cdot \left(\frac{-\delta}{(1 + \lambda \cdot \delta)^2} \right) < 0 \tag{13}$$

Higher values of λ imply sharper declines in responsiveness with ideological distance. A higher *lambda*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.

3.5.6 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
λ	$\frac{\partial \mu_i}{\partial \lambda}$	Negative	More rigidity reduces responsiveness across the board

3.6 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 14:

$$\text{AP}_i = L_i^{\text{in}} - L_i^{\text{out}} \tag{14}$$

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 $\Delta\theta_i = \theta_1 - \theta_0$ affects the change in affective polarisation:

$$\Delta\text{AP}_i = \Delta L_i^{\text{in}} - \Delta L_i^{\text{out}} \tag{15}$$

3.6.1 Asymmetric Mealleability 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. With greater in-group identification — or particularly high levels of animostity towards the out-group — it can be assumed that affective evaluations is a negative function of the initial strength of feeling. 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 **??** and Equation 17:

$$\Delta L_i^{\text{out}} = \frac{1}{|L_i^{\text{out}}| + \varepsilon} \cdot \Delta\theta_i \tag{16}$$

$$\Delta L_i^{\text{in}} = \frac{1}{|L_i^{\text{in}}| + \varepsilon} \cdot f(\Delta\theta_i) \tag{17}$$

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 18:

$$\Delta\text{AP}_i = \left(\frac{-\phi}{|L_i^{\text{in}}| + \varepsilon} - \frac{1}{|L_i^{\text{out}}| + \varepsilon} \right) \cdot \Delta\theta_i \tag{18}$$

3.6.2 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.

3.7 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 dervived in Equation 18, and restated below, we can define the responsiveness of affective polarisation to a belief chance A_i, and μ_i in Equation 19 to give a cleaner expression for the change in affective polarisation in Equation 20:

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

Letting:

$$A_i = \left(\frac{-\phi}{|L_i^{\text{in}}| + \varepsilon} - \frac{1}{|L_i^{\text{out}}| + \varepsilon} \right) \quad \text{and} \quad \mu_i = \bar{\beta}_i \cdot \frac{1}{1 + \lambda \cdot \delta} \tag{19}$$

we can write:

$$\Delta\text{AP}_i = A_i \cdot \mu_i \cdot (C - \theta_0) \tag{20}$$

3.7.1 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 21:

$$\frac{\partial \Delta\text{AP}_i}{\partial \delta} = A_i \cdot \frac{\partial \mu_i}{\partial \delta} \cdot (C - \theta_0) = A_i \cdot \bar{\beta}_i \cdot \left(\frac{-\lambda}{(1 + \lambda \cdot \delta)^2} \right) \cdot (C - \theta_0) < 0 \tag{21}$$

3.7.2 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 22:

$$\frac{\partial \Delta \text{AP}_i}{\partial d_i} = A_i \cdot \frac{\partial \mu_i}{\partial d_i} \cdot (C - \theta_0) = A_i \cdot (\beta_i - 1) \cdot \frac{1}{1 + \lambda \cdot \delta} \cdot (C - \theta_0) < 0 \quad \text{if } \beta_i < 1 \tag{22}$$

As shown in Section 3.2, 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.

3.7.3 Discount Factor β_i

The discount factor β_i captures the individual's trust in AI-generated content. Higher trust leads to greater belief updating and potentially greater affective polarisation. This is the case for both labelled and unlabelled AI-generated content, as shown in Equation 23 and Equation 24:

- **Labelled AI:**

$$\frac{\partial \Delta \text{AP}_i}{\partial \beta_i} = A_i \cdot w(\delta) \cdot (C - \theta_0) > 0 \tag{23}$$

- **Unlabelled AI:**

$$\frac{\partial \Delta \text{AP}_i}{\partial \beta_i} = A_i \cdot d_i \cdot w(\delta) \cdot (C - \theta_0) > 0 \tag{24}$$

3.7.4 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 25:

$$\frac{\partial \Delta \text{AP}_i}{\partial \phi} = \frac{-1}{|L_i^{\text{in}}| + \varepsilon} \cdot \mu_i \cdot (C - \theta_0) \tag{25}$$

3.7.5 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 26 and for out-group hostility by Equation 27:

$$\frac{\partial A_i}{\partial |L_i^{\text{in}}|} = \frac{\phi}{(|L_i^{\text{in}}| + \varepsilon)^2} > 0 \quad \Rightarrow \quad \frac{\partial \Delta \text{AP}_i}{\partial |L_i^{\text{in}}|} > 0 \text{ if } \Delta\theta_i > 0 \tag{26}$$

$$\frac{\partial A_i}{\partial |L_i^{\text{out}}|} = \frac{1}{(|L_i^{\text{out}}| + \varepsilon)^2} > 0 \quad \Rightarrow \quad \frac{\partial \Delta \text{AP}_i}{\partial |L_i^{\text{out}}|} < 0 \text{ if } \Delta\theta_i > 0 \tag{27}$$

3.7.6 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
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

3.8 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.

9. Experimental Conditions and Theoretical Predictions

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.

9.1 Treatment Structure

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

9.2 Condition-by-Condition Predictions and Heterogeneity

3.8.0.1 (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

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

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

Treatment effect heterogeneity:

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

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

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

- Affective polarisation depends on whether AI authorship is inferred and how strongly it is discounted.

Treatment effect heterogeneity:

- High d_i, low β_i: strong discounting → lower responsiveness
- Low d_i: perceived as human → closer to control
- High education → increases both detection and trust → ambiguous net effect

Key comparisons:

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

9.3 Summary of Theoretical Comparisons

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

- Outline the key assumptions of your argument, as well as the limits to your topic (temporally and spatially).

4 Case Selection and Data Gathering

4.1 YouGov UniOM Survey Experiment

- Case selection (why focus on the UK?)
  - I am to have external validity to my research/case?
  - Can I make inferences to other cases?
- What is the case?
  - What is the unit of analysis?
  - What is the time period?

  - What is the geographical scope?
  - What are the key variables?

- Survey Design
  - Note the use of a between-subjects survey experiment
  - Deliberately chosen to avoid sensitivity issues noted by Levendusky and Stecula (2021)
  - Include treatment matrix

- How is the data being collected?
  - What is the sampling strategy?
  - Note the UK weighting

4.2 Outcome Measures

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.[2] 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).[3] Moreover, Druckman and Levendusky's (2019: 119) findings show that respondents are more negative towards party elites rather than party voters; thus, the focus on party leaders here helps elicit the more visceral feelings. Alongside these in-

[2]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."

[3]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.

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.[4]

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`.

5 Data analysis

The following data analyses focus on all outcome measures of affective polarisation to give a holistic understading of both general and tangible prejudices, and discriminatory behaviours towards the opposing out-group to that of the respondent's identified in-group. The analysis is split by the treatments being tested: AI-generated content and AI-labelled content. Each treatment is analysed across the outcome measures of thermometer scores, trait-based measures, and social-distance measures.

[4] A full breakdown of the survey experiment variables and values can be found in the codebook Section 5.5 in the appendix.

5.1 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 (28):

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 \mathbf{X}_i + \beta_3 (D_i \times \mathbf{Z}_i) + \varepsilon_i \tag{28}$$

where:

- Y_i takes the outcome variables (`thermo_gap`, `MLthermoMean`, `LLthermoMean`, `agreedisagree`, `xtrust`, and `child`)
- D_i is the treatment recieved (`ai_treatment` or `label_treatment`)
- $\mathbf{X}_i$ is a vector of covariates (see Balance Check in Section 5.7 for details)
- $\mathbf{Z}_i$ is a vector of possible interaction terms between the treatment and moderators
- ε_i is the error term

In this full specitifcation, 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.

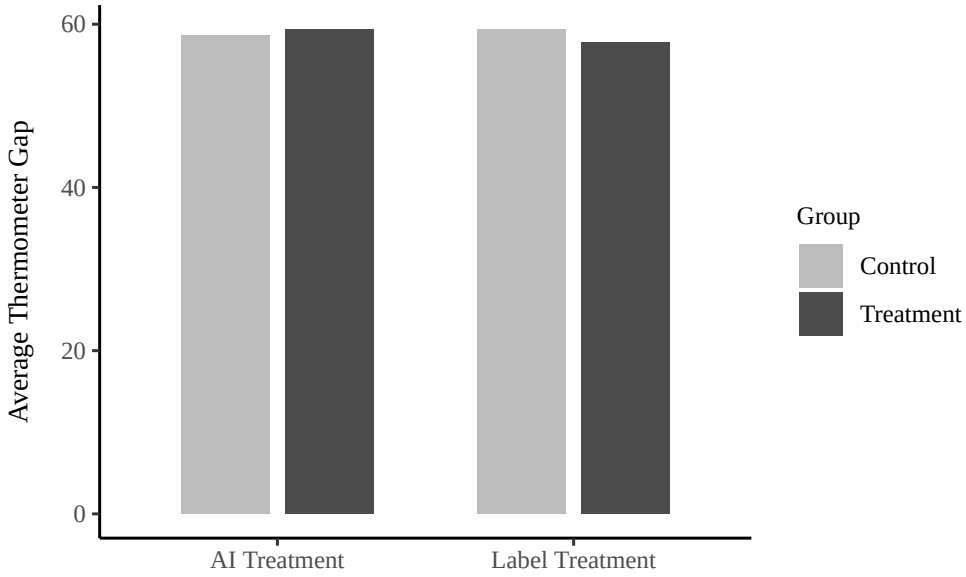
5.2 AI-Generated Content Treatment

The results show no statistically significant treatment effect of AI-generated content on in- and out-party, nor net-difference thermometer scores. However, it is found that Liberal Democrat voters are significantly susceptible to being polarised from expsure to AI-generated content. Trait-based and emotional ratings and views towards opposing parties and voters are also not signficantly affected by the deliberately divsive treatment of the AI-generated content. Nevertheless, the treatment is significantly more likely to polarise lower-educated respondents, and part-time workers on views of respect and emotional discomfort towards opposing partisans. Together, these results suggest that AI-generated content does not significantly polarise respondents' affective polarisation measures, but there are some sub-group differences in the treatment effect.

5.2.1 Thermometer Analysis

Thermometer analysis is one of the primary affective polarisation measures. Before determining a causal link between AI content exposure and the affective polarisation measures, a descriptive summary of the `thermo_gap` measures, averaged over all in- and out-party leaders, given for both treatments is presented in Figure 1. This shows how net-difference thermometer scores are similar across both control and treatment groups, suggesting causal effects are likely to be minimal.

Figure 1: Average in- and out-party thermometer net-difference scores



To test whether this descriptive expectation is causally salient, models for the outcome variables for in- and out-party, and net-difference thermometer scores are estimated. The thermometer outcome scores are continuous measures. Therefore, survey-weighted least squares regression models are estimated.

ATE models are presented in Table 4 for the outcome `thermo_gap`.[5] A first model (1) sets the benchmark without control for covariates and moderators. A full balanace check (Section 5.7) shows that the treatment and control groups were balanced across all convariates. Despite this, model (2) still includes a full set of pre-treatment covariates as each has theoretical justification for affecting the outcome independently of the treatment, and also to ensure the ATE estimates are efficient. To avoid multicollinearity, individual moderators were sequentially tested within the models; however, few showed any moderation effects. The moderators of party affiliation/warmth (`mostlikely`) and attentivness to politics (`political_attention`) showed the greatest moderation effects, thus are included in the final model (3) as interaction terms to test these groups for heterogeneity.

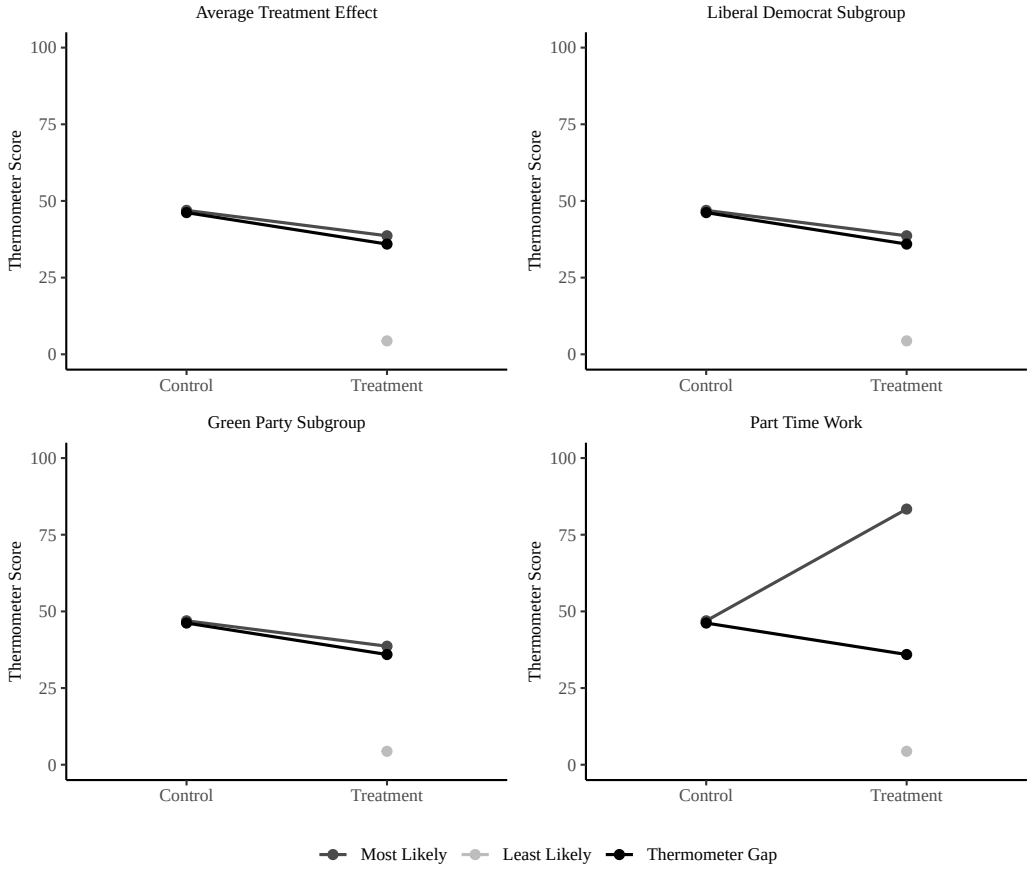
[5]The full models for the outcome variables of `MLthermoMean` and `LLthermoMean` are available in the appendix in Table 9 and

Table 4: AI-Generated Content: Thermometer Gap Results

	Treatment Only	Treatment + Covariates	Full Model
(Intercept)	59.184***	41.186***	46.200***
	(1.792)	(9.825)	(9.964)
AI Treatment	1.296	−1.144	−10.266+
	(2.596)	(2.430)	(5.829)
AI Treatment:Green			5.845
			(8.489)
AI Treatment:Labour			10.807
			(6.863)
AI Treatment:Lib Dem			17.988*
			(7.580)
AI Treatment:Other			25.890
			(18.232)
AI Treatment:Reform UK			10.710
			(8.344)
AI Treatment:SNP			25.815+
			(15.208)
Num.Obs.	554	479	479
R2	0.001	0.173	0.185
RMSE	28.78	25.80	25.60
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Note: Models weighted using YouGov survey weights. The coefficients are reported with robust standard errors in parentheses. Main effects of the included moderators are also reported as rows above the moderator treatment effects.

Figure 2: Thermometer Score Patchwork Plot for AI-Generated Content

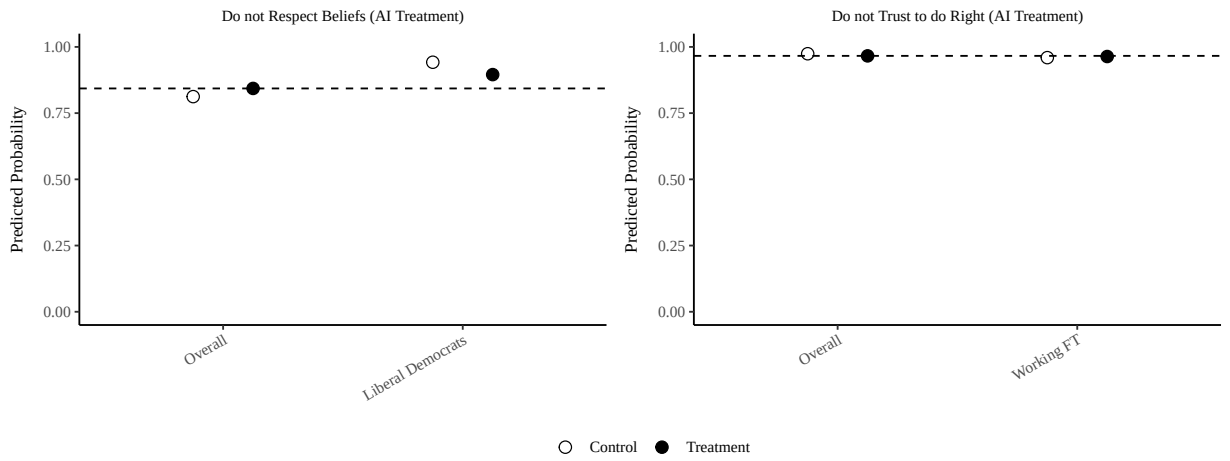


5.2.2 Ordinal Affective Polarisation Analysis

To get a better picture of the treatment effects of AI-generated content on the affective polarisation, models for the outcome variables of `agreedisagree`, `xtrust`, and `child` are also estimated. These models are ordinal logistic regression models, as the outcome variables are ordinal measures. The results of these models for each measure of *respect*, *trust*, and emotional, social-distance measures of *discomfort* towards opposing partisans are presented in full in Table 13, Table 14, and Table 15 respectively in the Appendix. None of the full models — estimated with relevant covariates and moderators — show any significant overall treatment effect of AI-generated content on the ordinal measures of affective polarisation, again giving credence to the viscocity of political attitudes and even a particularly divisive AI-generated article not being able to further polarise respondents' views.

Table 10.

Figure 3: Predicted Probabilities by Subgroup for AI-Generated Content



These results show that a single exposure to an AI-generated article does not significantly polarise UK respondents towards their political opponents. The treatment does not significantly polarise thermometer ratings, nor the trait- and emotional-based measures of *respect*, *trust*, and *discomfort* towards out-groips. While this appears to be an encouraging null result to help dampen fears towards AI-generated content, there are some subgroups who show significant susceptibility to polarisation. Noteably, Liberal Democrats show increase warmth towards their in-party leader; whereas, low-educated, part-time-working respondents show increased discomfort towards out-party voters. Explanations and theoretical implications of these results are to be investigated.

5.3 AI-Labelled Content Treatment

Building on Altay and Gilardi (2024), the AI-label treatment is designed to test whether the labelling of content as AI-generated can mitigate the polarising effects of AI-generated content, or whether the association with AI-generated content is enough to polarise respondents. The treatment group is shown the same article as the AI-generated content group, but labelled as AI-generated. The control group is shown the same article but labelled as human-generated. The treatment and control groups are compared to see if there is a significant difference in the thermometer scores, trait-based measures of affective polarisation, and emotional discomfort measures.

The results show that there is no significant treatment effect of AI-labelled content on thermometer scores, nor trait-based measures of affective polarisation. However, there is a significant treatment effect on the emotional discomfort measure of having a child marry an out-party voter. This suggests that labelling content as AI-generated can help mitigate the polarising effects of AI-generated content.

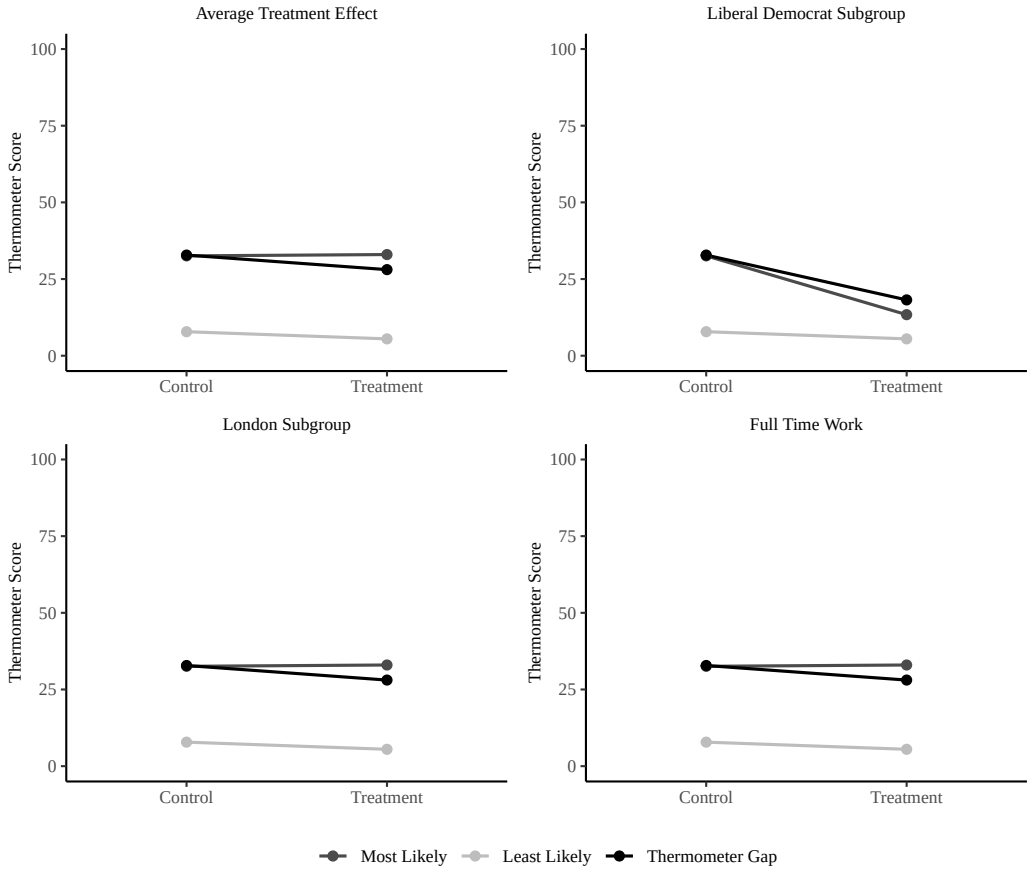
Table 5: AI-Labelled Content: Thermometer Gap Results

	Treatment Only	Treatment + Covariates	Full Model
(Intercept)	60.479***	32.546**	32.816**
	(1.879)	(10.460)	(11.730)
Label Treatment	−2.393	−1.331	−4.743
	(2.622)	(2.486)	(9.279)
Label Treatment:Greens			5.640
			(11.715)
Label Treatment:Labour			−6.368
			(6.634)
Label Treatment:LibDem			−9.873
			(8.261)
Label Treatment:Reform			−3.825
			(7.315)
Label Treatment:Political Attention			1.050
			(1.146)
Num.Obs.	543	470	470
R2	0.002	0.224	0.258
RMSE	28.24	25.53	24.94
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Note: Models weighted using YouGov survey weights. The coefficients are reported with robust standard errors in parentheses. Main effects of the included moderators are also reported as rows above the moderator treatment effects.

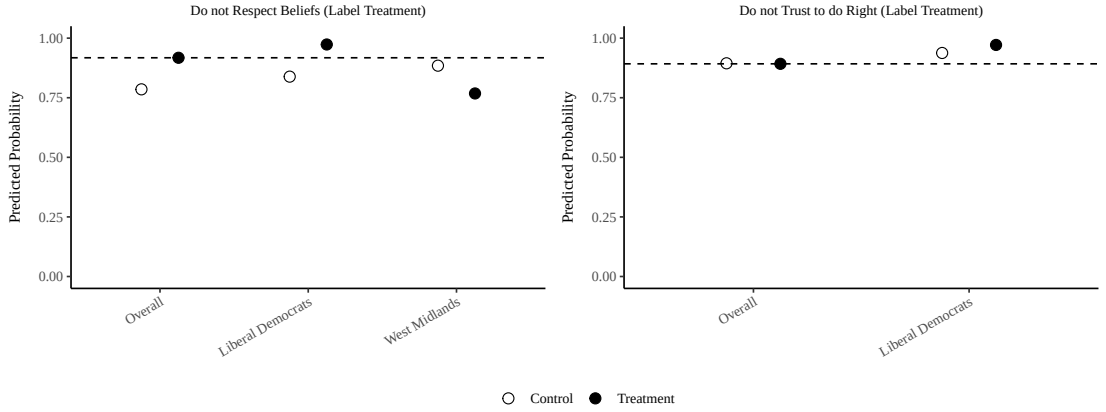
5.3.1 Thermometer Analysis

Figure 4: Thermometer Score Patchwork Plot for AI-Labelled Content



5.3.2 Ordinal Affective Polarisation Analysis

Figure 5: Predicted Probabilities by Subgroup for AI-Labelled Content



5.4 Additional Analysis

5.4.1 Causal Acyclic Testing

(see exeprimental analysis week 6 notes for details)

5.4.2 Agentic-based Modelling

- What is agentic-based modelling?
- Why is agentic-based modelling important?
- How will I use agentic-based modelling?

Appendix

5.5 Codebook

The codebook in Table **??** 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 6: (#tab:codebook-table)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)

Table 6: (#tab:codebook-table)YouGov UniOM Survey Codebook *(continued)*

Variable	Type	Description	Values
`political_attention`	Continuous	How much attention the respondent pays to politics	Scale (e.g., 0–10 or continuous 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

Table 6: (#tab:codebook-table)YouGov UniOM Survey Codebook *(continued)*

Variable	Type	Description	Values
`LLthermo_NF`	Continuous	Thermometer rating for Nigel Farage	0–100
`LLthermo_ED`	Continuous	Thermometer rating for Ed Davey	0–100
`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)

5.6 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.

5.7 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. Table 7 and Table 8 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.

5.8 Sensitivity Analysis

Given the nature of the results often being reported as null effects, a sensitivity analysis to determine what the smallest true effect that could have detected 80% of the time is calculated.

5.9 `MLthermoMean` and `LLthermoMean` Analysis

The models for the outcome variables of `MLthermoMean` and `LLthermoMean` are estimated using the same model specification as for `thermo_gap` in Table 4. These models are presented in Table 9 and Table 10 respectvely.

For the `label_treatment` models, the same model specification is used as for the `ai_treatment` models. The models for the outcome variables of `MLthermoMean` and `LLthermoMean` are estimated using the same model specification as for `thermo_gap` in Table 5. These models are presented in Table 11 and Table 12 respectvely.

5.10 Ordinal Affective Polarisation Results

5.10.1 AI-Generated Content

The following results presented in Table 13, Table 14, and Table 15 show the log-odds change in the probability of being in a higher level (a higher threshold cut point) of agreement, trust, or comfort respectively.

Table 7: (#tab:ai-balance)Balance Table of Covariates by AI Treatment Group

Variable	Control	Treatment	p-value	Signif.
Age	52.12 (16.74)	51.56 (16.75)	0.521	-
Political attention	6.69 (1.92)	6.61 (1.98)	0.452	-
Gender (male)	374 (50.1)	412 (54.8)	0.080	-
Female	372 (49.9)	340 (45.2)	NA	
Education level (High)	308 (41.3)	308 (41.0)	0.882	-
Low	151 (20.2)	160 (21.3)	NA	
Medium	287 (38.5)	284 (37.8)	NA	
Employment status (Full time student)	31 (4.2)	35 (4.7)	0.759	-
Not working	32 (4.3)	37 (4.9)	NA	
Other	16 (2.1)	13 (1.7)	NA	
Retired	222 (29.8)	210 (27.9)	NA	
Unemployed	12 (1.6)	21 (2.8)	NA	
Working full time (30 or more hours per week)	327 (43.8)	338 (44.9)	NA	
Working part time (8-29 hours a week)	94 (12.6)	87 (11.6)	NA	
Working part time (Less than 8 hours a week)	12 (1.6)	11 (1.5)	NA	
Voted in 2024 General Election (Don't know)	3 (0.4)	1 (0.1)	0.574	-
No, did not vote	97 (13.0)	102 (13.6)	NA	
Yes, voted	646 (86.6)	649 (86.3)	NA	
Vote in 2024 General Election (Conservative)	162 (25.1)	143 (22.0)	0.587	-
Don't know	2 (0.3)	6 (0.9)	NA	
Green	58 (9.0)	51 (7.9)	NA	
Labour	211 (32.7)	245 (37.8)	NA	
Liberal Democrat	90 (13.9)	84 (12.9)	NA	
Other	13 (2.0)	12 (1.8)	NA	
Plaid Cymru	2 (0.3)	2 (0.3)	NA	
Reform UK	98 (15.2)	96 (14.8)	NA	
Scottish National Party (SNP)	9 (1.4)	10 (1.5)	NA	
Skipped	1 (0.2)	0 (0.0)	NA	
Vote in EU Referendum (Can't remember)	125 (17.0)	132 (17.8)	0.669	-
I did not vote	287 (39.0)	273 (36.8)	NA	
I voted to Leave	323 (43.9)	337 (45.4)	NA	
Region (East Midlands)	49 (6.6)	61 (8.1)	0.376	-
I voted to Remain	89 (11.9)	79 (10.5)	NA	
East of England	94 (12.6)	73 (9.7)	NA	
London	34 (4.6)	26 (3.5)	NA	
North East	83 (11.1)	84 (11.2)	NA	
North West	44 (5.9)	64 (8.5)	NA	
Scotland	109 (14.6)	120 (16.0)	NA	
South East	79 (10.6)	70 (9.3)	NA	
South West	31 (4.2)	35 (4.7)	NA	
Wales	62 (8.3)	66 (8.8)	NA	
West Midlands	72 (9.7)	74 (9.8)	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 8: (#tab:ai-balance)Balance Table of Covariates by Label Treatment Group

Variable	Control	Treatment	p-value	Signif.
Age	51.84 (16.62)	51.84 (16.88)	0.996	-
Political attention	6.58 (1.94)	6.71 (1.96)	0.200	-
Gender (male)	408 (54.0)	378 (50.9)	0.240	-
Female	347 (46.0)	365 (49.1)	NA	
Education level (High)	321 (42.5)	295 (39.7)	0.542	-
Low	153 (20.3)	158 (21.3)	NA	
Medium	281 (37.2)	290 (39.0)	NA	
Employment status (Full time student)	31 (4.1)	35 (4.7)	0.966	-
Not working	37 (4.9)	32 (4.3)	NA	
Other	16 (2.1)	13 (1.7)	NA	
Retired	213 (28.2)	219 (29.5)	NA	
Unemployed	19 (2.5)	14 (1.9)	NA	
Working full time (30 or more hours per week)	338 (44.8)	327 (44.0)	NA	
Working part time (8-29 hours a week)	90 (11.9)	91 (12.2)	NA	
Working part time (Less than 8 hours a week)	11 (1.5)	12 (1.6)	NA	
Voted in 2024 General Election (Don't know)	2 (0.3)	2 (0.3)	0.154	-
No, did not vote	113 (15.0)	86 (11.6)	NA	
Yes, voted	640 (84.8)	655 (88.2)	NA	
Vote in 2024 General Election (Conservative)	148 (23.1)	157 (24.0)	0.927	-
Don't know	4 (0.6)	4 (0.6)	NA	
Green	55 (8.6)	54 (8.2)	NA	
Labour	233 (36.4)	223 (34.0)	NA	
Liberal Democrat	85 (13.3)	89 (13.6)	NA	
Other	10 (1.6)	15 (2.3)	NA	
Plaid Cymru	2 (0.3)	2 (0.3)	NA	
Reform UK	96 (15.0)	98 (15.0)	NA	
Scottish National Party (SNP)	7 (1.1)	12 (1.8)	NA	
Skipped	0 (0.0)	1 (0.2)	NA	
Vote in EU Referendum (Can't remember)	131 (17.6)	126 (17.2)	0.490	-
I did not vote	272 (36.5)	288 (39.4)	NA	
I voted to Leave	343 (46.0)	317 (43.4)	NA	
Region (East Midlands)	56 (7.4)	54 (7.3)	0.700	-
I voted to Remain	78 (10.3)	90 (12.1)	NA	
East of England	84 (11.1)	83 (11.2)	NA	
London	32 (4.2)	28 (3.8)	NA	
North East	86 (11.4)	81 (10.9)	NA	
North West	57 (7.5)	51 (6.9)	NA	
Scotland	116 (15.4)	113 (15.2)	NA	
South East	80 (10.6)	69 (9.3)	NA	
South West	28 (3.7)	38 (5.1)	NA	
Wales	72 (9.5)	56 (7.5)	NA	
West Midlands	66 (8.7)	80 (10.8)	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 9: AI-Generated Content: Thermometer (mostlikely) Results

	Treatment Only	Treatment + Covariates	Full Model
(Intercept)	69.079***	41.416***	46.928***
	(1.520)	(8.441)	(8.660)
AI Treatment	0.425	−0.717	−8.282
	(2.131)	(2.105)	(8.919)
age		0.109	0.108
		(0.107)	(0.106)
AI Treatment:Not Working			13.551
			(12.962)
AI Treatment:Other			19.183
			(12.961)
AI Treatment:Retired			4.055
			(9.718)
AI Treatment:Unemployed			10.120
			(12.380)
AI Treatment:Working Full Time			11.093
			(9.500)
AI Treatment:Working PT (8–29h)			−2.782
			(10.536)
AI Treatment:Working PT (<8h)			44.718***
			(11.350)
Num.Obs.	634	542	542
R2	0.000	0.160	0.175
RMSE	22.78	20.66	20.54
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Note: Models weighted using YouGov survey weights. The coefficients are reported with robust standard errors in parentheses. Main effects of the included moderators are also reported as rows above the moderator treatment effects.

Table 10: AI-Generated Content: Thermometer (leastlikely) Results

	Treatment Only	Treatment + Covariates	Full Model
(Intercept)	10.499***	0.140	−5.312
	(1.034)	(5.290)	(5.723)
AI Treatment	−1.142	0.301	9.687*
	(1.450)	(1.505)	(4.704)
AI Treatment:Age			−0.183*
			(0.087)
Num.Obs.	647	549	549
R2	0.001	0.096	0.105
RMSE	16.57	15.45	15.42
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Note: Models weighted using YouGov survey weights. The coefficients are reported with robust standard errors in parentheses. Main effects of the included moderators are also reported as rows above the moderator treatment effects.

5.10.2 AI-Labelled Content

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.

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.

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

Table 11: AI-Labelled Content: Thermometer (mostlikely) Results

	Treatment Only	Treatment + Covariates	Full Model
(Intercept)	69.504***	32.728***	32.559**
	(1.494)	(9.153)	(10.056)
Label Treatment	−3.251	−2.357	0.427
	(2.168)	(2.080)	(10.447)
age		0.256*	0.279*
		(0.103)	(0.113)
Label Treatment:Age			−0.097
			(0.121)
Label Treatment:Political Attention			1.899+
			(0.976)
Label Treatment:Greens			−8.272
			(10.014)
Label Treatment:Labour			−15.070**
			(5.507)
Label Treatment:LibDem			−19.598**
			(6.625)
Label Treatment:Reform			−9.345+
			(5.612)
Num.Obs.	626	541	541
R2	0.005	0.187	0.245
RMSE	23.57	21.76	21.18
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Note: Models weighted using YouGov survey weights. The coefficients are reported with robust standard errors in parentheses. Main effects of the included moderators are also reported as rows above the moderator treatment effects.

Table 12: AI-Labelled Content: Thermometer (leastlikely) Results

	Treatment Only	Treatment + Covariates	Full Model
(Intercept)	9.357***	5.785	7.826
	(1.017)	(7.242)	(7.730)
Label Treatment	1.264	0.420	−2.334
	(1.677)	(1.730)	(2.337)
Label Treatment:Male			5.420
			(3.806)
Num.Obs.	643	547	547
R2	0.001	0.111	0.116
RMSE	16.01	15.41	15.39
Model	(1)	(2)	(3)

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Note: Models weighted using YouGov survey weights. The coefficients are reported with robust standard errors in parentheses. Main effects of the included moderators are also reported as rows above the moderator treatment effects.

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.

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.

Table 13: AI-Generated Content: Agree Out-Party Respect Beliefs

	Treatment Only	Treatment + Covariates	Full Model
AI Treatment	−0.305	−0.306+	−0.538
	(0.185)	(0.185)	(0.342)
Strongly disagree|Tend to disagree	0.557***	−1.195*	−2.347***
	(0.126)	(0.594)	(0.624)
Tend to disagree|Neither agree nor disagree	1.842***	0.147	−0.947
	(0.163)	(0.588)	(0.620)
Neither agree nor disagree|Tend to agree	3.397***	1.726**	0.646
	(0.286)	(0.605)	(0.617)
Tend to agree|Strongly agree	4.615***	2.950***	1.867**
	(0.450)	(0.721)	(0.713)
mostlikelyGreen Party			−1.176**
			(0.364)
mostlikelyLabour Party			−1.085**
			(0.360)
mostlikelyLiberal Democrats			−1.528***
			(0.410)
mostlikelyReform UK			−0.551
			(0.338)
AI Treatment:mostlikelyGreen Party			0.089
			(0.662)
AI Treatment:mostlikelyLabour Party			−0.079
			(0.542)
AI Treatment:mostlikelyLiberal Democrats			1.437**
			(0.548)
AI Treatment:mostlikelyReform UK			0.026
			(0.533)
Num.Obs.	658	658	658
edf	5	17	25
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 included but have no substantive interpretation.

Table 14: AI-Generated Content: Trust in Out-Party to Do What Is Right

	Treatment Only	Treatment + Covariates	Full Model
AI Treatment	−0.188	−0.156	1.598
	(0.193)	(0.187)	(1.008)
Almost never|Once in a while	0.873***	0.658	1.934*
	(0.137)	(0.671)	(0.927)
Once in a while|About half of the time	2.846***	2.678***	3.966***
	(0.214)	(0.652)	(0.936)
About half of the time|Always	3.990***	3.820***	5.107***
	(0.385)	(0.760)	(1.033)
Always|Most of the time	4.146***	3.977***	5.264***
	(0.425)	(0.779)	(1.018)
AI Treatment:Not Working			−2.406+
			(1.306)
AI Treatment:Other			−2.388+
			(1.331)
AI Treatment:Retired			−1.434
			(1.050)
AI Treatment:Unemployed			−3.604+
			(1.936)
AI Treatment:Working Full Time			−2.022+
			(1.054)
AI Treatment:Working PT (8–29h)			−1.420
			(1.204)
AI Treatment:Working PT (<8h)			−2.019
			(1.911)
Num.Obs.	664	664	664
edf	5	17	24
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 included but have no substantive interpretation.

Table 15: AI-Generated Content: Comfort with Child Marrying Opposing Partisan

	Treatment Only	Treatment + Covariates	Full Model
AI Treatment	−0.015	0.060	0.446+
	(0.157)	(0.163)	(0.230)
AI Treatment:Education LevelLow			−1.024*
			(0.464)
AI Treatment:Education LevelMedium			−0.475
			(0.356)
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 included but have no substantive interpretation.

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.

Table 16: AI-Labelled Content: Agree Out-Party Respect Beliefs

	Treatment Only	Treatment + Covariates	Full Model
Label Treatment	0.007	0.007	−1.317
	(0.200)	(0.200)	(0.806)
Label Treatment:RegionEast of England			1.771+
			(0.990)
Label Treatment:RegionLondon			1.860+
			(0.989)
Label Treatment:RegionNorth East			1.654
			(1.187)
Label Treatment:RegionNorth West			2.217*
			(0.920)
Label Treatment:RegionScotland			0.996
			(0.950)
Label Treatment:RegionSouth East			1.797*
			(0.873)
Label Treatment:RegionSouth West			0.460
			(1.050)
Label Treatment:RegionWales			0.528
			(1.219)
Label Treatment:RegionWest Midlands			2.579**
			(0.996)
Label Treatment:RegionYorkshire and the Humber			1.866+
			(0.994)
Label Treatment:mostlikelyGreen Party			−1.330
			(0.857)
Label Treatment:mostlikelyLabour Party			0.633
			(0.576)
Label Treatment:mostlikelyLiberal Democrats			−1.082+
			(0.581)
Label Treatment:mostlikelyReform UK			−0.580
			(0.547)
Num.Obs.	669	669	669
edf	5	5	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 agreement that opposing partisans respect political beliefs. Threshold cutpoints are included but have no substantive interpretation.

Table 17: AI-Labelled Content: Trust in Out-Party to Do What Is Right

	Treatment Only	Treatment + Covariates	Full Model
Label Treatment	0.082	0.082	0.221
	(0.198)	(0.198)	(0.366)
Almost never|Once in a while	1.066***	1.066***	0.506+
	(0.139)	(0.139)	(0.270)
Once in a while|About half of the time	2.951***	2.951***	2.435***
	(0.191)	(0.191)	(0.311)
About half of the time|Always	4.212***	4.212***	3.702***
	(0.350)	(0.350)	(0.426)
Always|Most of the time	4.739***	4.739***	4.229***
	(0.380)	(0.380)	(0.457)
Label Treatment:mostlikelyGreen Party			−0.215
			(0.798)
Label Treatment:mostlikelyLabour Party			0.025
			(0.556)
Label Treatment:mostlikelyLiberal Democrats			−1.097+
			(0.634)
Label Treatment:mostlikelyReform UK			0.020
			(0.555)
Num.Obs.	678	678	678
edf	5	5	13
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 included but have no substantive interpretation.

Table 18: AI-Labelled Content: Comfort with Child Marrying Opposing Partisan

	Treatment Only	Treatment + Covariates	Full Model
Label Treatment	0.032	−0.106	−1.850*
	(0.165)	(0.187)	(0.742)
Label Treatment:GenderMale			1.172**
			(0.408)
Label Treatment:RegionEast of England			2.312**
			(0.883)
Label Treatment:RegionLondon			1.717+
			(0.977)
Label Treatment:RegionNorth East			0.780
			(1.247)
Label Treatment:RegionNorth West			2.553*
			(1.000)
Label Treatment:RegionScotland			0.350
			(0.890)
Label Treatment:RegionSouth East			0.858
			(0.846)
Label Treatment:RegionSouth West			2.039*
			(0.914)
Label Treatment:RegionWales			1.139
			(1.000)
Label Treatment:RegionWest Midlands			2.250*
			(0.933)
Label Treatment:RegionYorkshire and the Humber			2.086*
			(0.886)
Label Treatment:EU VoteI did not vote			−0.634
			(0.615)
Label Treatment:EU VoteI voted to Leave			−0.890*
			(0.427)
Num.Obs.	699	595	595
RMSE	2.33	2.31	2.31
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 included but have no substantive interpretation.

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.

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.

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.

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/.

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.

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

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/.