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

2 Literature Review

3 Theoretical Framework

3.1 Theory and Argumentation

- Develop the initial arguments and theoretical framework of your project.
- Discuss how your project relates to existing theoretical approaches in the literature and how these are further developed and/or applied in your research.
- This theoretical framework will most likely only be at the preliminary stage, but it is important to outline the relationships between the key actors or variables in your project.
- This 'model' does not have to be formal and explicit, but you may find it helpful to specify causal relationships in terms of dependent variable(s) (the outcome) and independent variables (the explanatory factors).
- Formulate some preliminary testable hypotheses derived from this 'model'.
- Outline the key assumptions of your argument, as well as the limits to your topic (temporally and spatially).

3.2 Hypotheses

- The hypotheses provided should state the relationship(s) expected to be observed between variables being used
- The formulation of a hypothesis should make clear whether it involves one- or two-tailed tests (i.e. predict an increase, decrease, or change in the outcome variable)
- There are two types of hypotheses to consider:
  - **Confirmatory Hypotheses**
    - The main focus of the study → what the study is designed to test
    - There should be well-powered analyses of this hypothesis
    - Should be backed up by strong theory leading to hypotheses *a priori*
  - **Exploratory Hypotheses**
    - Hypotheses wish to test but are not the main focus of the study
    - Often seen as secondary hypotheses looking at mechanisms, subgroups, heterogeneous effects, or downstream outcomes

- Should include as many hypotheses as relate to your theory or intervention
- With more than one hypothesis you will need to specify a procedure for handling multiple hypotheses in the inference criteria section of your PAP

Use the paper Affect, Not Ideology: A Social Identity Perspective on Polarization to look at theory. As is The Origins and Consequences of Affective Polarization in the United States. This has a lot of criqtues of the measures of polarization too.

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.[1] 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).[2] 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- 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.[3]

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

[1] 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."

[2] 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.

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

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.

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 (1):

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 \mathbf{X}_i + \beta_3 (D_i \times \mathbf{Z}_i) + \varepsilon_i \tag{1}$$

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.

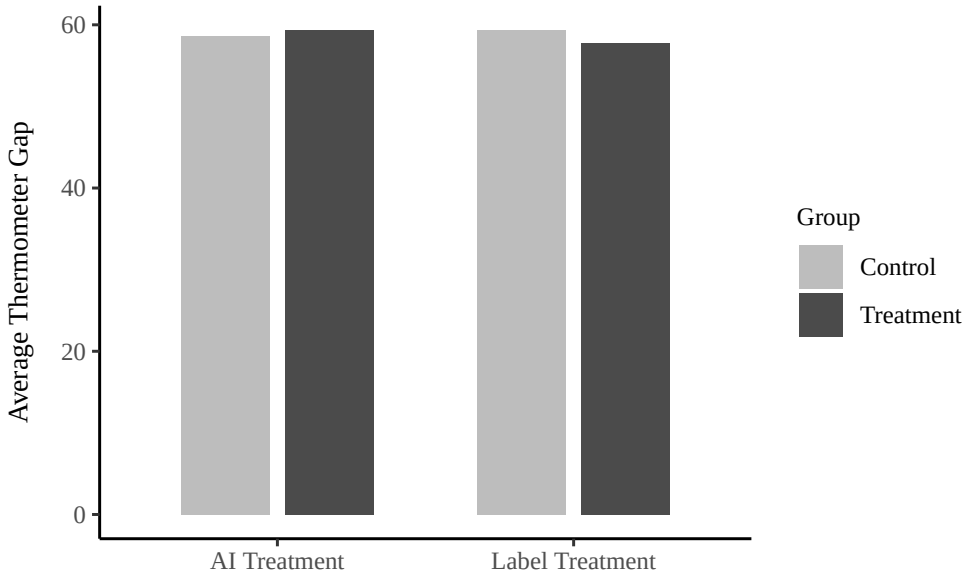
5.2 AI-Generated Content Treatment

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.

Table 1: AI-Generated Content: Thermometer Gap Results

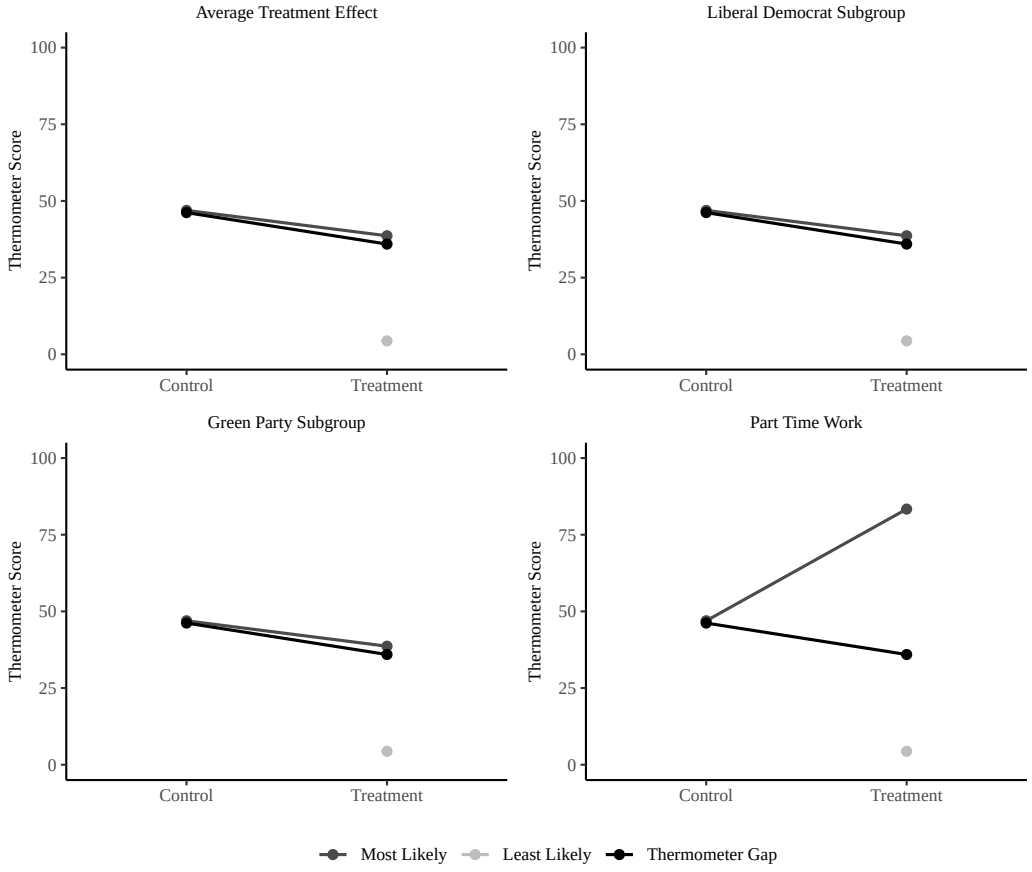
	Treatment Only	Treatment + Covariates	Full Model
(Intercept)	59.184***	41.186***	46.200***
	(1.792)	(9.825)	(9.964)
AI Treatment	1.296	−1.144	−10.266+
	(2.596)	(2.430)	(5.829)
AI Treatment:Green			5.845
			(8.489)
AI Treatment:Labour			10.807
			(6.863)
AI Treatment:Lib Dem			17.988*
			(7.580)
AI Treatment:Other			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.

ATE models are presented in Table 1 for the outcome `thermo_gap`.[4] 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.

[4]The full models for the outcome variables of `MLthermoMean` and `LLthermoMean` are available in the appendix in Table 6 and Table 7.

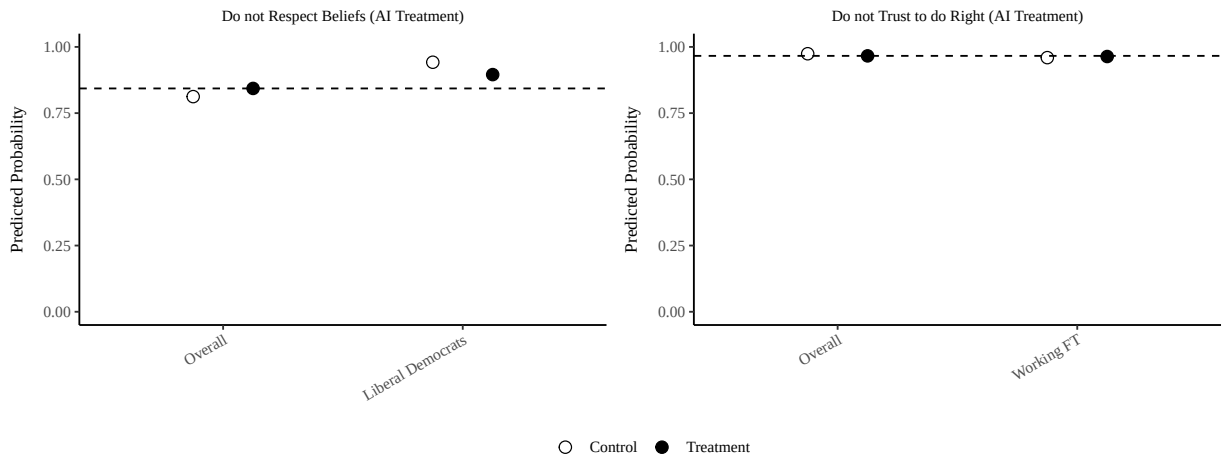
Figure 2: Thermometer Score Patchwork Plot for AI-Generated Content



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 10, Table 11, and Table 12 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.

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.

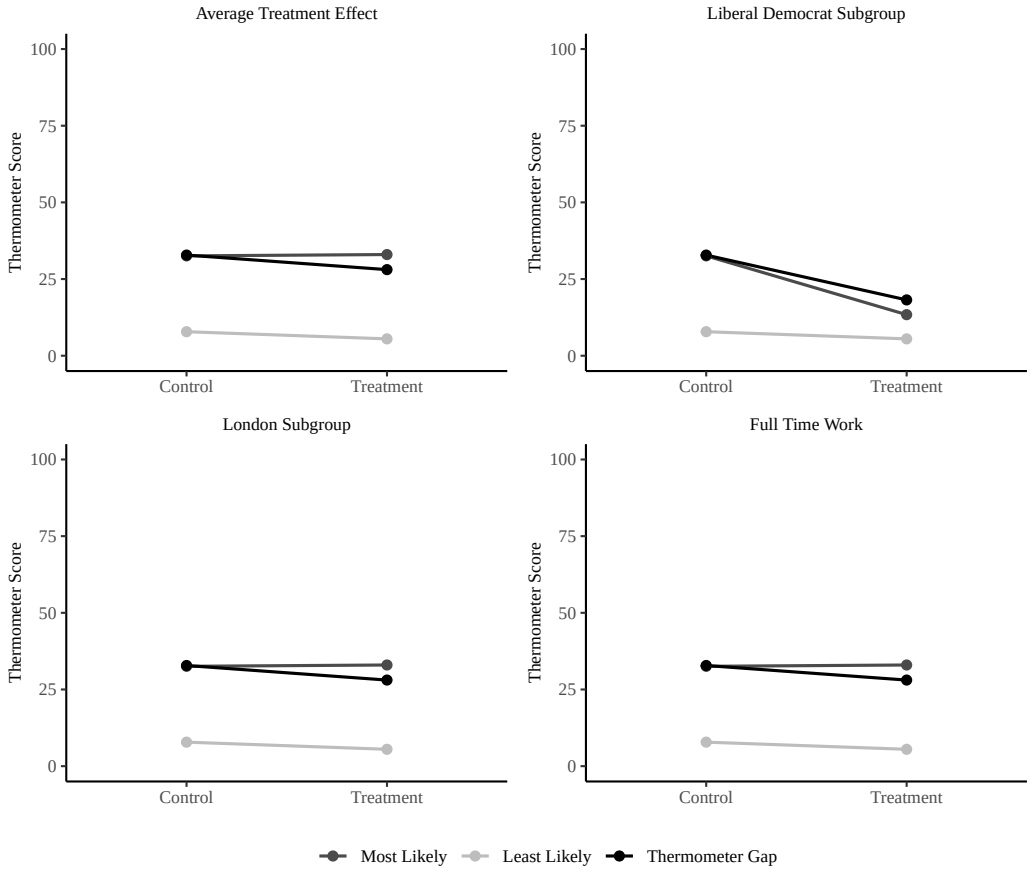
Table 2: AI-Labelled Content: Thermometer Gap Results

	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.

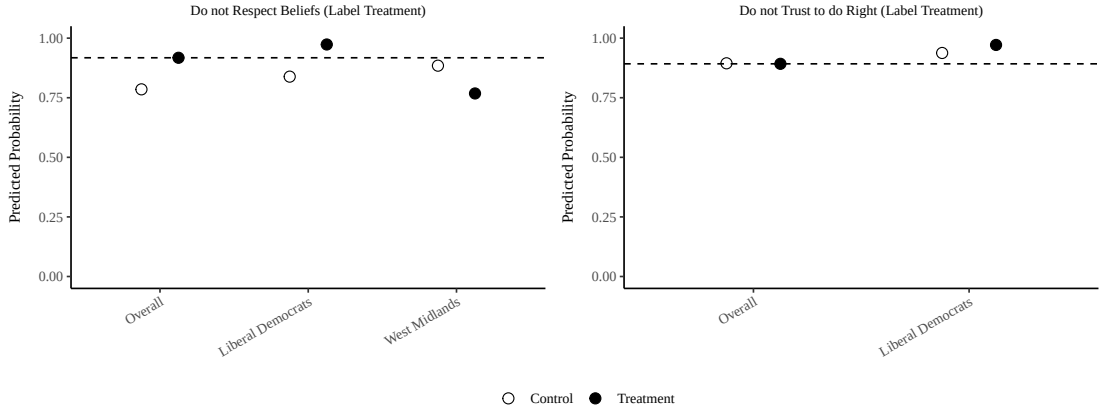
5.3.1 Thermometer Analysis

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 3: (#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 3: (#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 3: (#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 4 and Table 5 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 1. These models are presented in Table 6 and Table 7 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 2. These models are presented in Table 8 and Table 9 respectvely.

5.10 Ordinal Affective Polarisation Results

5.10.1 AI-Generated Content

The following results presented in Table 10, Table 11, and Table 12 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 4: (#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 5: (#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 6: 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 7: 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

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.

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.

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.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N. and Westwood, S.J. (2019) 'The Origins and Conse-

Table 8: 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 9: 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.

quences 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.

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].

Table 10: 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 11: 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 12: 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.

Table 13: 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 14: 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 15: 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.