PUSH*BACK*LASH

Funded by the European Union

UK Research and Innovation

DELIVERABLE D6.2

Integrated Report on Counter-Strategies to Mitigate the Impact of Anti-Gender and Anti-Feminist Discourse on Citizens

11.12.2025

Work Package WP6: The Effects of Anti-Gender Discourse and Counter Strategies

Authors: Florian Stoeckel, Monika Kovacs, Karolina Pieklo, Eszter Csordás, Ettore Pierre Sansivero
Contributor: Francesca Mignogna
Reviewer: Rebekka Kesberg

This project has received funding from the European Union's Horizon 2021 research and innovation program under grant agreement N° 101061687 and from the UKRI (grant number 10038117).

About Push*Back*Lash

Project: 101061687 — PushBackLash — HORIZON-CL2-2021-DEMOCRACY-01

Project summary:
The goal of PUSH*BACK*LASH is twofold: (1) Firstly, to systematically inquire into the present-day contestation of gender equality issues and policies at both elite and citizen levels. Approaching democracy from a global feminist perspective, we employ a rigorous, comparative, multi-method design (e.g., experiments, surveys, interviews, participatory theatre). Our project enables: (a) identifying anti-gender strategies as well as best practices in counteracting them across space and time; and (b) assessing the effects of anti-gender discourses by focusing on parties, social media, and public opinion. Secondly, to (2) develop and test strategies that can effectively counteract anti-gender and anti-feminist discursive strategies. Aiming at supporting the quality of democratic governance in more inclusive European societies, we acknowledge intersections between gender and other social categories at all stages of the project (composition of consortium and advisory board; theory formulation, empirical investigation, and policy recommendations) and thus engage with stakeholders. To develop sustainable solutions, we bring together gender activists, EU experts, and researchers from several fields of political science (political theory, public policy, political parties, public opinion, political behaviour), anthropology, communication and media, philosophy, sociology, and social psychology. PUSH*BACK*LASH is a transdisciplinary, gender-diverse consortium aiming at equipping pro-equality actors with practical toolkits for responding to anti-gender equality and anti-feminist discursive strategies and backlash tactics.

Keywords:

- Anti-gender discourse
- Survey experiment
- Counter-argument strategies
- Counterspeech
- Gender studies
- Gender equality
- Hate speech
- Poland
- Hungary
- Italy
- Communication
- Persuasion
- Misogyny
- Social media
- Threat perception

Document versions

Version	Date	Created/modified by	Comments / Changes
1	6.8.2025	Florian Stoeckel	Creation of document, structure, and first draft.
2	18.8.2025	Monika Kovacs	Revision
3	19.8.2025	Karolina Pieklo	Revision
4	20.8.2025	Eszter Csordas	Revision
5	22.8.2025	Florian Stoeckel	Revision
6	26.8.2025	Monika Kovacs	Revision
7	28.8.2025	Eszter Csordas	Revision
8	31.8.2025	Florian Stoeckel	Revision
9	11.9.2025	Monika Kovacs	Revision
10	7.10.2025	Florian Stoeckel	Revision
11	22.10.2025	Karolina Pieklo	Revision
12	13.11.2025	Florian Stoeckel	Revision
13	14.11.2025	Florian Stoeckel	Revision
14	20.11.2025	Florian Stoeckel	Revision based on reviewer comments
15	4.12.2025	Florian Stoeckel	Revision based on reviewer comments
16	5.12.2025	Ettore Pierre Sansivero	Revision: numbering of figures, tables etc.
17	8.12.2025	Monika Kovacs	Revision & checking
18	10.12.2025	Florian Stoeckel, Monika Kovacs, Ettore Pierre Sansivero	Revision
19	11.12.2025	Florian Stoeckel, Ettore Pierre Sansivero	Revision & final check

Table of Contents

About Push*Back*Lash 2
Document versions 3
List of Tables 5
List of Figures 6
About this document 7
List of Definitions and Abbreviations 8
Executive summary 9
1. Introduction 12
2. Country level context: Wave-1 14
3. Counter-argument strategies: rationale 20
3.1 The role of counter-argument strategies in challenging anti-gender egalitarian communication 21
3.2 The role of counter-arguments for norm perceptions 25
4. The experimental research design 27
4.1 The anti-gender discourse: original material 32
4.2 Testing of counter-arguments 35
5. Sample, fieldwork, descriptive results 37
6. Results 39
6.1 Stability and change in attitudes 40
6.1.1 Stability of attitudes across waves 41
6.1.2 Experimental treatment effects on attitudes 42
6.1.3 Heterogeneity of experimental treatment effects 44
6.2 Post-level results: counter-arguments strategies 47
6.3 Implications 56
7. Limitations, discussion, conclusion 58
References 61
Appendix A: Measures 73
Appendix B: Additional Results 75
Appendix C: Survey Questionnaire 94
Appendix D: Pre-registration 109

List of Tables

Table 1 Field work dates across the countries and observations per country 14
Table 2 Social media posts content and sources 34
Table 3 Field work dates and N per wave and country 37
Table 4. Modern sexism item mean score across survey waves by condition 42

Table A 1. OLS estimates for the outcome variable "Discrimination against women is no longer a problem in a country" - Hungary 78
Table A 2. OLS estimates for the outcome variable "Discrimination against women is no longer a problem in a country" - Poland 80
Table A 3. OLS estimates for the outcome variable "Discrimination against women is no longer a problem in a country"– Italy 81
Table A 4. OLS estimates for the outcome variable "Discrimination against women is no longer a problem in a country" - Hungary 82
Table A 5. OLS estimates for the outcome variable "Discrimination against women is no longer a problem in a country" - Poland 84
Table A 6. OLS estimates for the outcome variable "Discrimination against women is no longer a problem in a country" - Italy 86
Table A 7. OLS for gender system justification battery as the outcome - Hungary 88
Table A 8. OLS for gender system justification battery as the outcome - Poland 89
Table A 9. Pooled OLS for gender system justification battery as the outcome - Italy 90
Table A 10. Pooled OLS for the attitude towards gender equality as the outcome - Hungary 91
Table A 11. OLS, attitude towards gender equality as the outcome - Poland 92
Table A 12. OLS, attitude towards gender equality as outcome - Italy 93

List of Figures

Figure 1 Misogyny and zero-sum scale by country and gender 16
Figure 2 Perspective-taking scale and participation by country and gender 16
Figure 3. Levels of misogyny vs far-right vote share and gender inequality 17
Figure 4. Gender-stereotype item plotted against far-right party vote share and gender inequality 17
Figure 5. Modern sexism item plotted against far-right party vote share and gender inequality . 18
Figure 6. Levels of zero-sum thinking vs far-right party vote share and gender inequality..... 18
Figure 7. Levels of perspective taking vs far-right party vote share and gender inequality..... 19
Figure 8. Levels of willingness to participate vs far-right party vote share and gender inequality 19
Figure 9. Survey design 29
Figure 10. Example of the screen that respondents in the control track saw..... 30
Figure 11. Example of the screen that respondents in the treatment track saw. 31
Figure 12 Perceptions of counter-argument strategies in the pilot study..... 36
Figure 13 Descriptive evaluations of anti-gender posts across detection, agreement, liking, and perceived norms (control condition)..... 38
Figure 14. Effect of receiving treatment on gender system justification..... 43
Figure 15. Effect of receiving treatment on attitude towards gender equality..... 44
Figure 16. Predicted values (Wave-2) by treatment condition. 45
Figure 17 Marginal effect of treatment across the distribution of Wave-1 values. 46
Figure 18. Effect of treatment on detection of sexism by strategy averaged across posts 49
Figure 19. Effects of treatment on detection of sexism by post and strategy 50
Figure 20. Effects of treatment on agreement with anti-gender content by strategy, averaged across posts 52
Figure 21. Effects of treatment on agreement with anti-gender content by post and strategy 53
Figure 22. Effects of treatment on perceived normativity by strategy, averaged across posts..... 55
Figure 23. Effects of treatment on perceived normativity by post and strategy 56

Figure A 1. Marginal effect plots of treatment on agreement with statement "Discrimination against women is no longer a problem in a country"..... 75
Figure A 2. Marginal effect plots of treatment on gender system justification battery as outcome 76
Figure A 3. Marginal effects plots of treatment on attitude towards gender equality battery 77

About this document

Description:

This integrated report presents the main findings of Work Package 6 (WP6) of the PUSH*BACK*LASH project, which investigates the effects of anti-gender egalitarian discourse on citizens and evaluates the potential of counter-strategies to mitigate these effects. The report combines large-N cross-national survey evidence (Wave-1) from eight European countries with a preregistered experiment (Wave-2) conducted in Hungary, Italy, and Poland.

The document examines:
(1) the attitudinal landscape surrounding gender equality across national contexts;
(2) the short-term causal effects of exposure to anti-gender communication on individuals; and
(3) the effectiveness of five theoretically grounded counter-argument strategies (empathy, constructivist challenges to essentialism, humour, fairness, and norms) in shaping how citizens perceive anti-gender messages.

The report provides a detailed account of the experimental design, measurement strategy, and analytical approach, followed by results on both attitude-level change and message-level evaluations. It further discusses cross-national variation in susceptibility to counter-argument strategies, the stability of gender-related attitudes over time, and the potential unintended consequences of counterspeech, including effects on citizens' perceptions of social norms.

In addition to presenting the empirical findings, this document interprets their implications for policy, public communication, and future research on countering anti-gender discourse. It highlights how counter-arguments can influence immediate reception of anti-gender messages, even when attitudes remain stable in the short term.

Acknowledgments

We are grateful to Rebekka Kesberg for her very careful review of this document and constructive feedback. We also thank the audiences at the EPOP Conference 2025 and the Center for European Studies Webinar Series at the University of Exeter for their thoughtful and constructive feedback on earlier versions of this work. We additionally acknowledge Joseph Phillips as well as Ayşe Iraz Kural Başkan for expert methodological advice and support at all stages and Jason Reifler for his insightful comments. Finally, we thank Francesca Mignogna for her excellent research assistance.

List of Definitions and Abbreviations

Anti-gender
In this deliverable, anti-gender (or anti-gender egalitarian) discourse is used as an umbrella term for communication opposing gender equality, feminism, and LGBT+ rights. Explicitly anti-feminist messages are treated as one element within this broader anti-gender repertoire.

Control track
The experimental condition in which respondents viewed anti-gender posts without any accompanying counter-arguments.

Counter-argument strategy
In this deliverable, we refer to the experimental messages attached to anti-gender posts as counter-argument strategies (short: counter-arguments). We use the term counterspeech when referring to the broader practice of responding to harmful or discriminatory speech in public communication.

Filler post
A social media post included in the experiment that addresses a topic unrelated to gender.

GenAI
Generative Artificial Intelligence. In this deliverable, GenAI refers to generative Artificial Intelligence used to support the drafting of counter-arguments.

LGBT+
An abbreviation and collective term for lesbian, gay, bisexual, trans, queer, intersex, and asexual people.

Perspective taking
The cognitive ability to understand a situation from another person's viewpoint by considering their thoughts and feelings.

Random assignment
The procedure by which respondents were allocated to the control or treatment track, and within the treatment track to one of the five counter-argument strategies when seeing anti-gender egalitarian social media posts.

Threat perception
In this report, threat perception refers to beliefs that women's status gains correspond directly with men's status losses. We measure these beliefs with items about zero-sum thinking related to gender equality.

Treatment track
The experimental condition in which respondents viewed anti-gender posts accompanied by one type of counter-argument.

Executive summary

Background/Summary of Contexts

Anti-gender discourse has become an increasingly prominent feature of political communication across Europe. We focus on a variety of topics that are salient in this context, such as: the protection of the traditional family, the portrayal of gender equality as a liberal or foreign threat, challenges to national sovereignty, claims that men are now the disadvantaged group, and opposition to abortion and reproductive rights.

Against this backdrop, Work Package 6 investigates two central questions. The first concerns the causal effects that exposure to anti-gender egalitarian communication has on citizens' gender equality related attitudes (Objective O6.1). The second concerns the potential of counter-argument strategies to mitigate these effects (Objective O6.2). To address these questions, WP6 combines two complementary sources of data. Wave-1 of the PUSH*BACK*LASH survey provides a cross-national overview of public attitudes towards gender equality, threat perceptions, perspective taking, and willingness to support gender-equality initiatives in eight European countries. Wave-2 consists of survey experiments conducted in Hungary, Italy, and Poland, in which respondents were exposed to anti-gender social media posts either unchallenged (control) or accompanied by one of five counter-argument strategies (treatments).

Together, these data offer both a descriptive account of the gender-equality climate across national contexts and experimental evidence on the effectiveness of multiple counter strategies.

Added value and summary of deliverable

1. The results present new data from a cross-national experimental study that identifies the causal effects of anti-gender communication and evaluates whether these effects extend to a range of outcomes covered by Objective O6.1.
2. The report provides comprehensive comparison of five theoretically grounded counter-argument strategies: empathy, constructivist arguments, humour, fairness, and social-norms appeals, thereby addressing Objective O6.2.
3. The deliverable introduces a methodological approach that can be replicated by civil-society organisations, citizens, and policymakers, including the use of AI-assisted drafting of counter-arguments.
4. The report offers one of the first detailed analyses of unintended consequences of counter-arguments, including systematic effects on perceptions of normativity, an issue relevant to both Objective O6.1 and Objective O6.2.
5. The findings provide evidence-based recommendations for designing effective counter-argument interventions in contexts where anti-gender narratives are widespread.

Key findings

1. Stability of attitudes
Across all countries in which we conduct survey experiments (HU, IT, PL), attitudes towards gender equality show little change between Wave-1 and Wave-2. This stability holds even in Poland, where Wave-2 was conducted during a polarised presidential election. Exposure to unchallenged anti-gender messages does not shift attitudes in the short term, nor does exposure to counter-arguments. This finding speaks directly to Objective O6.1, which focuses on identifying the causal effects of anti-gender communication on gender equality related attitudes.

2. Counter-strategies can help maintain support for gender equality
Although attitudes remain stable on average, the counter-argument strategies we tested play a "protective" role in Poland. Among respondents who already recognised discrimination at Wave-1, exposure to unchallenged anti-gender communication reduced this recognition. By contrast, respondents who saw counter-arguments did not show this decline. This indicates that counter-strategies could help maintain support for gender equality by preventing its erosion. This finding addresses both Objective O6.1 (effects of anti-gender communication) and Objective O6.2 (effectiveness of counter-strategies).

3. Counter-arguments influence how individual messages are perceived
While attitudes do not shift, counter-argument strategies shape the interpretation of anti-gender communication. Several strategies increase recognition that messages contain sexist intent, reduce agreement with anti-gender claims, and lower willingness to like related content on social media. This result addresses Objective O6.2 by showing that counter-arguments can alter how anti-gender messages are evaluated.

4. No single counter-argument strategy consistently outperforms others
All five counter-argument strategies tend to succeed or fail together. When counter-arguments reduce agreement or increase recognition of sexism, this occurs across multiple strategies. When no effect is observed, this also applies across strategies. This pattern suggests that topic and context are more important than strategy type. This finding contributes to Objective O6.2, which calls for evaluating the performance of different counter-strategies.

5. Counter-arguments increase perceived societal support for anti-gender views
A consistent and unexpected finding across all countries is that counter-arguments increase perceived normativity. Respondents who viewed anti-gender communication accompanied by counter-arguments estimated that a larger share of the public would agree with the content. This unintended consequence highlights the importance of norm-corrective information in counter-argument interventions. This speaks to both objectives by showing a causal effect of exposure (Objective O6.1) and by identifying a boundary condition and risk associated with counter-strategies (Objective O6.2).

Policy recommendations

Inclusive communication strategies addressing anti-gender narratives should:

- Recognise that counter-arguments can be effective in shaping immediate evaluations of anti-gender messages, even if attitudes remain stable. Counter-arguments can therefore help reduce the traction of communication that challenges gender equality in public debates and online spaces.

- Tailor counter-arguments to national contexts and specific issue areas. A strategy that is effective for one topic in one country may not produce similar effects elsewhere.

- Consider counter-arguments as a long-term engagement tool. Repeated exposure may gradually accumulate into broader attitude shifts, even when single interventions do not produce measurable short-term change.

- Use the knowledge generated in this project on how counter-arguments work, and how they can be designed, in educational, civic, and digital literacy programmes. Policymakers and civil-society organisations can draw on these insights to help citizens understand effective counter-arguments and to build confidence in responding to anti-gender claims in everyday interactions. Teaching individuals why certain counter-strategies are effective can empower them to challenge harmful narratives both online and offline.

- Be aware that counter-arguments may unintentionally increase perceptions of how widespread anti-gender views are. Policymakers and practitioners should take these unintended consequences into account. Further research is needed to identify communication strategies that can mitigate these norm perception effects without reducing the effectiveness of counter-arguments.

1. Introduction

Overview

- This section outlines the central aim of Work Package 6: to identify the causal effects of anti-gender egalitarian communication and to evaluate whether counter-argument strategies can mitigate these effects.
- The research design combines two complementary sources of data: a cross-national baseline survey (Wave-1) capturing gender-equality attitudes across eight European countries, and a preregistered experiment (Wave-2) in Hungary, Italy, and Poland.
- In Wave 2, respondents were randomly assigned either to view anti-gender social media posts without challenge or to see the same material accompanied by one of five counter-argument strategies.
- The findings show that attitudes towards gender equality are stable. At the same time, counter-arguments do influence how anti-gender content is perceived, reducing agreement with anti-gender content and increasing recognition of discriminatory intent.
- Attitudes and post-level evaluations operate on different levels of malleability: while attitudes anchor people's broader worldview, evaluations of individual posts are more flexible and responsive to counter-arguments.

Work Package 6 addresses two key questions: what are the causal effects of anti-gender discourse on citizens, and can counter-strategies mitigate these effects? Across Europe, anti-gender narratives have become a salient part of public discourse. Objective O6.1 is to identify the causal effects of anti-gender egalitarian communication. Objective O6.2 is to develop and test counter-strategies that can mitigate these effects.

The Objective refers to anti-gender and anti-feminist discursive strategies. In the empirical contexts that we analysed, political actors predominantly frame their opposition to equality and equality policies using the language of "gender ideology." The literature shows that this term functions as an "empty signifier" or "symbolic glue" that brings together diverse forms of resistance to gender equality, feminism, reproductive rights, and LGBT+ rights (Mayer and Sauer, 2017; Grzebalska et al., 2017). Anti-feminist claims constitute an important element of this repertoire, but they are typically not articulated as a separate category. Instead, they are incorporated into broader anti-gender mobilisation (Korolczuk et al., 2025). For this reason, and reflecting how these discourses appear in practice, we use anti-gender egalitarian discourse as our primary term in this deliverable and treat anti-feminist rhetoric as one component within it.

To meet these objectives, WP6 draws on two complementary sources of data. First, a large-N cross-national survey in eight countries (Wave-1) establishes baseline attitudes: levels of support for gender equality, threat perceptions, perspective taking, and willingness to get involved. Second, a follow-up panel survey with embedded experiments in Hungary, Italy, and Poland (Wave-2) allows us to identify causal effects and, most importantly, the effectiveness of counter-argument strategies. In Wave-2, respondents were randomly assigned to a control track, where they viewed anti-gender content without challenge, or to treatment tracks, where they saw the same content accompanied by counter-arguments.

We begin by examining country level differences, using the data from Wave-1. Next, we zoom in on the role of anti-gender communication on people and examine person-level outcomes: do citizens' attitudes change when exposed to anti-gender content with or without counter-arguments? These analyses map directly onto Objective O6.1. We also test whether assignment to treatment versus control tracks affected recognition of discrimination, net of baseline attitudes in Wave-1. The results show stability. Neither exposure to unchallenged anti-gender discourse (control) nor to counter-arguments (treatments) produced significant change in person-level attitudes. Moreover, comparison between Wave-1 and Wave-2 shows little evidence of aggregate shifts, even in Poland where Wave-2 overlapped with a presidential election campaign.

This stability is itself an important finding for Objective O6.1. It suggests that attitudes toward gender equality and recognition of discrimination function as deep-seated dispositions rather than opinions swayed by short-term communication. Put differently, the causal effect of exposure to anti-gender discourse on attitudes is negligible in the short run. This does not mean that anti-gender communication has no effect. Instead, it indicates that its impact is more likely to accumulate over time through repeated exposure.

The second part of our analysis shifts from person-level attitudes to post-level evaluations, which speaks to Objective O6.2. Here, counter-arguments demonstrably make a difference. Respondents shown anti-gender posts with counter-arguments were less likely to agree with or "like" the posts, and more likely to perceive them as threatening to gender equality. These are direct indicators of the effectiveness of counter-strategies.

We tested several strategies, including empathy, humour, and fairness-based appeals. Contrary to our preregistered expectation that some strategies would systematically outperform others, we find no clear hierarchy. Instead, strategies appear to succeed or fail together, depending on context and topic. This suggests that counter-arguments can be effective, but that their success depends strongly on the environment in which they are deployed.

Taken together, the findings provide an answer to the two objectives of WP6. For Objective O6.1, the causal effects of (short term) exposure to anti-gender discourse or a counter discourse on attitudes are minimal. For Objective O6.2, counter-arguments do affect how individual messages are perceived: they reduce acceptance and amplification of anti-gender posts and enhance recognition of their threat to gender equality.

The broader implication is that attitudes and communication evaluations operate on different levels of malleability. Attitudes towards gender equality are relatively stable dispositions, while evaluations of individual messages are more flexible and created "on the spot". Counter-arguments may not immediately shift attitudes, but by changing the reception of anti-gender content they have the potential to alter discourse dynamics. Over time, repeated counterspeech could accumulate into broader attitudinal change, which is a key question for future research.

2. Country level context: Wave-1

Overview

- This section summarises the landscape of gender-equality attitudes in the eight countries included in Wave-1 of the survey fieldwork. A more detailed presentation of the data can be found in Deliverable D2.1
- Data is shown for cross country variation in misogyny, gender stereotypes, perceptions of discrimination, threat perceptions, perspective-taking, and willingness to get involved.
- Countries with stronger far-right support and higher structural gender inequality (e.g., Poland) tend to score higher on misogynistic attitudes, underscoring why we focus on them in the experimental work. However, there is also much variation in attitudes towards gender equality that does not map onto support for far-right parties or gender inequality.
- Across all countries, women consistently express more egalitarian views than men, including lower levels of misogyny and greater willingness to get involved in action promoting gender-equality.
- These descriptive results establish the broader societal context in which the Wave-2 experiments discussed in the subsequent sections take place.

Before presenting the experimental findings, we provide a descriptive overview of a selection of indicators for attitudes towards gender equality. This overview presents data collected within Work Package 2, and which is discussed in greater details in the respective Deliverable (D2.1). The survey was fielded online by YouGov. Sampling included quotas for key indicators (age, gender, education, region, vote choice); later weighting allowed inference from the sample to the respective populations. Fieldwork took place between March and June 2025. The questionnaire was developed as part of the work conducted in Work Package 2.

Table 1 Field work dates across the countries and observations per country

Field work Wave-1 dates	16/05/25-10/06/25	16/05/25-10/06/25	16/05/25-06/06/25	19/05/25-05/06/25	20/05/25-04/06/25	21/05/25-30/05/25	26/03/25-28/03/25	19/05/25-01/06/25
N	2015	2048	2026	1020	1059	1057	1061	1138

Below we present descriptive patterns for the following measures (see Appendix for an overview of all items included in the respective measures):

- Misogyny scale
  A nine-item scale adapted from Rottweiler and Gill (2021), capturing hostile attitudes towards women (e.g., "Women seek to gain power by getting control over men.").
- Gender-stereotype endorsement
  An item originating from the Special Eurobarometer 2017, assessing agreement with statements questioning women's suitability for leadership roles (e.g., "Women do not have the necessary qualities and skills to fill positions of responsibility in politics.").

- Modern sexism
  A single item adapted from Swim et al. (1995), assessing the belief that gender discrimination is no longer a societal problem (e.g., "Discrimination against women is no longer a problem in COUNTRY.").
- Threat perceptions measured via zero-sum thinking
  A three-item measure based on Kuchynka et al. (2018), capturing perceptions that progress for women entails losses for men (e.g., "As women have gained status in the workplace, men have lost status in the workplace.").
- Willingness to get involved (participation measure)
  A behavioural intention measure assessing support for signing a petition promoting gender parity in parliament.
- Perspective taking
  A six-item scale adapted from Davis (1980), capturing the tendency to adopt and consider others' perspectives in social situations.

Figures 1–8 plot country means (weighted data), shown separately for men (blue) and women (red), against two context level variables: the vote share of far-right populist parties (at the last general election in the respective country) and the Gender Inequality Index (GII). For categorisation of parties, we rely on the widely used Popu-list data (Rooduijn et al. 2023). With only eight countries in the dataset, correlations are not statistically meaningful (values above about r = .7 would be required to reach conventional significance). The purpose of these figures is therefore to provide a descriptive mapping of distributions rather than inferential testing.

Misogyny, the view that women lack the necessary skills to hold political office, and the belief that discrimination against women is no longer a problem are measures of attitudes towards gender equality in our data. Misogyny and the belief that women lack political skills tend to be higher in countries with larger far-right vote shares (Figure 3). Hungary and Poland, which have the highest far-right vote shares in the dataset, are also positioned at the higher end of the scale for these items. By contrast, countries such as Spain and the United Kingdom, with comparatively lower far-right support, score lower. The comparison with the GII points in a similar direction, with higher levels of misogyny observed in contexts with greater gender inequality. We find a similar pattern for the outcome "women lack the necessary skills to hold political office" (Figure 4). The third measure, the belief that discrimination is no longer a problem, varies across countries but does not align consistently with either contextual benchmark (Figure 5). Zero-sum thinking is used here as a measure of gender-related threat perceptions. Figure 6 shows no strong relationship between levels of zero-sum thinking and far-right support or GII scores.

Perspective-taking, plotted in Figure 7, shows a comparatively consistent descriptive pattern. Countries with higher far-right vote shares, such as Hungary and Poland, show lower levels of perspective-taking among both men and women. Conversely, countries with lower far-right support, such as Spain and the United Kingdom, tend to score higher. A similar but somewhat weaker pattern is visible in relation to the GII: perspective-taking is lower in countries with higher structural inequality. The difference between men and women is small, and the two groups track each other closely across contexts.

Willingness to participate in gender-equality related activities is shown in Figure 8. Here the cross-national pattern is less systematic. Countries with both high and low far-right support can be found at different points on the participation scale, and the relationship with the GII is similarly weak. The most consistent feature is the gender difference: women in nearly all countries report higher willingness to participate than men, although the size of this gap varies.

Taken together, the descriptive figures show that higher levels of misogyny and the perception that women lack the skills for political office are observed in countries with higher far-right support and greater structural gender inequality. Perspective-taking is lower where far-right support and gender inequality are higher. Perceptions that discrimination is no longer a problem and willingness to participate do not display clear alignment with the contextual indicators. These descriptive observations provide a baseline understanding of how attitudes towards gender equality, perceptions of threat, perspective-taking, and willingness to participate are distributed across the national contexts in our study.

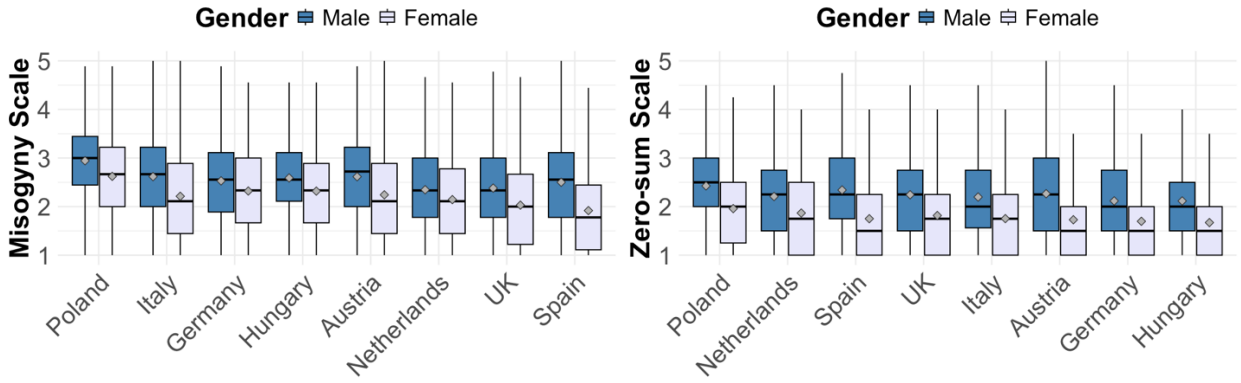


Figure 1 Misogyny and zero-sum scale by country and gender
Weighted country means for misogyny (scale following Rottweiler and Gill, 2021) and zero-sum threat perceptions (Kuchynka et al., 2018). Results are shown separately for men and women.

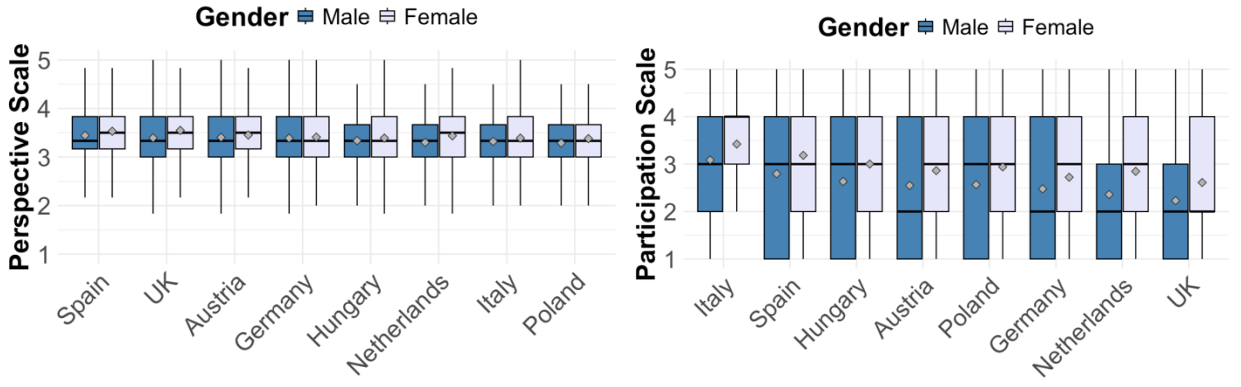


Figure 2 Perspective-taking scale and participation by country and gender
Weighted country means for perspective taking (scale following Davis, 1980) and participation (willingness to get involved). Results are shown separately for men and women.

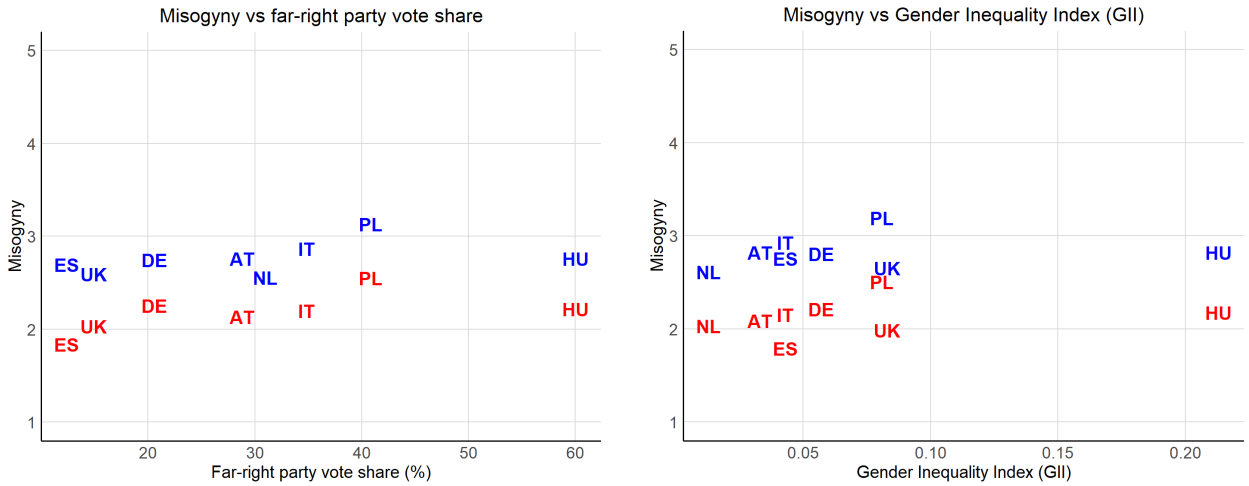


Figure 3. Levels of misogyny vs far-right vote share and gender inequality
Weighted country means of the misogyny scale for men (blue) and women (red), plotted against far-right party vote shares (most recent national election; coding following Popu-List.org, left panel) and Gender Inequality Index scores (2025, right panel).

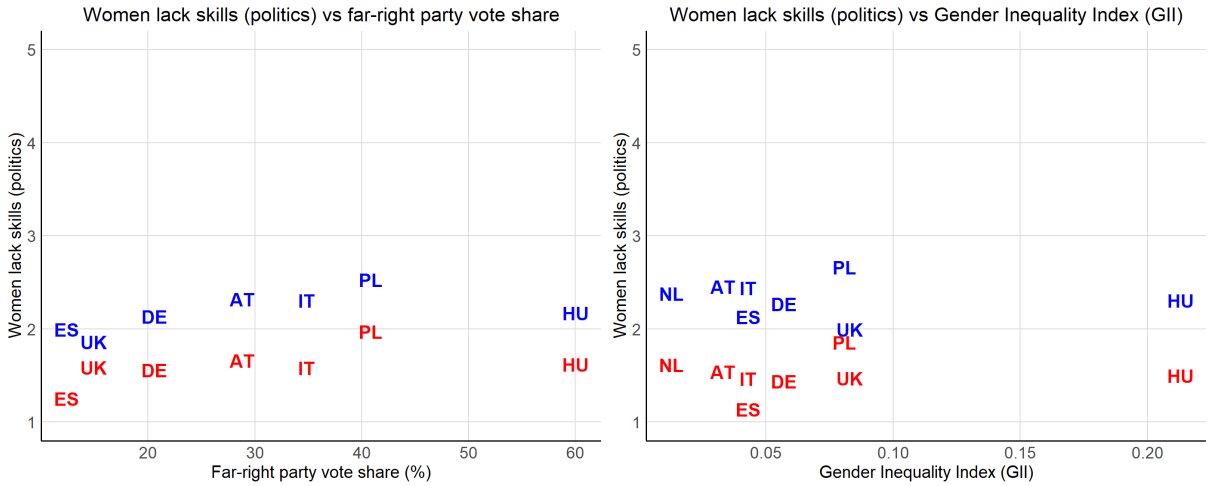


Figure 4. Gender-stereotype item plotted against far-right party vote share and gender inequality
Weighted country means for agreement with the statement "Women do not have the necessary qualities and skills to fill positions of responsibility in politics", shown separately for men and women and plotted against far-right/populist party vote shares (most recent national election; Popu-List.org coding) and Gender Inequality Index scores (2025).

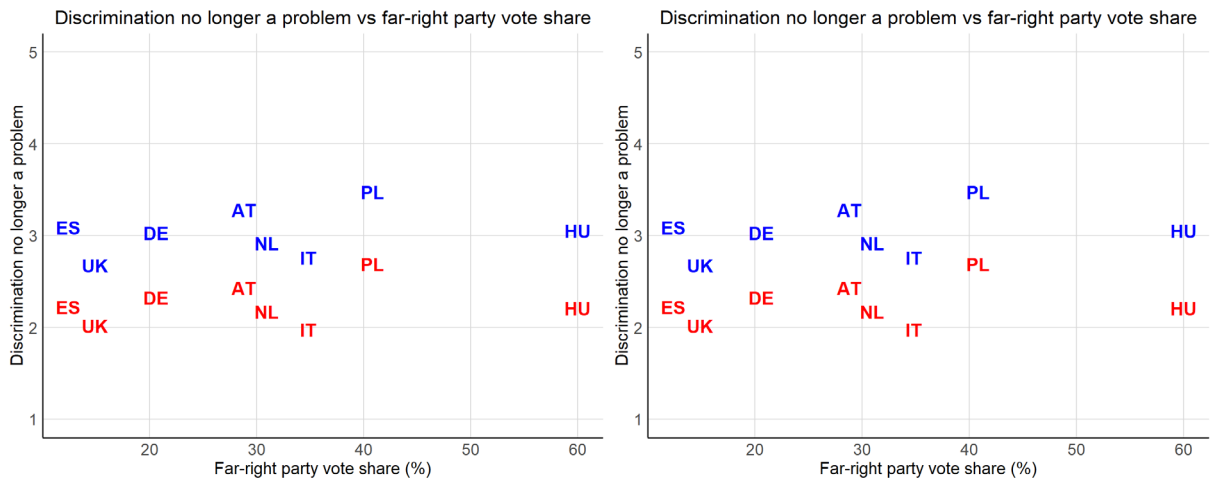


Figure 5. Modern sexism item plotted against far-right party vote share and gender inequality

Weighted country means for agreement with the statement *"Discrimination against women is no longer a problem in [COUNTRY]"* (modern sexism item), shown separately for men and women and plotted against far-right party vote shares (most recent national election; Popu-List.org coding) and Gender Inequality Index scores (2025).

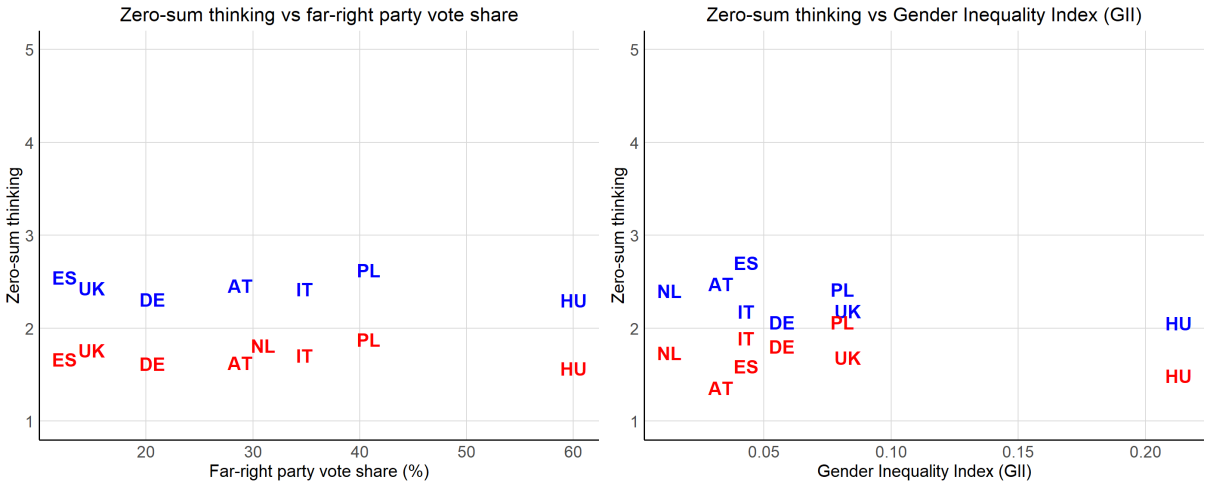


Figure 6. Levels of zero-sum thinking vs far-right party vote share and gender inequality

Weighted country means of the zero-sum thinking scale (our measure for threat perceptions, Kuchynka et al. 2018), shown separately for men (blue) and women (red). Values are plotted against far-right party vote shares (most recent national election; Popu-List.org coding, left panel) and Gender Inequality Index scores (2025, right panel).

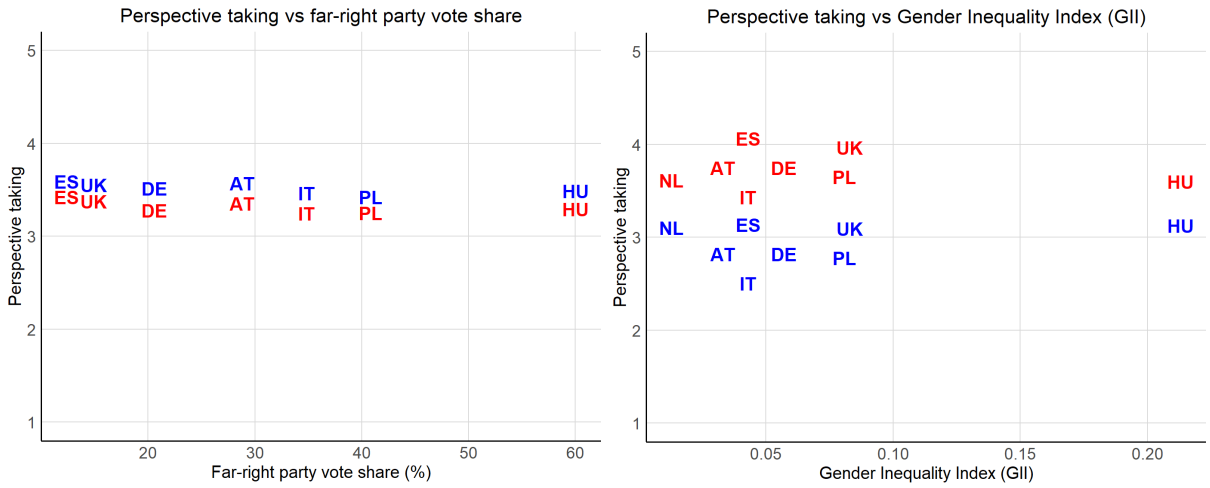


Figure 7. Levels of perspective taking vs far-right party vote share and gender inequality
Weighted country means of the perspective-taking scale (Davis, 1980), shown separately for men (blue) and women (red). Values are plotted against far-right party vote shares (most recent national election; Popu-List.org coding, left panel) and Gender Inequality Index scores (2025, right panel).

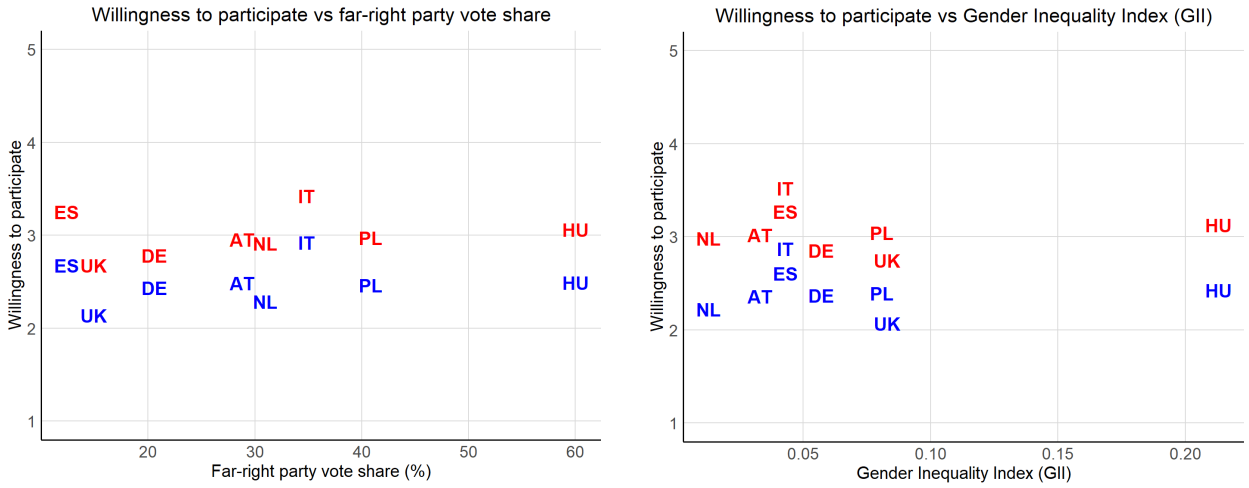


Figure 8. Levels of willingness to participate vs far-right party vote share and gender inequality
Weighted country means for willingness to participate in actions promoting gender equality, shown separately for men (blue) and women (red). Values are plotted against far-right/populist party vote shares (most recent national election; Popu-List.org coding, left panel) and Gender Inequality Index scores (2025, right panel).

3. Counter-argument strategies: rationale

Overview

- This section outlines the theoretical foundations for the five counter-argument strategies examined in WP6: empathy-based, constructivist, fairness-based, humour-based, and norms-based responses. Each strategy draws on established research in social and political psychology on persuasion, prejudice reduction, and how people process potentially harmful messages.
- The section explains why these particular strategies were selected. Each operates through a different psychological mechanism: empathy encourages perspective-taking; constructivist counters challenge essentialist claims about "natural" gender roles; fairness-based appeals highlight justice and equal treatment; humour can reduce defensiveness; and norms-based messages correct misconceptions about what most people support.
- Rather than assuming that one strategy will be effective across all audiences and topics, the section argues for evaluating a diverse set of theoretically grounded approaches. This allows the project to compare the potential of different mechanisms.
- In doing so, Section 3 provides the conceptual groundwork for the experimental work described in Section 4, where these strategies are operationalised into concrete counter-arguments and tested empirically.

We evaluate five counter-argument strategies that literature suggests can effectively promote openness, reduce defensiveness, and foster critical engagement: humour, which can disarm hostility and engage audiences more receptively (Meyer, 2000; Samson & Gross, 2012); empathy, which encourages perspective-taking and humanises those targeted by discrimination (Batson et al., 1997; Turner & Cameron, 2016); constructivist explanations, which challenge essentialist and fixed ideas about gender (Haslam et al., 2002; Taylor et al., 2009); fairness appeals, which invoke principles of justice and equity central to many moral frameworks (Fehr & Schmidt, 1999; Tyler & Smith, 1998); and social norms, which leverage perceived majority opinions to encourage conformity with egalitarian values (Cialdini et al., 1990; Tankard & Paluck, 2016). Our pre-registration uses slightly different terminology for two of these strategies. Specifically, the constructivist strategy is referred to as *essentialism-challenging* in the pre-registration, and the norms-based strategy is labelled *consensus-based*. We adopt the updated terminology here to reflect feedback received since the pre-registration and to ensure consistency with recent theoretical developments. We also revised the numbering.

Particularly, we expose participants either to unchallenged anti-gender messages or to messages paired with one of these counter-arguments. This methodology enables us to compare how each strategy influences individuals' ability to recognise and critically evaluate anti-gender communication.

3.1 The role of counter-argument strategies in challenging anti-gender egalitarian communication

While anti-gender narratives are pervasive and context-sensitive, emerging evidence from social and political psychology highlights the potential for strategically targeted counter-arguments to disrupt their influence. Counter-arguments operate through multiple psychological mechanisms, including cognitive reframing, emotional engagement, and social signalling, which collectively facilitate critical evaluation of biased messages (Martel et al., 2020; Zhou & Shen, 2024). These processes are particularly relevant in politically polarised or digitally mediated environments, where unchallenged misinformation can reinforce prejudices, non-egalitarian social norms, and systemic inequalities (Wardle & Derakhshan, 2023; Guess et al., 2024).

Counterspeech is one approach shown to influence attitudes and behaviours related to prejudice. Interpersonal confrontation, for instance, can reduce prejudice through self-regulation mechanisms (e.g., eliciting guilt, moral reflection, and motivation to align with egalitarian values) and can generate secondary transfer effects across unrelated social groups (Chaney et al., 2021). Online activism offers opportunities for rapid dissemination, community building, and coordinated collective action, yet is constrained by slacktivism, harassment, and performative engagement (Gomez & Kaiser, 2019). Emerging AI-mediated interventions, such as bots and algorithmic messaging, can deliver egalitarian norm-based or empathic counter-messages at scale, producing short-term reductions in aggression and bias (Bilewicz et al., 2021; Crabtree et al., 2025). Building on this evidence, counter-arguments are expected to enhance individuals' ability to recognise biased messages. Accordingly, our main hypothesis (H1) proposes that exposure to counter-arguments increases the extent to which a communication is identified as challenging gender equality. To test this, we examine five theoretically grounded counter-argument strategies, namely using (1) empathy, (2) constructivist challenges to essentialism, (3) humour, (4) fairness, and (5) references to egalitarian norms, each of which leverages distinct psychological mechanisms to target anti-gender messages.

3.1.1 Empathy-based counter-arguments

Empathy enables individuals to understand and resonate with others' perspectives and emotions, fostering prosocial behaviour and reducing intergroup bias (Batson, 2009; Davis, 1983). Cognitive empathy, or perspective taking, facilitates recognition of marginalised experiences, while affective empathy (empathic concern) elicits compassionate responses that motivate supportive action. Empathy has been shown to mediate the relationship between intergroup contact and reduced prejudice toward outgroups (Pettigrew & Tropp, 2008), suggesting that it can serve as a key mechanism through which targeted communication fosters openness and solidarity. Building on this, empathy-based messages can bridge divides and foster greater understanding of marginalised perspectives, increasing engagement with social justice issues, enhancing willingness to confront discrimination, and predicting collective action tendencies among disadvantaged groups (Montada & Schneider, 1989; Thomas & McGarty, 2018; Selvanathan et al., 2017). Research also indicates that individuals who endorse sexist attitudes or hold traditional gender ideologies tend to score lower on both perspective taking and empathic concern (Garaigordobil, 2014), suggesting that empathy deficits may reinforce resistance to gender egalitarian values. Indeed, in context related to gender equality, empathy predicts a lower

level of hierarchy-legitimizing beliefs regarding gender and sexual minorities (Sinclair & Saklofske, 2019).

Although findings on the relationship between perspective taking and intergroup attitudes have been mixed (Paluck et al., 2021), several interventions demonstrate its potential to reduce prejudice and bias. For example, Simonovits, Kézdi, and Kardos (2018) developed an online choose-your-own-adventure game in Hungary where participants assumed the role of a Roma individual; this intervention reduced anti-Roma prejudice by about 0.3 standard deviations immediately and effects persisted at least one month later, also reducing antipathy toward refugees and far-right political intentions. Furthermore, Gutierrez and colleagues (2014) developed the videogame Fair Play, where players experience subtle racial bias through the eyes of a Black graduate student; their study revealed that this immersive, empathy-driven intervention reduced implicit racial bias, likely because it actively engaged players in perspective-taking. However, intervention success depends on the inclusion of reflective components and the consideration of individual differences in baseline attitudes highlighting the need for empathy training to be combined with explicit challenges to sexist beliefs and social norms (van der Graaff et al., 2014; Szekeres et al., 2024). These findings provide a foundation for predicting that empathy-focused interventions can meaningfully increase awareness of gender inequality.

Hypothesis 1.1: Empathy-based counter-arguments increase the extent to which individuals identify challenges to gender equality.

3.1.2 Constructivist counter-arguments

Constructivist arguments challenge essentialist beliefs which explain gender characteristics as inherent, natural, or biologically fixed (Skewes et al., 2018) by emphasizing the dynamic interaction between biological and social influences. These perspectives stress that gender traits are shaped through socialization, culture, and lived experience rather than fixed at birth. However, when gender differences are framed as biologically innate (e.g., such as assuming men are naturally more competitive or women inherently nurturing) social hierarchies are legitimised as inevitable, limiting support for structural change (Eagly & Wood, 2013).

Essentialist beliefs are also linked to greater acceptance of discriminatory practices and differential treatment based on gender (Skewes et al., 2018). Furthermore, essentialist views often operate subtly through generic language, which serves as a powerful conduit for transmitting such beliefs across generations (Leshin et al., 2021). Girls as young as six, for example, have already internalised the idea that brilliance is more typical of boys (Bian et al., 2017). If left unchallenged, such essentialist statements perpetuate stereotypes and sustain prejudice and inequality (Rhodes et al., 2012; Leshin et al., 2021).

In their research Broemer and Grabowski (2018) investigated how nature versus nurture explanations shape the processing of gender stereotypes. Following four studies, they concluded that both frames could trigger stereotypical expectations but their influence on social reasoning differs. Nurture-based frames lead to more flexible interpretations and greater openness to counter-stereotypical evidence, as perceivers adjust their views based on context. In contrast, nature-based frames reinforce stereotypical thinking, especially when faced with disconfirming evidence. These findings suggest that emphasizing nurture-based explanations, which highlight

the role of socialization and context, can reduce endorsement of essentialist beliefs and increase openness to counter-stereotypical information. Thus:

Hypothesis 1.2: Counter-arguments that criticise essentialist beliefs by highlighting the socially constructed nature of gender differences will increase the extent to which individuals recognise challenges to gender equality.

3.1.3 Humour-based counter-arguments

Humour, especially subversive or satirical, is a powerful tool for challenging social injustice, including sexism, racism, and power inequalities (Sharpe & Hynes, 2016; Dobai & Hopkins, 2020; Hylton, 2018). Subversive feminist humour can subtly yet powerfully challenge sexism by strengthening women's sense of collective efficacy. Cohen-Chen et al. (2024) found that humour addressing sexist injustice increased women's belief in their shared ability to create social change, which in turn raised their intentions to engage in collective action. Importantly, women with low feminist identification showed a stronger increase in collective action intentions in response to humour than they did in response to serious messages, whereas women with strong feminist commitment gained little additional benefit. These findings highlight humour's potential to engage those who are less politically active.

Similarly, Vizcaíno-Cuenca et al. (2024) demonstrated that humorous feminist messages were more effective than serious ones in mobilizing low-identifiers. They attributed this to the reduced psychological resistance humour elicits, and its ability to activate the peripheral route of persuasion – less confrontational, more engaging, and thus more persuasive for ambivalent audiences. Riquelme et al. (2021) further confirmed this pattern, emphasising that humour can mobilise ambivalent people towards egalitarian attitudes while leaving high-identifiers unaffected.

Furthermore, for marginalised groups, humour serves as both coping and resistance: Roma participants used irony to ease tension and subvert hierarchies (Dobai & Hopkins, 2020), Black coaches reframed racist experiences through parody (Hylton, 2018) and racially charged performance art exposed absurdities of racism (Sharpe & Hynes, 2016). Humour can also generate positive emotions that support activism, such as body-positive parody boosting gratitude and engagement (Fasoli et al., 2025), and correlating with empathic concern (Hampes, 2001).
However, humour carries risks. Sexist jokes may mask prejudice, reduce confrontation, and reinforce stereotypes, especially among women with hostile sexist beliefs (Mallett et al., 2016). Benevolent sexism can lead women to enjoy sexist jokes about men, subtly sustaining gender hierarchies (Koszałkowska & Wróbel, 2022). Humour also reduced women's negative emotions, such as anger or outrage toward men, which slightly lowered motivation to act (Cohen-Chen et al., 2024). Careful calibration to audience and context is therefore essential to maximise humour's counter-prejudice potential.

Hypothesis 1.3: Humour-based counter-arguments increase the extent to which individuals identify challenges to gender equality.

3.1.4 Fairness-based counter-arguments

Fairness appeals emphasise structural inequities and social injustices, drawing attention to the unequal distribution of resources, opportunities, and power. By highlighting these systemic disparities, such messages can elicit moral emotions (including guilt, sympathy, and outrage) which serve as important drivers of prosocial behaviour and support for gender equality (Montada & Schneider, 1989; Selvanathan et al., 2017). When individuals recognise that social systems advantage some groups over others, they are more likely to reflect on their own role in perpetuating or challenging inequality, reducing defensiveness and fostering openness to egalitarian values (Powell et al., 2005; Szóstakowski & Besta, 2024).

The effectiveness of fairness-based interventions is strengthened when people perceive their actions as personally meaningful and capable of producing tangible outcomes. Linking moral responsibility to concrete, actionable steps (such as advocating for inclusive policies, supporting underrepresented groups, or engaging in allyship) enhances engagement and sustained behavioural change (Renger et al., 2020). Moreover, these appeals can be deployed across multiple levels of social interaction, including interpersonal conversations, online campaigns, and institutional initiatives, allowing for coordinated strategies that reinforce egalitarian norms over time (Hartman et al., 2022; Li et al., 2025). By combining moral framing with actionable guidance, fairness-based counter-arguments create a psychologically compelling pathway for motivating both awareness and practical commitment to reducing inequality.

Hypothesis 1.4: Fairness-based counter-arguments will increase recognition of challenges to gender equality.

3.1.5 Norms-based counter-arguments

Social norms, i.e. the shared perceptions of what behaviours are typical (descriptive norms) or socially approved (injunctive norms), exert a powerful influence on individual and collective behaviour. Importantly, norms are not static; they can be framed as dynamic and evolving, a feature that can be leveraged to promote social change. Mortensen et al. (2019) demonstrated that dynamic minority norms (perceived as gaining momentum) elicited greater conformity than static norms, as individuals interpret them as indicators of future majority behaviour. This "trend effect" suggests that framing prosocial behaviours as growing movements may enhance persuasiveness. In parallel, Álvarez-Benjumea (2024) found that expressions of prejudice are shaped by perceptions of social acceptability, which are sensitive to shifting political and cultural signals. Together, these findings highlight the strategic value of portraying egalitarian norms as both prevalent and ascendant.

In the context of climate action, Constantino et al. (2022) emphasise the need to use descriptive and injunctive norms deliberately, cautioning that highlighting harmful behaviours can inadvertently normalise them. Norm interventions are most effective when they are visible, relevant to the in-group, and supported by broader structural change. Wenzel and Woodyatt (2024) note that internalised norms can sustain cooperation but may provoke backlash when enforced too rigidly, recommending that norm messages be aligned with personal values and that misperceptions be corrected to reduce pluralistic ignorance, i.e. a misperception about the prevalence of others' attitudes or behaviours (Sargent & Newman, 2021).

Pluralistic ignorance has been documented in domains such as alcohol use, mental health, climate change, and gender bias. Such misperceptions can enable harmful practices, including excessive drinking, or suppress positive behaviours like seeking mental health support. In the gender equality context, DeSouza and Schmader (2022) showed that both men and women underestimate the extent to which men care about sexism. This misperception discourages men, especially those concerned with masculinity norms, from acting as allies despite privately endorsing egalitarian values. Correcting such misperceptions can increase allyship intentions, underscoring the need to address both private beliefs and perceived group norms when promoting collective action.

Research on confrontation styles further informs norms-based strategies. Lewis and Yoshimura (2017) found that clear, explanatory, and polite confrontations reduce prejudice while minimizing defensiveness. At the institutional level, Schuster et al. (2023) demonstrated that introducing egalitarian norms reduced gender pay bias in HR decisions, increased offers to women, and fostered a shared commitment to equity.

Messages that emphasise widespread agreement or shared values can also reduce polarisation and bolster adherence to egalitarian norms (Li et al., 2025; Gelfand et al., 2024). When backlash statements are confronted, observers are more likely to infer that egalitarian norms are dominant (Tirion et al., 2023). Even when perceptions of consensus are subjective, norm-violating statements are more likely to be judged unacceptable by bystanders (Bicchieri et al., 2022; Tirion et al., 2023). Conversely, failing to address norm violations can lead to imitation of deviant behaviour and norm erosion (Li et al., 2025). Taken together, these findings suggest that highlighting widespread agreement on gender equality can strengthen perceptions of norms, thereby increasing individuals' ability to recognise and critically evaluate challenges to gender equality.

Hypothesis 1.5: Norms-based counter-arguments increase the extent to which individuals identify challenges to gender equality.

3.2 The role of counter-arguments for norm perceptions

Perceptions of social norms play a central role in how individuals evaluate and align with public messages. Research on norm perception shows that people frequently misjudge the prevalence of others' attitudes and behaviours, often assuming that problematic views are more widely shared than they actually are (Tankard & Paluck, 2016). This miscalibration is closely related to the *false uniqueness effect*, whereby individuals underestimate the extent to which others share their own egalitarian beliefs and overestimate the popularity of rival, anti-egalitarian positions (Suls & Wan, 1987). In the context of anti-gender communication, such biases may foster a sense of minority status among egalitarian respondents, reducing willingness to challenge dominant rhetoric. Counter-arguments can function as corrective signals in this regard: by making visible that disagreement exists, they provide tangible evidence that others reject the anti-gender message, thereby reducing misperceptions of consensus. Seeing a counter-argument not only signals that "at least one other person" opposes the biased statement but also makes it more plausible that such opposition is shared by a

broader collective (Cialdini & Goldstein, 2004). In this way, counter-arguments may weaken pluralistic ignorance and prompt individuals to infer that the original message is not as widely endorsed as it might appear without challenge.

At the same time, prior research on corrective communication also cautions about potential backfire effects. In some cases, corrections can inadvertently amplify the salience of the original claim, leading audiences to infer that a message is more widely discussed, and perhaps more widely endorsed, than they initially believed (Nyhan & Reifler, 2010; Walter & Tukachinsky, 2020). A similar mechanism is observed in media studies of *false balance*, where presenting opposing perspectives can unintentionally convey that both positions enjoy equal societal support (Boykoff & Boykoff, 2004). Thus, while counter-arguments are expected to reduce inflated perceptions of societal agreement with anti-gender messages, they may in some contexts produce the opposite effect by enhancing the visibility and perceived prevalence of the original rhetoric.

Hypothesis 2: Counter-arguments reduce the estimated share of others in society who agree with the original post.

4. The experimental research design

Overview

- This section explains how WP6 implements a preregistered experiment to identify the causal effects of anti-gender communication and examine whether counter-arguments can mitigate these effects. All respondents were exposed to anti-gender social media posts reflecting five common themes identified in contemporary anti-gender discourse:
    - gender equality as a threat to the traditional family;
    - gender equality as a liberal or elite-driven project detached from ordinary people;
    - gender equality as an ideology threatening the nation, national identity or sovereignty;
    - claims that gender equality oppresses men;
    - portrayals of abortion rights and reproductive autonomy as dangerous or immoral
- The counter-arguments accompanying these posts in the treatment groups were developed using a combination of GenAI-assisted drafting and pilot testing.
- Two pilot studies informed the selection of counter-arguments for the main experiment. Pilot Study 1 asked respondents to identify the most convincing counter-arguments among several candidates created by GenAI. Pilot Study 2 tested the extent to which each counter-argument expresses the intended strategy, ensuring that stimuli were both credible and theoretically valid before full fieldwork.
- The experimental outcomes operate at two levels. Attitudinal measures capture broader views about gender equality (which are expected to be relatively stable). Post-level evaluations, in contrast, measure respondents' immediate reactions to messages (e.g., agreement with the post, recognition of sexism), where change is more likely. This dual structure allows testing whether counter-arguments influence how communication is interpreted, even when attitudes are stable.

The experiment at the heart of Work Package 6 was designed to investigate two core questions: what effects anti-gender discourse has on citizens, and which counter-strategies can effectively mitigate those effects. The pre-registration (Appendix) anchored the design. It reflected a key insight that shaped our approach: the experiment needed to balance the ambition of tracing attitudinal change with the feasibility of detecting more immediate communication effects.

Respondents who had already taken part in Wave-1 of the survey were re-contacted for Wave-2. They were randomly assigned to a control track or to one of several treatment tracks. All saw anti-gender content in the form of social media posts. Those in the control track saw all posts (five substantively important ones and two filler ones) unchallenged, while those in treatment tracks saw the same posts followed by counter-arguments. These counter-argument interventions drew on strategies identified in the literature: appeals to empathy, fairness, humour, constructivism, or norms. This design allowed us to examine both the impact of exposure to unchallenged anti-gender communication and the potential of different strategies to counteract it.

The objectives of the work package refer to a broad array of potential effects: attitudes toward gender equality, threat perceptions, perspective taking, and willingness to engage in politics. In selecting which measures to include post-treatment, we were guided by two considerations. First,

the outcomes needed to be conceptually close to the manipulation in order to maximise the chance of detecting effects. Second, we needed to ensure sufficient statistical power to compare counter-strategies (a core task of Objective O6.2). On this basis, we focused on two sets of attitudinal measures: a gender system-justification battery that taps perceptions of fairness, legitimacy, and the persistence of discrimination, and a smaller set of items capturing direct attitudes toward gender equality. These measures are both theoretically relevant and empirically tractable. They also connect to threat perceptions, since those who see gender equality as threatening are more likely to endorse gender system-justifying claims.

Allocating survey space in this way came with a trade-off: fewer items directly on political engagement, perspective taking, or other downstream outcomes. Power calculations showed that the experimental comparison of multiple counter-argument strategies required large samples, and the design had to ensure this. At the same time, the measures we included allow us to speak to the dynamics that underpin willingness to engage: gender attitudes serve as anchors that shape how people process anti-gender content and counter-arguments.

The results we present later on underscore our reasoning of focusing much on message-level (or, more precisely, post-level) variation. Across the panel, attitudes toward gender equality proved stable. Even in Poland, where the timing of Wave-2 coincided with a presidential election campaign between candidates with distinct positions on gender equality related issues (European Policy Centre 2025), there was no aggregate shift between waves. This stability suggests that attitudes are not easily shifted in the short term but rather serve as anchors through which communication is filtered. In the short term, attitudes shape how content is perceived, and our results show that counter-arguments can change perceptions of anti-gender content. Over the longer term, repeated exposure to such content and its contestation could plausibly reshape attitudes themselves. In this sense, attitudes are simultaneously anchors and outcomes: stable in the short run, but potentially subject to change through cumulative discursive processes.

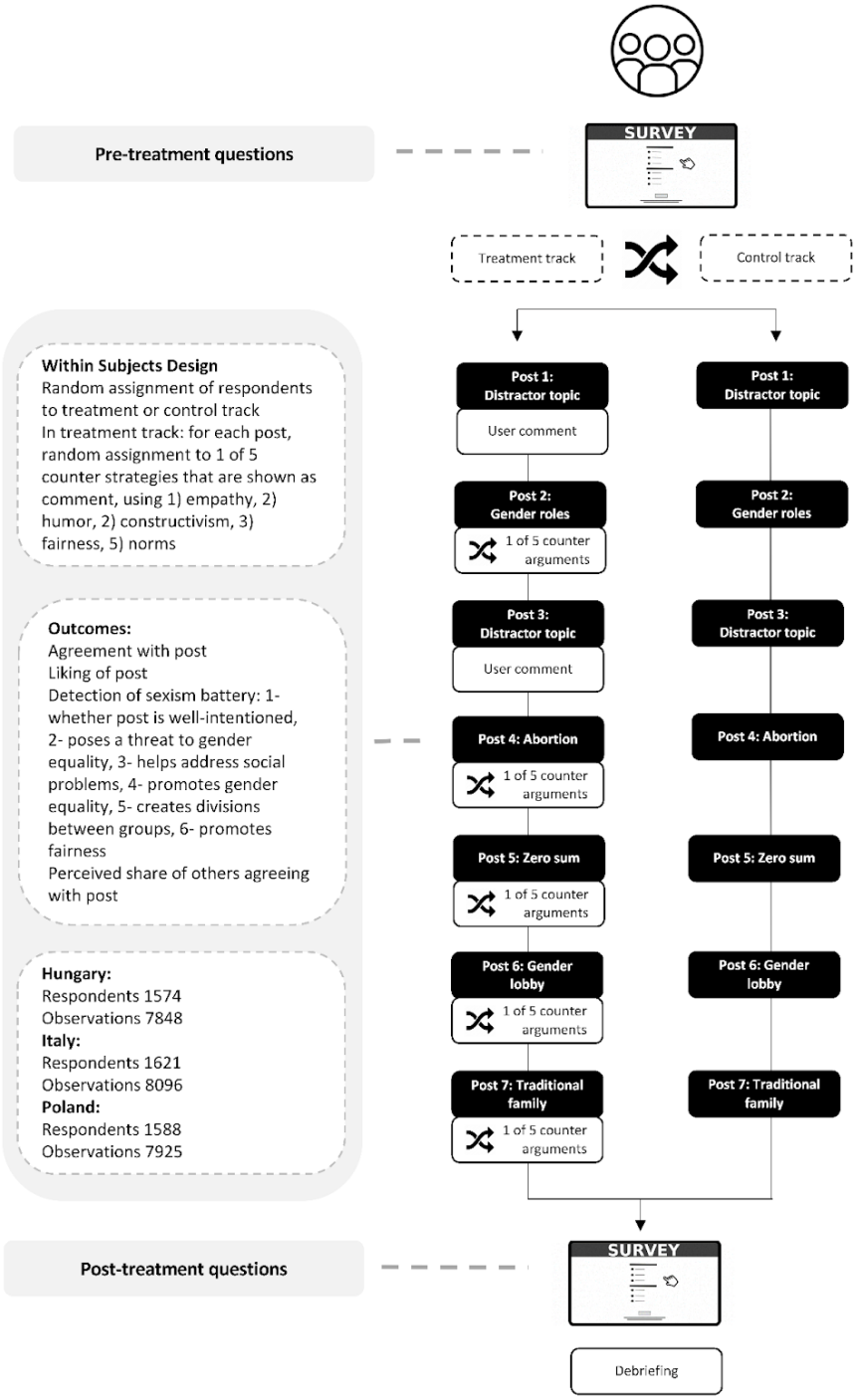
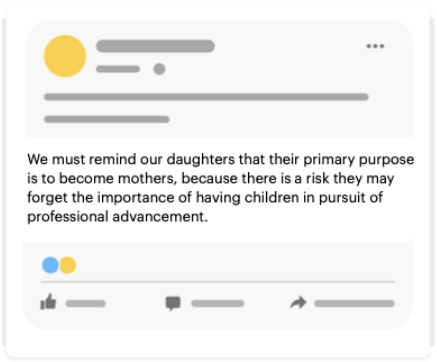


Figure 9. Survey design

Respondents were randomly assigned either to the control track, where they viewed seven anti-gender posts without commentary, or to the treatment track, where the same seven posts were shown accompanied by counter-arguments. Two of the seven posts addressed filler topics unrelated to the topic of the study to reduce transparency of the experimental purpose. Within the treatment track, respondents were randomly assigned to one of five counter-argument strategy conditions for each social media post (empathy, fairness, humour, constructivist, or norms).

YouGov



Please place yourself on this scale.

I completely disagree with the post	1	2	3	4	5	6	I completely agree with the post
Schools should not address this topic at all	1	2	3	4	5	6	Schools definitely should address this topic

To what extent do you agree with the following statements about the original post?

	Strongly disagree	Disagree	Somewhat disagree	Somewhat agree	Agree	Strongly agree
The post promotes fairness	○	○	○	○	○	○
The post is well-intentioned	○	○	○	○	○	○
The post promotes gender equality	○	○	○	○	○	○
The post creates division between groups	○	○	○	○	○	○
The post helps address social problems	○	○	○	○	○	○
The post poses a threat to gender equality	○	○	○	○	○	○
If I saw a post like this, I would click the "like" button on social media	○	○	○	○	○	○

When thinking about people living in COUNTRY, what percentage of them do you think agree with the original post?

No one agrees — Everyone agrees

Figure 10. Example of the screen that respondents in the control track saw.
Post includes anti-gender equality content and no counter-argument. Respondents were asked a series of post-related questions on the same page.

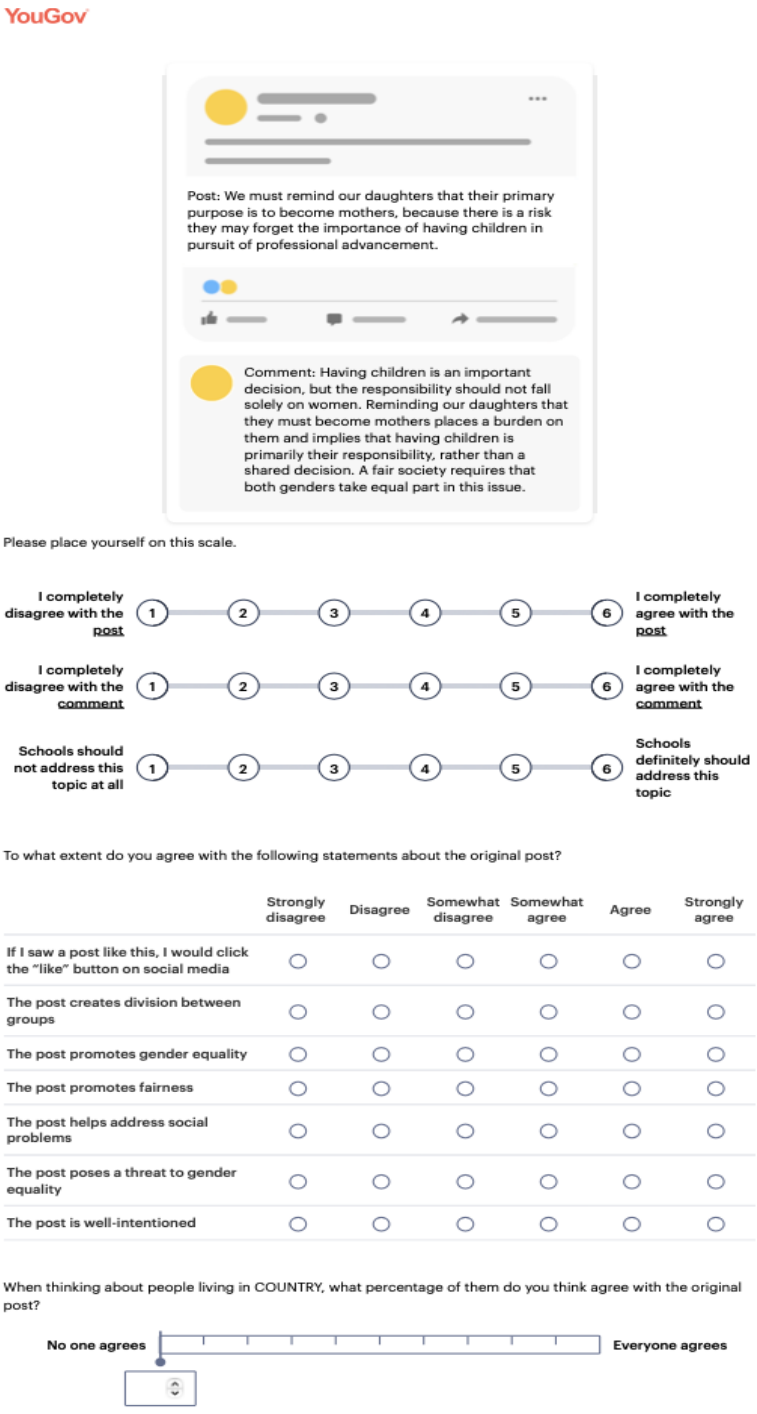


Figure 11. Example of the screen that respondents in the treatment track saw.
Post includes anti-gender equality content and a counter-argument. Respondents were asked a series of post-related questions on the same page.

4.1 The anti-gender discourse: original material

Political actors, religious authorities, and conservative movements in Poland, Hungary, and Italy increasingly promote a worldview rooted in gender essentialism, i.e. the belief that men and women have distinct, biologically determined roles that must be preserved for the good of society (Graff & Korolczuk, 2022). This vision, long championed by the Vatican under the doctrine of "sex complementarity," has been reframed in contemporary politics as a defensive struggle against a nebulous threat called "gender ideology" (Garbagnoli, 2016; Grzebalska & Pető, 2018). The term functions as a unifying label for a range of anxieties: fears about the breakdown of the traditional family, the erosion of national identity, and the influence of liberal or foreign agendas.

In Hungary, the Fidesz government has elevated anti-gender rhetoric to a core governing principle, explicitly linking "gender" with migration and war as existential threats (Kuhar & Smrdelj, 2025). Prime minister Orbán has characterised "gender ideology" as a "progressive virus" and celebrated Hungary's development of a "vaccine" against it, reinforcing a narrative that ties equality politics to external interference and cultural subversion (Orbán, 2023 as cited in Dömötör & Horváth-Kávai, 2023; Pető, 2022). Institutional measures, such as the 2018 revocation of gender studies accreditation, have further entrenched the perception that gender scholarship is "ideology, not science" (Balogh, 2018 as cited in Balázs, 2018; Pető, 2022).
In Italy, anti-gender mobilisation has been sustained by strong church–movement–party linkages. Since the early 2010s, groups such as La Manif Pour Tous Italia and the Committee to Defend Our Children have mobilised large-scale protests against civil unions, LGBT+ rights, and comprehensive sexuality education, most notably the 2015 and 2016 Family Day rallies (Lavizzari & Prearo, 2019; Prearo, 2020). Catholic lobbying has historically influenced both right-wing and centre-left agendas, framing gender equality as a partisan, urban-liberal project at odds with "traditional" Italian values (Garbagnoli & Prearo, 2018; Lavizzari, 2025).

In Poland, anti-gender politics form part of the Law and Justice Party's (PiS) ideological core, bolstered by Catholic media and organizations such as Ordo Iuris. Following the 2020 Constitutional Tribunal ruling that introduced a near-total abortion ban, campaigns intensified linking "gender ideology" to threats against children, family, and nation (Graff & Korolczuk, 2017, 2022). Measures such as "LGBT-free zones" and "Stop LGBT" initiatives have recast equality claims as external impositions akin to cultural colonisation (Grzebalska & Pető, 2018; Paternotte & Kuhar, 2017).

Across these contexts, the language of "gender ideology" condenses a wide spectrum of opposition into a simple moral and political binary: defending tradition versus succumbing to foreign decadence. This framing has proven politically potent because it resonates with existing conservative values while presenting progressive reforms as dangerous, imposed, and culturally alien (Grzebalska et al., 2017; Paternotte, 2023). Taken together, these national climates provide the discursive repertoire that our study examines. In each, "gender ideology" serves as a catch-all frame, merging the defines of the "natural" family, nationalist sovereignty, men-as-victims claims, and opposition to reproductive rights into a single, emotionally charged narrative. Against this backdrop, our analysis identifies five recurring narrative frames that operationalise this discourse in concrete and persuasive ways. These frames (outlined below) distil the broad

anti-gender climate into specific, repeatable storylines that challenge gender equality in everyday political and cultural communication.

The key treatment material that includes challenges to gender equality mirrors the following five core frames of anti-gender discourse. The first theme we based our study on is the portrayal of gender equality as a threat to the “nuclear” or “traditional” family. This argument posits that gender equality undermines the “natural” family structure, particularly by challenging traditional gender roles and questioning the primacy of heterosexual relationships. Many anti-gender narratives argue that the nuclear family is under attack from the promotion of gender equality, leading to fears about the destruction of the social fabric and endangering children’s well-being (e.g., Villa 2017; Blum, 2015).

The second theme frames gender equality as a liberal project. By casting gender equality as inherently "liberal," proponents of these arguments suggest that support for gender equality is incompatible with conservatism, patriotism, or adherence to traditional cultural norms. This portrayal also helps to paint gender equality as an elitist, top-down imposition rather than a grassroots effort to address inequality and injustice. The frame exploits political divides and fears of cultural change, thereby undermining the legitimacy of gender equality initiatives (e.g., Graff & Korolczuk, 2017; Grabowska, 2022).

The third theme presents gender equality as an ideology that threatens the nation, national identity or national sovereignty. It is often depicted as a foreign influence that imposes non-native values on local (national) cultures. This narrative constructs gender equality as a form of colonialism or global elitism, aiming to weaken national identity and erode cultural sovereignty (e.g., Gwiazda, 2020; Graff & Korolczuk, 2022). In this frame, gender equality is portrayed as a Western, foreign, or transnational imposition that does not align with the traditions and values of a country.

The fourth theme claims that gender equality oppresses men. This narrative is built around the idea that promoting gender equality marginalises men and undermines their roles in society, suggesting that men are being deprived of their rights or unfairly treated. It suggests that efforts to achieve equality go too far, resulting in discrimination against men, and claims that masculinity is under siege (e.g., Graff & Korolczuk, 2022).

The fifth theme focuses on the issue of abortion rights, specifically portraying efforts to secure reproductive rights for women as dangerous or immoral. Narratives in this category argue that the promotion of abortion rights is part of an agenda that threatens the sanctity of motherhood and devalues traditional family values (e.g., Paternotte & Kuhar, 2017; Mayer & Sauer, 2017). These arguments often intersect with broader anti-gender narratives.

The materials were initially translated from the original language into English by native-speaker members of the research team. They were then translated into Polish, Hungarian, and Italian by native speakers of the respective languages.

Table 2 Social media posts content and sources

English translation	source
We must remind our daughters that their primary purpose is to become mothers, because there is a risk they may forget the importance of having children in pursuit of professional advancement.	https://www.la7.it/coffee-break/video/calo-demografico-lavinia-mennuni-fdi-la-prima-aspirazione-deve-essere-quella-di-diventare-mamma-28-12-2023-520264
It doesn't matter whether a fetus has already been born or not — they should have rights, and their life is equal in value to that of their mother, and to all of us.	ATV. (2022, September 14). Dóra Dúró on the abortion law – Interview on "Egyenes beszéd" [Video]. YouTube. https://www.youtube.com/watch?v=IPnyF1EneTo&t=89s
Meanwhile, the international gender lobby has launched a campaign of conquest — and this conquest threatens our families as well. The goal of gender lobby's conquest is to sow confusion. The aim is to make us uncertain about even the most fundamental things — like whether a father is a man and a mother is a woman.	Alapjogokért Központ. (2021, March 4). Book launch of "The Critique of Gender Theory" by Gergely Szilvay [Video]. YouTube. https://www.youtube.com/watch?v=C2YSlnfwDz0&t=3103s
What do gender equality enforcement measures lead to? They will result in men being pushed out by force, and positions being filled based not on competence, but on gender. As a consequence, competent men will be replaced by less competent women.	https://x.com/StopDMiP/status/1800848091458810202
We don't need to abandon our traditions in the name of progress. People often say that if we want to live in Europe, we must live like Europeans — with two mothers or two fathers. But that's not how it's going to be. We stand for the traditional family. The family is the foundation of everything. We will do everything we can to preserve it in its traditional form: one woman, one man, and children.	https://www.youtube.com/watch?v=5jJwJPGCgns

4.2 Testing of counter-arguments

To operationalise counter-arguments, we first translated five theoretically grounded strategies (empathy, constructivist, humour, fairness, and norms) into short, social-media-style comments that could naturally appear beneath anti-gender posts. Draft counter-arguments were generated with ChatGPT and then validated in two pilot studies. This workflow produced a diverse pool of concise counters that could be vetted empirically before inclusion in the field experiment. Machine generation was strictly a drafting aid: only validated wording entered the final fieldwork. Participation in all studies was voluntary and anonymous under institutional ethical guidance. Only the counter-arguments identified as most convincing in Pilot Study 1 and validated in Pilot Study 2 were carried forward into the main fieldwork. Given that the pilot studies were conducted in Hungary, the finalised counter-arguments were translated by native speakers into English and subsequently into Polish and Italian.

4.2.1 Pilot study 1: choice-based selection of counter-arguments

Before testing effects, we first identified which counter arguments are perceived as most convincing. Between the 12th of March and the 18th of April 2025, an online survey with students from ELTE University (Hungary) randomly assigned participants to one of five questionnaire versions. Each version presented five anti-gender claims, and beneath each claim, three candidate counters drawn from the designated strategy family for that claim. Respondents selected the single counter they considered "most effective". Of 625 starters, 543 completed the questionnaire (female 242; male 110; non-binary/other 9). The sample was young (M age = 24.66, SD = 8.99) and education skewed toward secondary and university degrees, consistent with a student-heavy convenience sample. For each claim-by-strategy combination we calculated the share selecting each option and retained the top-rated wording. Because the sample is not representative and the task is pre-selection rather than population inference, we treat these results as a pragmatic screen to prune weaker options, not as evidence of general persuasive rankings.

4.2.2 Pilot study 2: manipulation checks of strategies

Having selected candidate counters, we then verified that items labelled as humour, fairness, empathy, constructivist, or norms were in fact perceived as expressing the intended cue. From the 27th of April to the 22nd of May 2025, a second online survey presented respondents (again students from ELTE University) with each retained counter and asked them to rate on five 5-point items the degree to which it "uses humour", "emphasises social justice", "evokes empathy", "challenges the rigidity of gender roles", and "appeals to social norms" (1 = strongly disagree; 5 = strongly agree). Of 306 starters, 244 completed (female 186; male 53; non-binary/other 5; M age = 23.58, SD = 6.40). Education again skewed higher, and average feminist identity was 4.45 (SD = 1.69) on a 1–7 scale. For each counter we compared the intended rating to the other four; we considered the manipulation successful when the intended characteristic scored higher at conventional thresholds. Most counters passed this check. Humour and fairness were most consistently identified as distinct by respondents. Empathy, constructivist, and norms cues overlapped more often, which is unsurprising given that several counters simultaneously invite compassion. Across types, language invoking justice tended to receive high ratings even when not the primary intent, reflecting the moral framing that runs through much counterspeech.

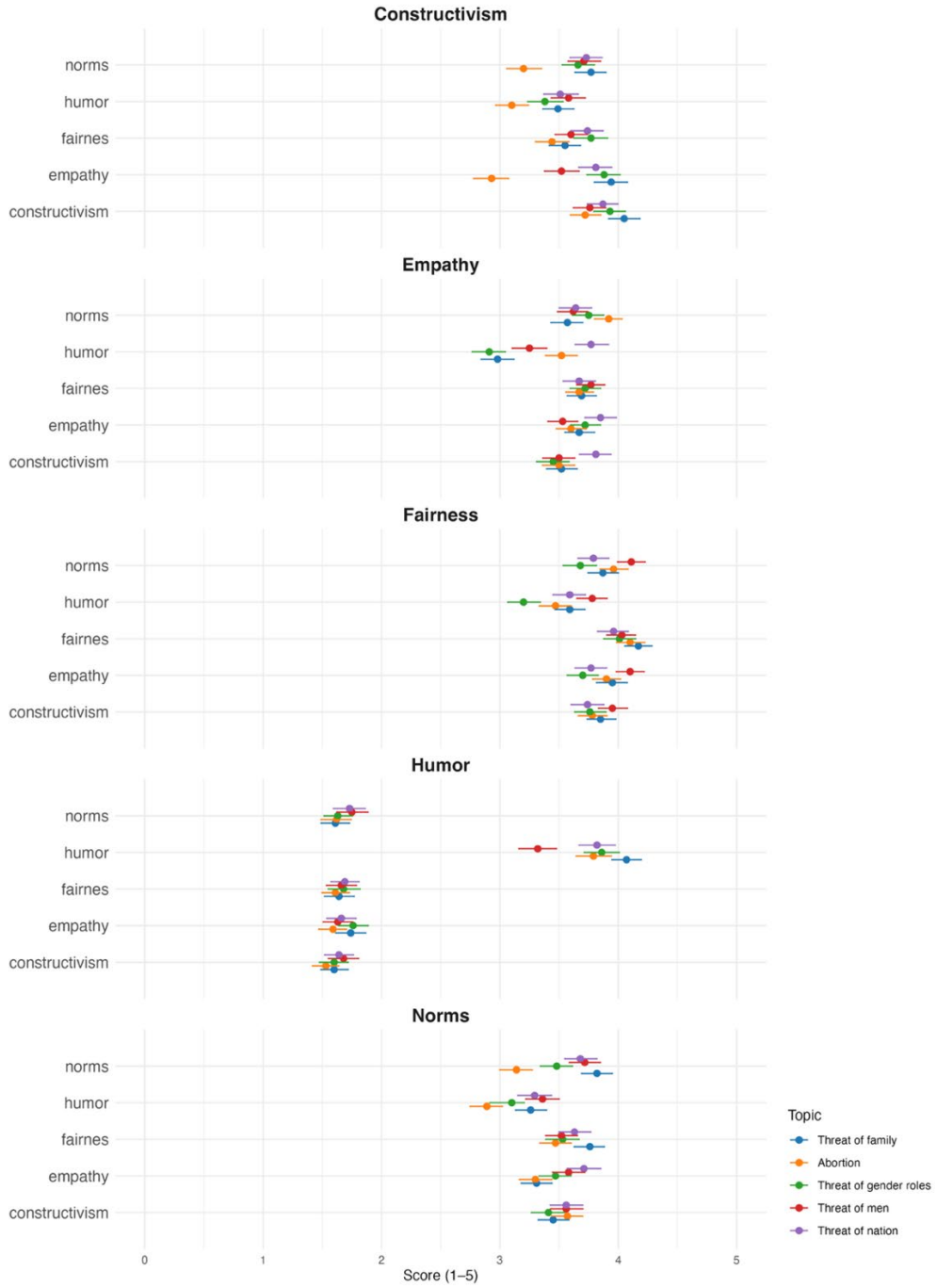


Figure 12 Perceptions of counter-argument strategies in the pilot study.
Dots represent point estimates and lines indicate 95% confidence intervals. Respondents evaluated each counter-argument strategy on the extent to which it was perceived as drawing on constructivism, empathy, fairness, humour, or norms. The results indicate overlap: for example, empathy-based counter-arguments are also rated highly on fairness. This pattern is not surprising, as it is difficult (if not impossible) to construct a fairness-oriented counter-argument that does not also express elements of empathy. The humour dimension stands out: most counter-argument strategies score low on humour, with only the humour-based messages receiving consistently high ratings on this attribute.

5. Sample, fieldwork, descriptive results

Overview

- Wave-2 fieldwork took place in July 2025 in Hungary, Italy and Poland, i.e. it was launched about seven weeks after Wave-1 fieldwork was started.
- The final samples include around 1,600 respondents in Hungary, Italy and Poland.
- Descriptive results from the control group (who saw only the original anti-gender posts) show that respondents generally recognised the posts as containing sexist claims, though the degree of recognition varied by country.

Ethics approval for the pilot studies and main fieldwork was obtained at ELTE University. Participants were recruited by YouGov in Hungary, Italy, and Poland. YouGov handled participant remuneration in line with its standard procedures. The final sample comprised 1,574 respondents in Hungary (55.02% female, Mage = 49.77, SDage = 14.71), 1,621 in Italy (52.31% female, Mage = 51.24, SDage = 15.34), and 1,588 in Poland (53.97% female, Mage = 50.07, SDage = 15.70).

Table 3 Field work dates and N per wave and country.

Country	Wave-1 fieldwork period	Wave-1 N	Wave-2 fieldwork period	Wave-2 N
Hungary	16/05/2025 – 10/06/2025	2,015	07/07/2025 – 22/07/2025	1,574
Italy	16/05/2025 – 10/06/2025	2,048	07/07/2025 – 22/07/2025	1,621
Poland	16/05/2025 – 06/06/2025	2,026	07/07/2025 – 22/07/2025	1,588

Participants were randomly assigned to one of six between-subject conditions: a control group viewing unchallenged anti-gender posts, or one of five treatment groups in which the same posts were accompanied by a counter-argument using a single strategy. We aimed to recruit 1,300 participants per country, based on a power analysis targeting 80% power to detect a small effect size (Cohen's f = 0.10). We oversampled slightly to account for potential exclusions due to incomplete responses or quality-control failures.

Descriptive statistics from respondents in the control condition are depicted in Figure 13 below. In general, detection of sexism in posts was above the midpoint. Hungarian participants generally detected posts as the most sexist, followed by Italian participants, then Polish participants. Agreement with the posts varied based on the item, with agreement on traditional gender roles lower, and agreement on a pro-traditional family frame (among Polish respondents) higher. Agreement with the posts was higher among Polish participants than among other

participants. Post liking was generally below the scale midpoint, except for the traditional family post among Polish respondents. Polish participants again liked posts more than other participants. Participants generally saw the posts as minority viewpoints, below the scale midpoint on the norms measure, with country differences small and inconsistent.

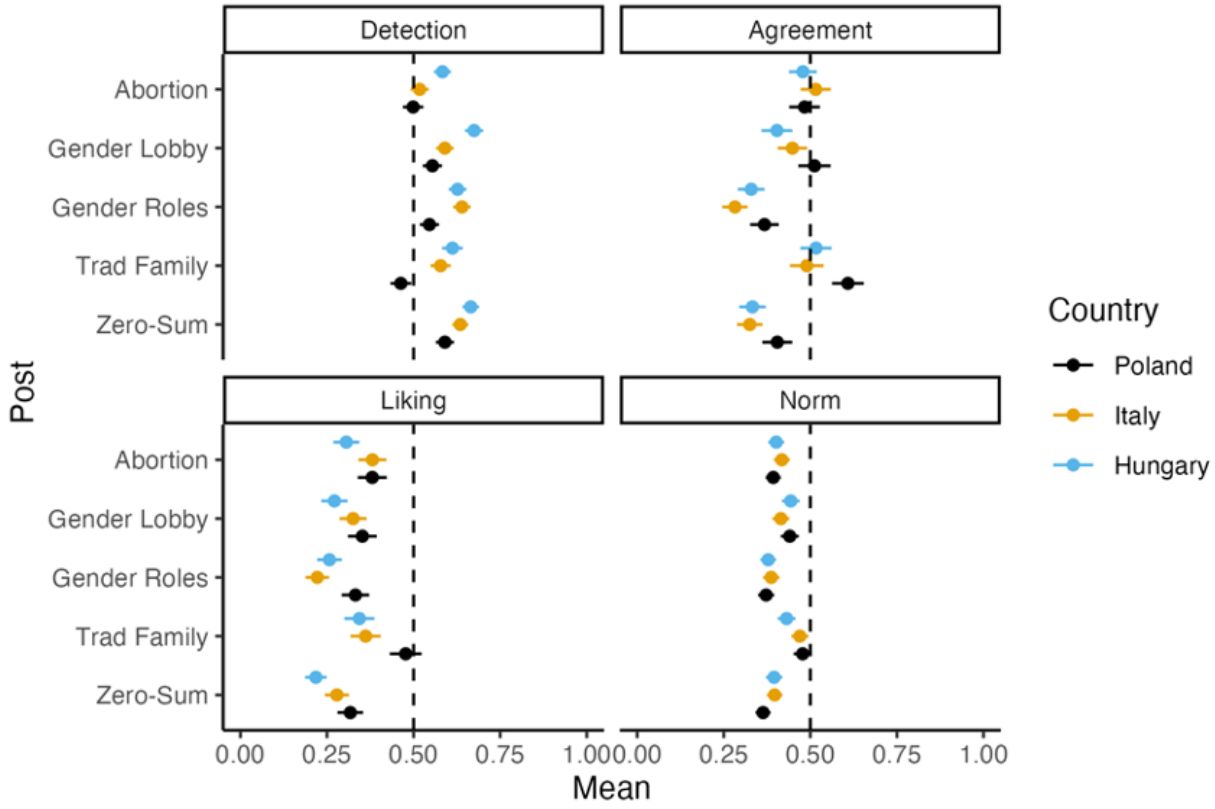


Figure 13 Descriptive evaluations of anti-gender posts across detection, agreement, liking, and perceived norms (control condition)
The figure presents means for four evaluative measures of the original posts among respondents in the control condition (i.e., without counter-arguments) in Poland, Italy, and Hungary. For each post topic (Abortion, Gender Lobby, Gender Roles, Traditional Family, Zero-Sum), the top panels show (1) detection of sexism, measured using items on perceived threat to gender equality, fairness, intention, and divisiveness, and (2) agreement with the post (1–6 scale). The bottom panels show (3) liking of the post (likelihood of clicking "like," 1–6 scale) and (4) perceived social norms, measured as the estimated percentage of people in the respondent's country who would agree with the post (0–100 slider). Across countries, mean scores for sexism detection were generally above the scale midpoint, with Hungarian respondents tending to detect the most sexism, followed by Italian and then Polish respondents. Agreement varied strongly by post type: respondents agreed least with gender-role items and more with traditional-family content, with Polish respondents showing generally higher agreement than Italian or Hungarian respondents. Liking was typically below the scale midpoint, except for the traditional-family post among Polish respondents. Perceived social norms were generally below 50%, indicating that most respondents viewed the posts as minority opinions, with only small and inconsistent differences between countries.

6. Results

Overview

- This section presents the main findings of WP6, distinguishing between treatment effects on attitudes (Section 6.1) and treatment effects on evaluations of individual posts (Section 6.2). Section 6.1 speaks directly to Objective O6.1, the effect of anti-gender communication on attitudes, while Section 6.2 speaks to Objective O6.2, the identification of effective counter-strategies to anti-gender communication.
- Section 6.1 tests whether exposure to anti-gender messages, either unchallenged in the control track or challenged by counter-arguments in the treatment track, affects respondents' gender-related attitudes. Because all counter-arguments are aggregated into a single treatment track, this design allows us to assess whether anti-gender communication itself moves attitudes and whether adding counter-arguments mitigates such effects. It does not allow comparing the effectiveness of individual counter-strategies. This comparison is the focus of Section 6.2.
- Across countries, attitudes remain highly stable (Section 6.1). Exposure to anti-gender messages, whether challenged or unchallenged, does not produce measurable attitudinal change. This suggests that gender-related attitudes function as deep-seated predispositions that are not easily shifted by brief exposure to social-media content.
- In contrast, evaluations of individual posts are malleable (Section 6.2). Counter-arguments alter how respondents perceive the anti-gender posts shown in the experiment, including their perceived sexism, fairness, intention, likability, and perceived normative support.
- This divergence highlights an important pattern: attitudes tend to be stable, whereas message-level reactions can shift, depending on whether counter-arguments are present.

The reporting of our results begins with person-level attitudes for Objective O6.1 and turns then to post-level evaluations for Objective O6.2. We start with the simplest tests: we compare attitudes in Wave-1 and 2, and then we compare respondents assigned to any treatment track (who saw anti-gender content followed by counter-arguments) to those in the control track (who saw the same content unchallenged). Because these Wave-2 outcomes were measured at the respondent level and assignment to treatment versus control occurred at the respondent level, these comparisons directly test whether counter-arguments shift broad attitudes in the short run.

Next, we then leverage the panel link to increase precision and align more closely with our causal objective. For the repeated item "Discrimination against women is no longer a problem in [COUNTRY]" (reverse-coded so higher values indicate more recognition of discrimination), we estimate regressions with the following specification: Wave-2 outcome as a function of treatment assignment, the Wave-1 value of the same item, and, where indicated, age, gender, and tertiary education. This approach adjusts for baseline dispersion and yields a cleaner estimate of the treatment effect than a Wave-2-only comparison. As a complementary check, we estimate a difference model with the change score (Wave-2 minus Wave-1 on the same item) as the outcome and treatment as the predictor, optionally adding the same covariates; results from lagged DV regressions and the change-score model are presented together to show that conclusions do not hinge on functional form.

Having established average effects, we examine the heterogeneity of these effects across individuals. We interact treatment with Wave-1 dispositions (misogyny, zero-sum threat beliefs, perspective-taking, and willingness to sign a pro-quota petition) to test whether counter-arguments protect or shift outcomes more among some baseline types than others. We visualise these interactions with marginal-effect plots that trace the estimated treatment effect across the observed moderator range with 95% confidence intervals.

We then turn to post-level evaluations for Objective 2. Because each respondent rated multiple posts, we estimate multilevel models with posts nested in participants, using random intercepts for participants and fixed effects for treatment strategy arms relative to the unchallenged control baseline. These models assess how counter-argument strategies (empathy, fairness, humour, constructivism, norms) shape immediate evaluations of the specific message (sexism detection, agreement, "like" intention, and perceived normativity) both pooled across posts and, in exploratory analyses, disaggregated by post topic (abortion, traditional family, gender roles, "gender lobby," zero-sum). Across all analyses we use listwise deletion for missing covariates, maintain consistent coding (higher values = more egalitarian recognition/support where applicable), and report two-sided tests at $\alpha = .05$.

6.1 Stability and change in attitudes

Overview

- Section 6.1 examines whether exposure to anti-gender messages, either unchallenged or accompanied by counter-arguments, produces changes in respondents' attitudes toward gender equality. This analysis directly addresses Objective O6.1 of WP6, which concerns the effect of anti-gender communication on attitudes.
- Respondents were randomly assigned to a control track, where they saw anti-gender posts without counter-arguments, or to a treatment track, where they saw the same posts with counter-arguments. After viewing several posts, respondents answered two batteries that capture gender-related attitudes: a four-item gender equality scale and a six-item gender system justification scale.
- The design allows us to test whether anti-gender communication shifts attitudes and whether the presence of counter-arguments mitigates any such effects. However, because all counter-arguments are pooled into one treatment track, Section 6.1 cannot examine the relative effectiveness of different counter-strategies. That analysis is the focus of Section 6.2.
- Across all three countries, the results show no evidence of treatment effects at the attitude level. There are no substantively meaningful differences between respondents in the control and treatment tracks.
- Attitudes measured in Wave 2 replicate those recorded in Wave 1. This reinforces the conclusion that gender-related attitudes are stable over time.
- This stability holds even in politically charged contexts, such as Poland during the presidential election campaign.

- The findings indicate that short-term exposure to anti-gender content does not affect gender-related attitudes, regardless of whether respondents see the material unchallenged or accompanied by counter-arguments. These results help clarify that while attitudes appear resistant to short-term influence, message-level evaluations (examined in Section 6.2) show greater malleability.

To address Objective O6.1, we examine whether exposure to anti-gender content, unchallenged or accompanied by counter-arguments, alters citizens' attitudes toward gender equality. The design allows us to test this in several complementary ways: by comparing averages across two waves, by contrasting treatment and control groups in Wave-2, and by probing interactions between baseline attitudes (measured at Wave-1) and experimental assignment (in Wave-2). Together, these approaches speak directly to whether entrenched attitudes are malleable in the short run, and whether counter-arguments can mitigate any effects of anti-gender discourse.

6.1.1 Stability of attitudes across waves

We begin by showing descriptive statistics for how our modern sexism item develops over time, i.e. comparing values of the repeated item "Discrimination against women is no longer a problem in [COUNTRY]" between Wave-1 and Wave-2 in Poland, Hungary, and Italy. We reversed the item in the presentations below so that higher values indicate greater recognition of discrimination, and lower levels of modern sexism. In Poland, the two waves coincided with a presidential election (first round on the 18th of May, second round on the 1st of June). Respondents from Wave-1 were recontacted in Wave-2 and randomly allocated to either the control track (viewing unchallenged anti-gender posts) or one of several treatment tracks (viewing the same posts with counter-arguments).

The comparison of the values of our modern sexism item scores highlights stability. We do not find statistically significant differences between Wave-1 and Wave-2 averages. In other words, levels of the modern sexism item remain strikingly constant. This is noteworthy given that the Polish respondents were surveyed during an election campaign in which candidates with distinct positions on cultural issues were competing (European Policy Centre 2025), in contrast to those in Italy and Hungary. The result suggests that recognition of discrimination is less a volatile opinion and more a stable disposition that does not shift easily over short time horizons.

This stability has two implications for Objective O6.1. First, if attitudes remain unchanged even in the midst of an intense national campaign where cultural issues generally and abortion rights in particular played a role (European Policy Centre 2025), it is unlikely that a short experimental manipulation like our 15-minute survey with treatments alone would move them substantially. This makes the repeated item and thus, as expected, outcomes such as batteries for threat perceptions or perspective taking, more valuable as moderators that help identify for whom counter-arguments are more or less effective, rather than as outcomes in their own right that might change as a result of our treatments. Second, the high test–retest reliability underscores the broader conclusion that some attitudes related to gender equality are deeply entrenched, while other outcomes (such as post-level evaluations of sexism, fairness, or willingness to amplify content) are more malleable. In this sense, the Wave-1–Wave-2 panel comparison provides both

an anchor for causal analysis and a benchmark against which more sensitive measures can be interpreted.

Table 4. Modern sexism item mean score across survey waves by condition

Country	Condition	Mean Wave-1	Mean Wave-2	N
Hungary	Control	3.45	3.44	281
Hungary	Treatment	3.48	3.46	1293
Italy	Control	3.68	3.71	265
Italy	Treatment	3.71	3.67	1356
Poland	Control	3.04	2.94	259
Poland	Treatment	2.86	2.98	1328

Item: *"Discrimination against women is no longer a problem in [COUNTRY]."* We refer to this as the modern sexism item (reverse-coded so that higher values indicate lower modern sexism, i.e., greater recognition of discrimination). Means for Wave-1 represent only those respondents who were successfully recontacted in Wave-2 and assigned to the control or treatment track; full Wave-1 sample sizes are therefore not reported because they are not used in the paired comparison. Paired t-tests show that none of the comparisons between Wave-1 and Wave-2, nor between treatment and control tracks within Wave-2, reach statistical significance.

6.1.2 Experimental treatment effects on attitudes

Having established the overall stability of the modern sexism item from Wave-1 to Wave-2, we next consider whether exposure to counter-arguments at Wave-2 produced differences on a wider set of items which we only collected at Wave-2. This test is about whether being in the control condition or in the treatment condition has an effect on respondents' attitudes. The key comparison here is between respondents in the control track, who viewed anti-gender content unchallenged, and those in the treatment tracks, who saw the same posts with counter-arguments.

We tested effects on two sets of outcomes. The first is the six-item gender system justification battery adapted from Jost and Kay (2005), which measures perceptions of fairness and legitimacy in gender relations (e.g., "In general, relations between men and women are fair"; "Society is set up so that men and women usually get what they deserve"; see Appendix for full scale). Responses were on five-point scales, with higher scores indicating stronger system justification. Reliability was acceptable in all three countries (Hungary $\alpha = .75$, Italy $\alpha = .71$, Poland $\alpha = .78$).

The second is a four-item attitudes toward gender equality scale adapted from Pratto et al. (2013). Items included "The equality between men and women should be our goal" and "Men should dominate over women" (reverse-coded, see Appendix for full scale). Reliability was adequate in Italy and Poland ($\alpha = .77$ and .76 respectively), but low in Hungary ($\alpha = .11$), so in the Hungarian sample items were analysed separately.

The effect of receiving any treatment on gender system justification is depicted in Figure 14. The effect of experimental condition on attitudes toward gender equality is depicted in Figure 15. Receiving a counter-argument did not have a significant effect on gender system justification in any country's sample (ps ≥ .148). Similarly, treatment had no detectable effect on the four-item gender equality scale in Italy or Poland (ps ≥ .313). In Hungary, the low reliability of the scale meant that items were examined separately; here, treatment increased support for one equality item ("The equality between men and women should be our goal") with statistical significance (p = .036), but no consistent pattern was found across the others.

Taken together, these results show that, consistent with the two-wave comparison, entrenched attitudes such as system justification and gender equality orientations do not shift in the short term in response to either unchallenged anti-gender statements or to counter-arguments.

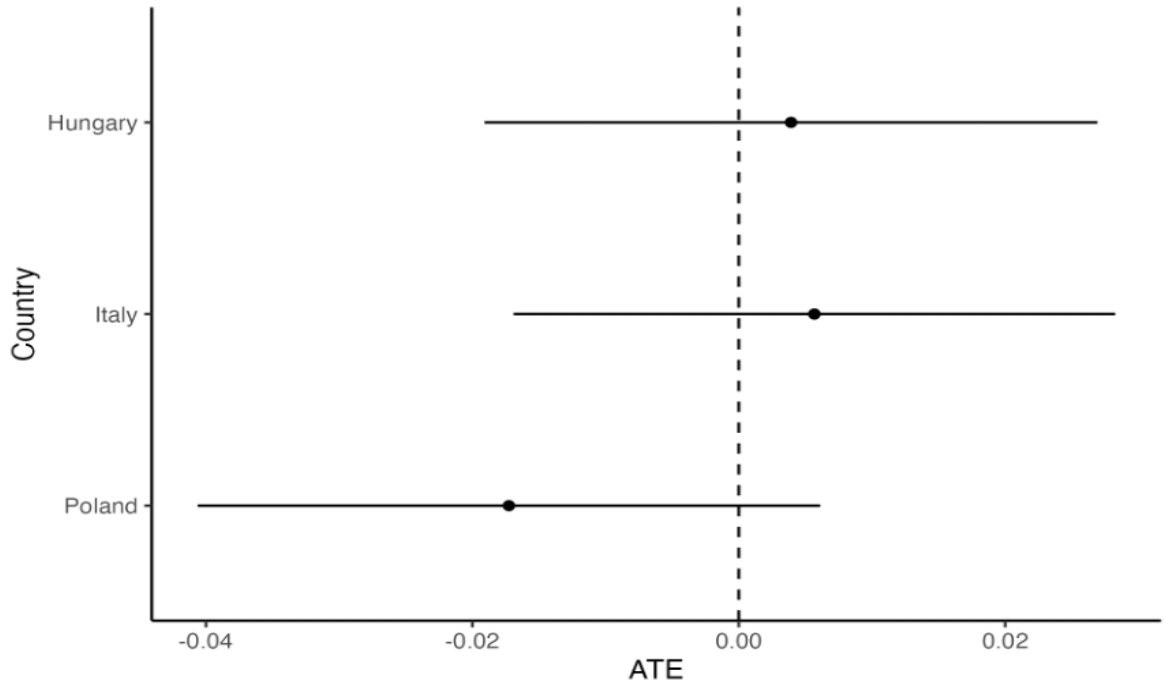


Figure 14. Effect of receiving treatment on gender system justification
The figure displays the average treatment effect (ATE) of receiving counter-arguments on the gender system justification battery. Dots represent point estimates and horizontal lines represent 95 percent confidence intervals. Across all three countries, the estimated effects are close to zero and the confidence intervals include zero, indicating no detectable impact of the treatment on system justification. The battery includes six items measuring perceived fairness and legitimacy of gender relations (response scale: 1 = strongly disagree to 5 = strongly agree). Items are: "In general, relations between men and women are fair," "The division of labor in families generally operates as it should," "Gender roles need to be radically restructured" (reverse-coded), "Sexism in society is getting worse every year" (reverse-coded), "Society is set up so that men and women usually get what they deserve," and "Discrimination against women is no longer a problem in [COUNTRY]."

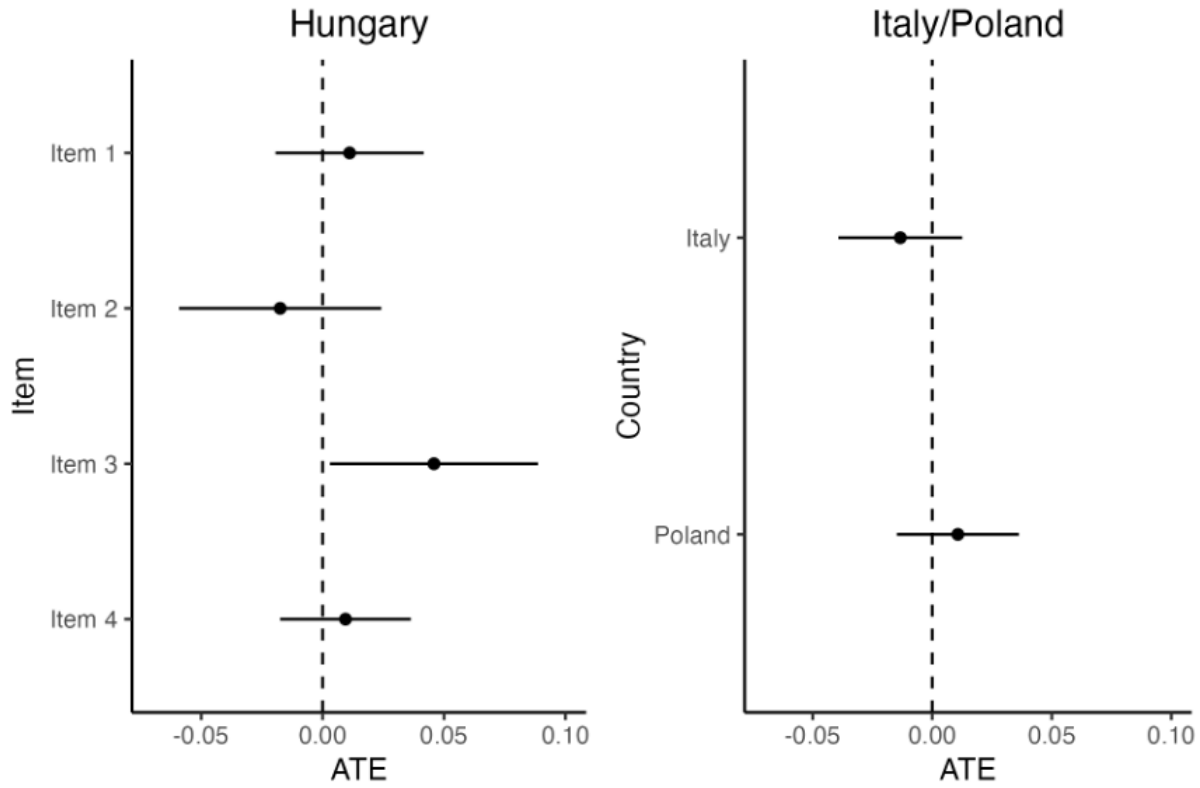


Figure 15. Effect of receiving treatment on attitude towards gender equality
The figure shows the average treatment effects (ATEs) of receiving counter-arguments on four items measuring attitudes toward gender equality. Dots indicate point estimates and horizontal lines indicate 95 percent confidence intervals. For Italy and Poland, the items form a reliable scale and results are shown at the country level. In Hungary, scale reliability was low, so items were analysed separately. Across countries, most estimated effects are close to zero and confidence intervals include zero, indicating no detectable impact of treatment on attitudes toward gender equality. The four items (response scale: 1 = strongly disagree to 5 = strongly agree) are: (1) "In setting priorities, we must consider men and women alike," (2) "We should not push for equality between men and women" (reverse-coded), (3) "The equality between men and women should be our goal," and (4) "Men should dominate over women" (reverse-coded).

6.1.3 Heterogeneity of experimental treatment effects

Although average treatment effects are negligible, treatment may nonetheless operate differently across subgroups. We therefore examine heterogeneity of treatment effects in two ways. In section 6.1.3.1, we test the treatment effects by levels of our modern sexism item that is included in both waves. In Section 6.1.3.2, we briefly discuss the results of analyses that test whether treatment effects vary by various characteristics people vary on at Wave-1: levels of misogyny, threat perceptions (zero-sum thinking), perspective taking, and willingness to get involved.

We estimated regression models whose outcome is modern sexism at Wave-2 (discrimination recognition) with treatment as main independent variable. We test if the effect of treatment varies by modern sexism item scores at Wave-1, using an interaction (treatment x modern sexism

at Wave-1). Figure 16 displays predicted values by treatment condition, and Figure 17 shows the marginal effect of treatment across the distribution of Wave-1 values of modern sexism. In Poland, one of the interactions is substantively meaningful. In the control track, the predicted values plot shows a clear downward slope: respondents with higher baseline recognition of discrimination in Wave-1 (lower sexism) reported lower recognition (higher sexism) at Wave-2 when exposed to unchallenged anti-gender content. In contrast, in the treatment track, the slope is flat: exposure to counter-arguments prevents this erosion. The marginal effect plot confirms this pattern: the effect of treatment grows with higher Wave-1 scores, becoming positive and statistically significant at the upper end of the baseline distribution. Counter-arguments help those who already recognised discrimination at Wave-1, while it does not meaningfully shift the views of those starting from a low baseline. In Italy and Hungary, the results look different. The marginal effect plots show no discernible treatment effect at any baseline level; the lines are flat and the confidence intervals overlap zero throughout. The predicted values plots in these two countries show upward slopes for both treatment and control groups. This suggests a positive baseline persistence: respondents who already recognised discrimination in Wave-1 also tended to report higher recognition at Wave-2, regardless of treatment assignment. In other words, the experimental manipulation did not significantly alter outcomes in these contexts.

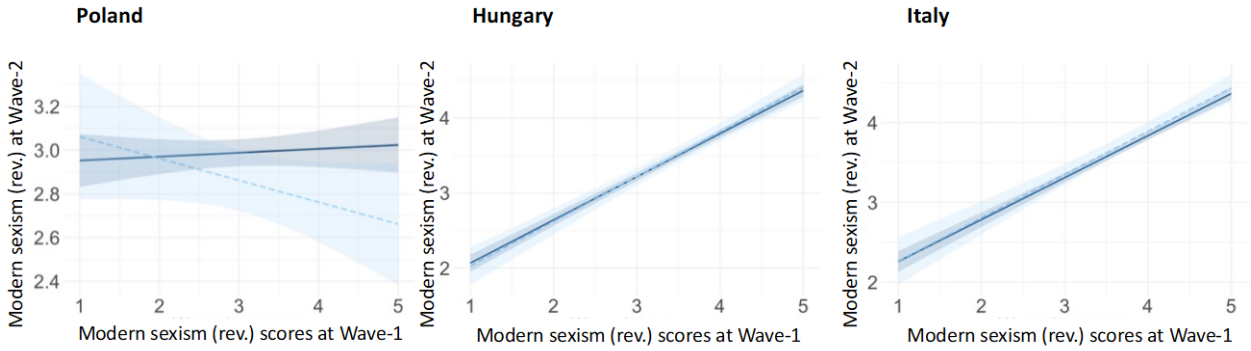


Figure 16. Predicted values (Wave-2) by treatment condition.
The figure displays predicted values of Wave-2 modern sexism (reverse-coded so that higher values indicate greater recognition of discrimination) across the full range of Wave-1 scores, based on regression models that interact treatment assignment with baseline attitudes. The item "Discrimination against women is no longer a problem in [COUNTRY]" was measured in both waves. The light blue lines show predicted values for respondents in the control track who viewed unchallenged anti-gender posts, while the darker blue line shows predictions for respondents who viewed counter-arguments. Shaded bands represent 95 percent confidence intervals. In Poland, the two lines diverge: the control line slopes downward, indicating that respondents with higher baseline recognition of discrimination show lower recognition at Wave-2 when exposed to unchallenged anti-gender content, suggesting an erosion of recognition over time. In contrast, the treatment line is flat, meaning that counter-arguments prevent this downward shift among respondents who start out with higher awareness of discrimination. The difference between the downward-sloping control line and the stable treatment line shows that, in Poland, counter-arguments play a protective role for individuals already sensitive to discrimination. In Hungary and Italy, both the control and treatment lines rise in parallel, meaning that respondents' Wave-2 recognition of discrimination mirrors their initial Wave-1 levels regardless of treatment assignment. This indicates no treatment effects in these contexts.

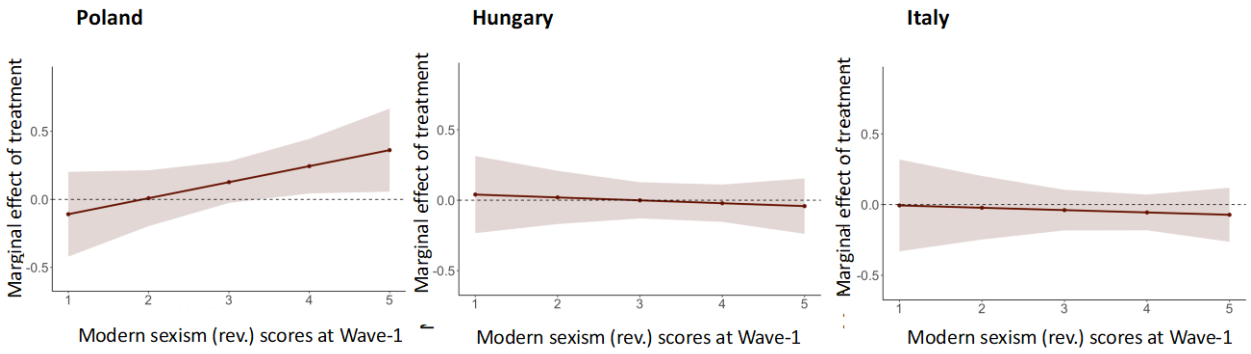


Figure 17 Marginal effect of treatment across the distribution of Wave-1 values.
The figure plots the marginal effect of receiving counter-arguments on Wave-2 recognition of discrimination across the full distribution of Wave-1 scores. The outcome is the reverse-coded modern sexism item "Discrimination against women is no longer a problem in [COUNTRY]," where higher values indicate greater recognition of discrimination. The solid line represents the estimated treatment effect at each baseline value, the shaded band shows the 95 percent confidence interval, and the horizontal line at zero marks the threshold indicating no effect. In Poland, the marginal effect becomes increasingly positive as Wave-1 recognition of discrimination increases, and at higher baseline values the effect is statistically significant because the confidence interval lies entirely above zero. This pattern indicates that counter-arguments help respondents who already recognised discrimination at Wave-1 maintain or increase this recognition in Wave-2, whereas exposure to unchallenged anti-gender content in the control group is associated with declining recognition among the same subgroup. In Hungary and Italy, the marginal effect remains flat and close to zero across the entire baseline range, with confidence intervals consistently overlapping zero, indicating no evidence that the impact of treatment varies by prior levels of modern sexism in these two contexts.

We expected that the short experimental intervention would only have a limited impact on threat perceptions, gender equality, and gender equality related attitudes more generally. This is indeed what we find, and why we did not replicate the full set of batteries from Wave-1 (e.g. perspective taking battery) but only include one modern sexism item in both waves. However, Wave-1 attitudes might serve as important screen that shapes how people perceive anti-gender communication, and this screen might in fact moderate the effectiveness of the counter arguments. We can leverage Wave-1 to gain a better understanding of the heterogeneity of treatment effects. We therefore interacted treatment assignment with misogyny, zero-sum threat perceptions, perspective taking, and involvement, i.e. willingness to support a petition for gender quotas. The results are presented graphically in the Appendix (Figures A1-A3).

Across these models, we do not find evidence that treatment effects vary systematically by any of our measures (misogyny, zero-sum threat perceptions, perspective taking, and involvement). The marginal effect plots for each moderator show flat lines with confidence intervals overlapping zero. Neither misogyny, threat perceptions, perspective taking, nor willingness to engage significantly conditioned the effect of counter-arguments on gender system justification or gender equality attitudes.

6.2 Post-level results: counter-arguments strategies

Overview

- Section 6.2 examines whether counter-arguments influence how respondents evaluate the specific anti-gender posts shown during the experiment. These post-level measures capture respondents' immediate reactions, such as agreement with the post and whether they perceive it as containing sexist claims.
- Respondents were assigned either to the control track, where they saw unchallenged anti-gender content, or to a treatment track, where the same content was presented with counter-arguments. After each post, respondents rated the message on several post-level dimensions.
- In contrast to the attitude-level results (Section 6.1), treatment effects emerge at the post level. Counter-arguments reduce agreement with anti-gender messages and increase recognition that the posts challenge gender equality.
- These effects vary across countries: they are strongest and most consistent in Poland, appear in a more selective pattern in Italy, and are very limited in Hungary.
- No single counter-argument strategy consistently outperforms the others across all countries and themes.
- While counter-arguments do not shift attitudes, they do affect how respondents perceive anti-gender content, limiting the persuasive impact of such messages at the post level.

Objective O6.2 concerns whether counter-arguments can mitigate the influence of anti-gender discourse at the level of individual posts. Whereas the attitudinal outcomes examined in Section 6.1 capture broad and relatively stable dispositions, the outcomes in this section are post-level evaluations: how participants perceive, agree with, and would engage with specific anti-gender messages when these appear unchallenged or accompanied by counter-arguments. This design allows us to assess more immediate communication effects. Even if core attitudes are resistant to change in the short run, counter-arguments may still alter how people respond to anti-gender posts in the moment of exposure.

6.2.1 Detection of sexist intention

We examine several outcomes at the level of posts, i.e. after people see each of the anti-gender posts we use in this study. First, detection of sexist intentions, which captures whether participants perceive the post as unfair, threatening, or poorly intentioned. Second, agreement with the post, which measures whether participants endorse its content, with an exploratory behavioural item asking whether they would "like" the post on social media.

For each post, participants completed a six-item scale measuring perceived intent and fairness (e.g., "The post is well-intentioned" [reversed], "The post poses a threat to gender equality"). Responses ranged from 1 (strongly disagree) to 6 (strongly agree). Reliability was high across countries.

When averaging across all posts, counter-arguments increased detection of sexism among Polish participants. The effect was consistent across most strategies, with the exception of

constructivism, which approached but did not reach conventional significance. By contrast, in Hungary and Italy, counter-arguments did not significantly affect sexism detection.
Looking more closely at individual posts reveals a more differentiated picture. In the case of abortion-related content, empathy, constructivist, and fairness counter-arguments increased detection of sexism among Italian participants, while empathy, constructivist, humour, and fairness counter-arguments did so among Polish participants. For the traditional family frame, multiple strategies again raised sexism detection in Poland, but had no effect in Hungary or Italy. A similar pattern held for the gender roles post, where norms-based counters increased detection in Poland but had no impact elsewhere.

Some exceptions emerged. For the gender lobby post, constructivist counters actually reduced sexism detection in Hungary and Italy, suggesting that some strategies may backfire depending on the topic and context. For the zero-sum post, none of the counter-arguments produced significant effects in any country.

Taken together, these results show that counter-arguments can increase recognition of sexist intent in some contexts, particularly in Poland, and especially on salient issues such as abortion or traditional family roles. However, effects are not uniform across countries or topics, and certain strategies may even reduce recognition in some cases.

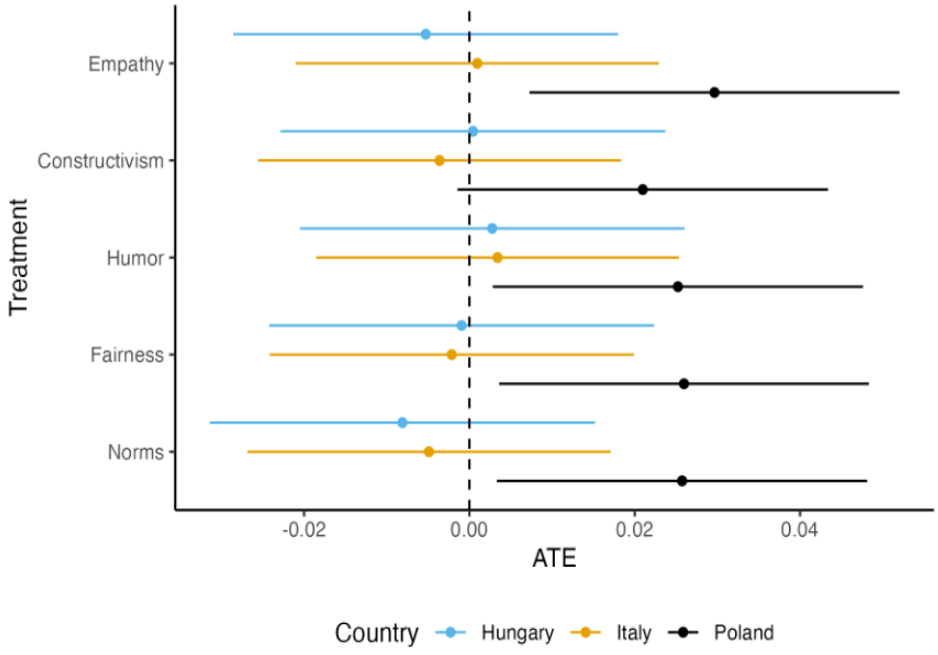


Figure 18. Effect of treatment on detection of sexism by strategy averaged across posts
The figure presents the average treatment effects (ATEs) of each counter-argument strategy on respondents' detection of sexism, averaged across the five anti-gender posts included in the experiment. Dots indicate point estimates and horizontal lines indicate 95 percent confidence intervals. Results are shown separately for Hungary, Italy and Poland. Detection of sexism is measured using items that capture whether a post is perceived as threatening gender equality, unfair, divisive, or promoting equality and fairness. Higher scores indicate greater detection of sexism. Across countries, most estimated effects are small and confidence intervals overlap zero, indicating limited change in sexism detection for most strategies. In Poland, several strategies show modest positive point estimates. In Hungary and Italy, the effects are close to zero for all strategies.

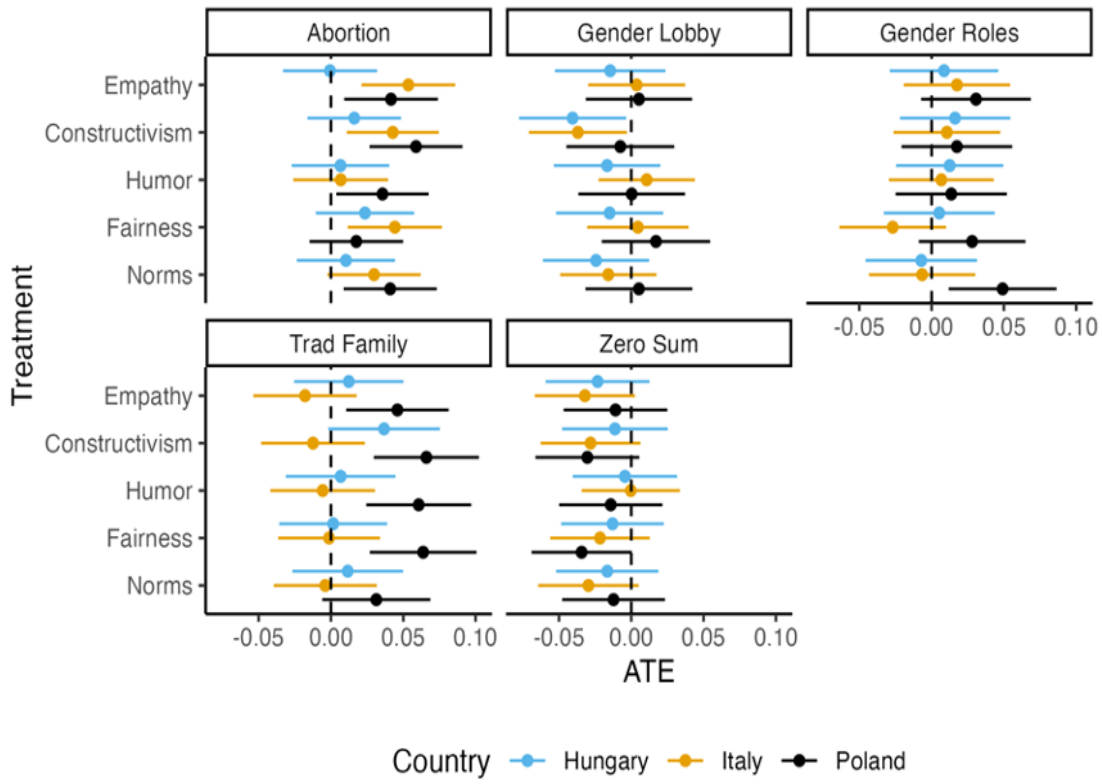


Figure 19. Effects of treatment on detection of sexism by post and strategy
The figure presents average treatment effects (ATEs) of each counter-argument strategy on respondents' detection of sexism, shown separately for the five anti-gender posts (Abortion, Gender Lobby, Gender Roles, Traditional Family, Zero-Sum). Dots indicate point estimates and horizontal lines indicate 95 percent confidence intervals. Results are shown separately for Hungary, Italy and Poland. Detection of sexism is measured using items that assess whether a post is perceived as threatening gender equality, divisive, unfair or promoting equality and fairness. Higher values indicate greater detection of sexism. Most estimated effects cluster around zero, but some consistent patterns emerge. For several posts, particularly *Abortion* and *Traditional Family*, multiple strategies yield positive effects, indicating increased recognition of sexism relative to the control condition. For example, in the context of communication on abortion, the empathy and constructivism strategies show statistically significant effects in Italy and Poland. For communication on traditional family roles, all strategies except the norms-based strategy show positive effects in Poland.

6.2.2 Post agreement

Agreement with the post was measured using a six-point scale (1 = completely disagree to 6 = completely agree), complemented by a behavioural item asking whether participants would "like" the post.

At the aggregate level, counter-arguments reduced agreement with anti-gender content in Poland. Constructivist, humour, and fairness counters were most effective. In Hungary and Italy, no consistent overall effects were observed.

Disaggregated by topic, however, clear patterns emerge. In the abortion post, fairness counters reduced agreement in Hungary and Poland, while in Italy nearly all strategies (empathy, constructivism, fairness, norms) reduced agreement. In the traditional family post, humour reduced agreement in Hungary, and constructivist, humour, and fairness reduced agreement in Poland. For the gender roles post, norms-based counters lowered agreement in Poland but had no effect elsewhere. For the gender lobby post, humour reduced agreement in Poland but not in Hungary or Italy. Again, the zero-sum post showed no effects across any country.

These findings map directly onto Objective O6.2. While entrenched attitudes remain stable in the short run, counter-arguments can alter how people evaluate individual anti-gender messages. Respondents exposed to counter-arguments were more likely to recognise sexism and less likely to agree with or endorse the posts, particularly in Poland and on topics such as abortion and traditional family roles. The absence of consistent effects in Hungary and Italy, as well as occasional backfire effects, highlights the importance of context. Counter-arguments are not a universal remedy; their success depends on the issue at stake, the framing of both the anti-gender message and the counter-argument, and the national context in which it is received.
Overall, the results show that counter-arguments have the potential to weaken the immediate persuasive power of anti-gender communication. Even if this does not translate into measurable changes in core attitudes within the timeframe of our study, it matters for the dynamics of public discourse. By reducing agreement and increasing recognition of sexist intent at the post-level, counterspeech can change which narratives gain traction in online spaces. Over time, repeated exposure to such interventions may accumulate into broader attitudinal change, which is a possibility that our results underline as a key avenue for future research.

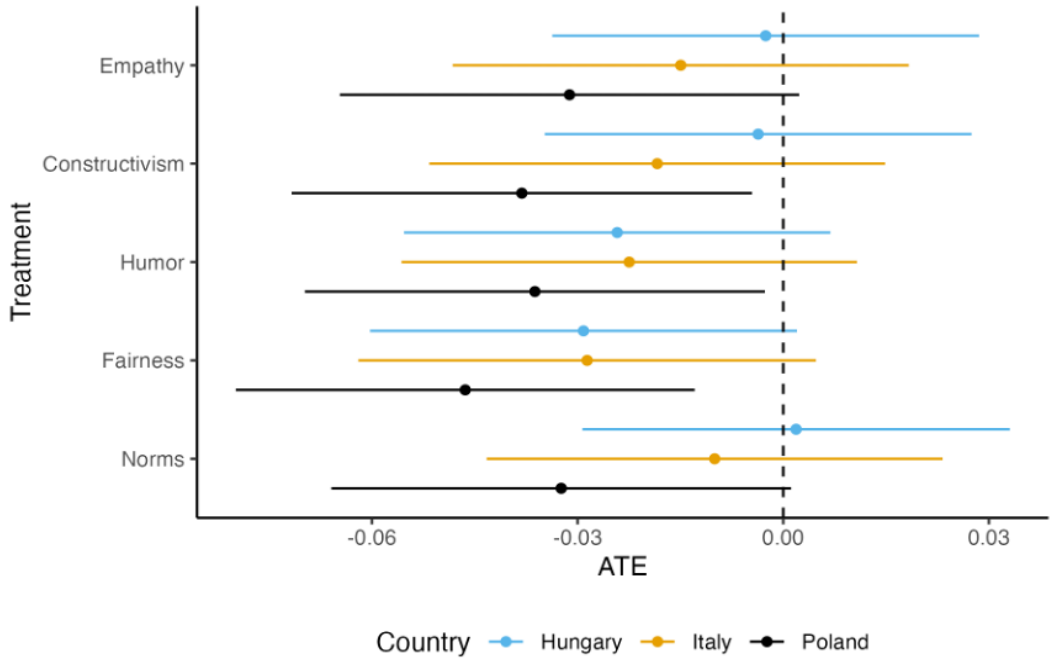


Figure 20. Effects of treatment on agreement with anti-gender content by strategy, averaged across posts

The figure presents average treatment effects (ATEs) of each counter-argument strategy on respondents' agreement with the anti-gender posts, averaged across all five posts included in the experiment. Dots indicate point estimates and horizontal lines represent 95 percent confidence intervals. Results are shown separately for Hungary, Italy and Poland. Agreement is measured on a six-point scale (1 = strongly disagree to 6 = strongly agree), with lower values indicating weaker agreement with the anti-gender content. Across countries, most estimated effects are small and negative, indicating that counter-arguments tend to reduce agreement with anti-gender content relative to the control condition. In Hungary and Italy, the empathy, constructivism and humour strategies show consistently negative point estimates, suggesting reduced agreement on average. In Poland, effects are also negative for several strategies, including humour and fairness, although estimates remain modest in size.

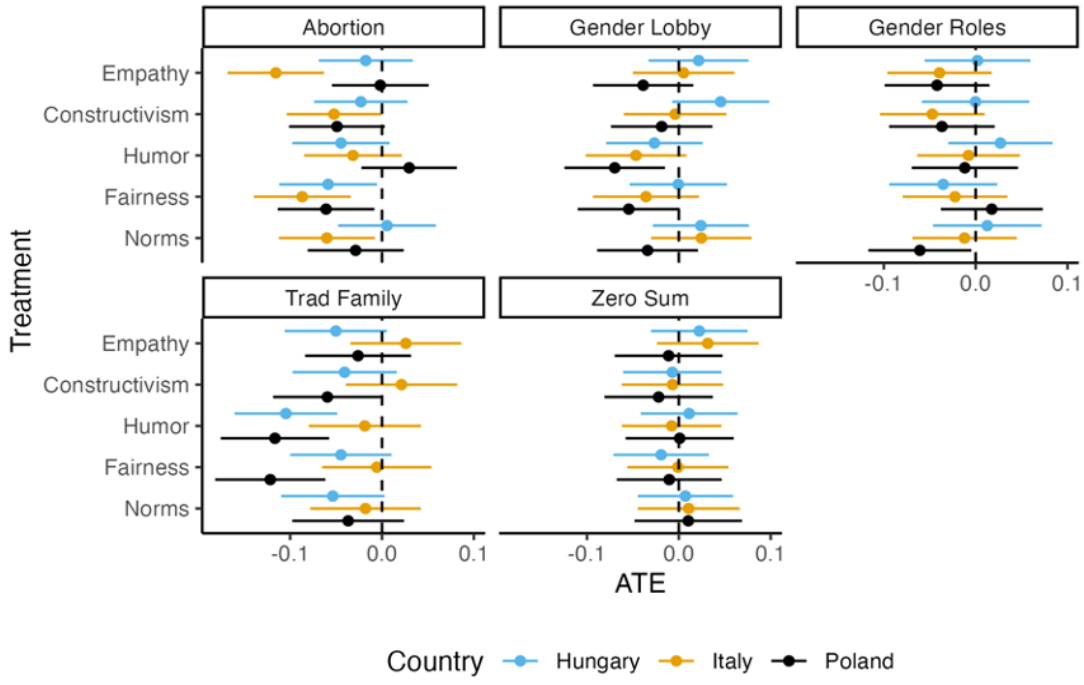


Figure 21. Effects of treatment on agreement with anti-gender content by post and strategy
The figure presents average treatment effects (ATEs) of each counter-argument strategy on respondents' agreement with the anti-gender posts, shown separately for the five posts (Abortion, Gender Lobby, Gender Roles, Traditional Family, Zero-Sum). Dots indicate point estimates and horizontal lines represent 95 percent confidence intervals. Results are reported separately for Hungary, Italy and Poland. Agreement is measured on a six-point scale (1 = strongly disagree to 6 = strongly agree), with lower scores indicating reduced agreement with the anti-gender message. Effects vary by post and strategy, but several consistent patterns emerge. Across countries, many strategies yield negative point estimates, suggesting that counter-arguments often reduce agreement with anti-gender content relative to the control condition. These patterns are particularly visible for the *Abortion* and *Gender Lobby* posts, where strategies such as empathy, constructivism and humour frequently produce negative effects across multiple country samples. In Poland, the fairness strategy also shows negative effects for several posts.

6.2.3 Effects on norm perceptions

A third outcome concerns perceived normativity, or the extent to which participants believed that others in their country would endorse the anti-gender posts. After evaluating each post, respondents estimated what percentage of people in their country would agree with the message on a 0–100% scale. This measure captures a different dimension of influence: not whether respondents themselves agree, but whether they perceive sexist discourse as socially widespread and acceptable.

Across country samples, a striking and unexpected pattern emerged. Exposure to counter-arguments consistently increased the perceived prevalence of agreement with anti-gender posts. In other words, counter-arguments did not make participants think "fewer people would agree

with this," but rather the opposite. This pattern held across nearly all posts and strategies, with only minor exceptions.

To unpack this finding, we examined how effects varied depending on respondents' prior endorsement of sexist attitudes. In Poland, interactions between treatment and baseline attitudes were consistently significant. Counter-arguments raised perceived normativity across the board, but the increase was strongest among participants who were initially least supportive of sexist content. For these respondents, the presence of counter-arguments appeared to signal that anti-gender discourse is common enough to require challenge, thereby inflating their estimates of societal prevalence. Similar though less consistent patterns appeared in Italy, particularly for empathy- and fairness-based counters, where low-baseline respondents again reported the sharpest increases in perceived normativity. In Hungary, by contrast, baseline attitudes did not moderate the effect of treatments, and no systematic interactions emerged.

Disaggregated by post, the same structure repeats. For example, in Poland, humour-based counters increased perceived normativity of both gender roles and traditional family posts, with the strongest effects again among respondents who were least supportive of those frames at baseline. In Italy, empathy and constructivist counters increased perceived normativity of the gender roles post among those low in traditional gender-role endorsement, while fairness and humour had similar effects on abortion and zero-sum posts. These results suggest that when counter-arguments are present, audiences who already reject anti-gender arguments may infer that anti-gender narratives are widespread in the population: otherwise, why would they need to be challenged?

This dynamic highlights a paradoxical side effect of counter-arguments: while it can reduce personal agreement with anti-gender content and increase recognition of its sexism, it may simultaneously amplify perceptions of its social reach. For audiences opposed to anti-gender rhetoric, this could strengthen perceptions of being in the minority, potentially normalising the idea that such rhetoric is more widely accepted than it truly is.

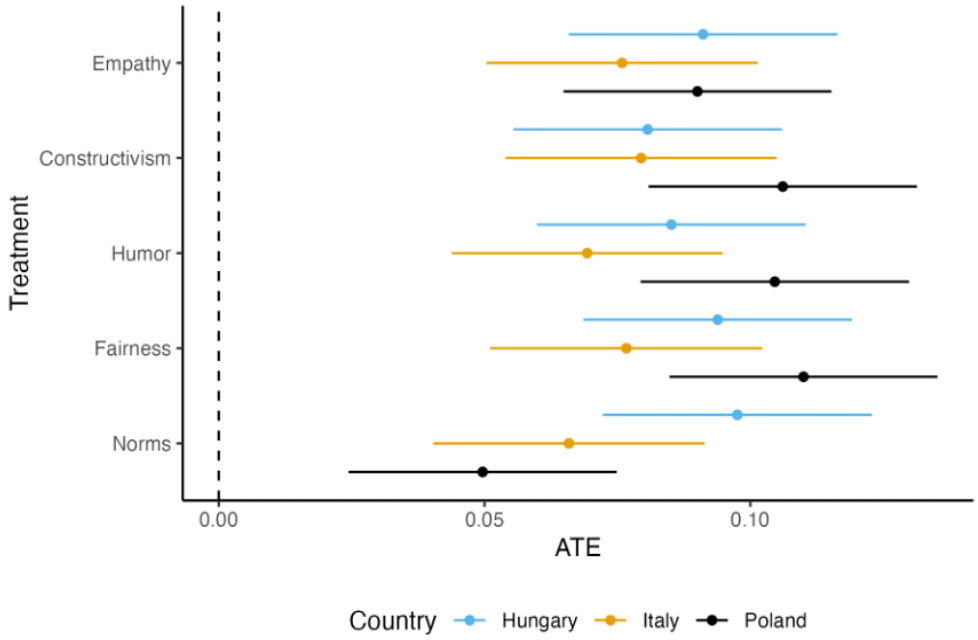


Figure 22. Effects of treatment on perceived normativity by strategy, averaged across posts
The figure presents average treatment effects (ATEs) of each counter-argument strategy on respondents' perceptions of how widely supported the anti-gender posts are in their country. Results are averaged across all five posts and shown separately for Hungary, Italy and Poland. Dots represent point estimates and horizontal lines indicate 95 percent confidence intervals. Perceived normativity is measured using a 0–100 scale asking respondents what percentage of people in their country they believe would agree with the post. Across countries, all strategies produce positive point estimates, indicating that counter-arguments tend to increase perceptions of the share of people agreeing with the anti-gender content shown, relative to the control condition. The effects are strongest and most consistent in Poland, where empathy, constructivism, humour and fairness show positive estimates.

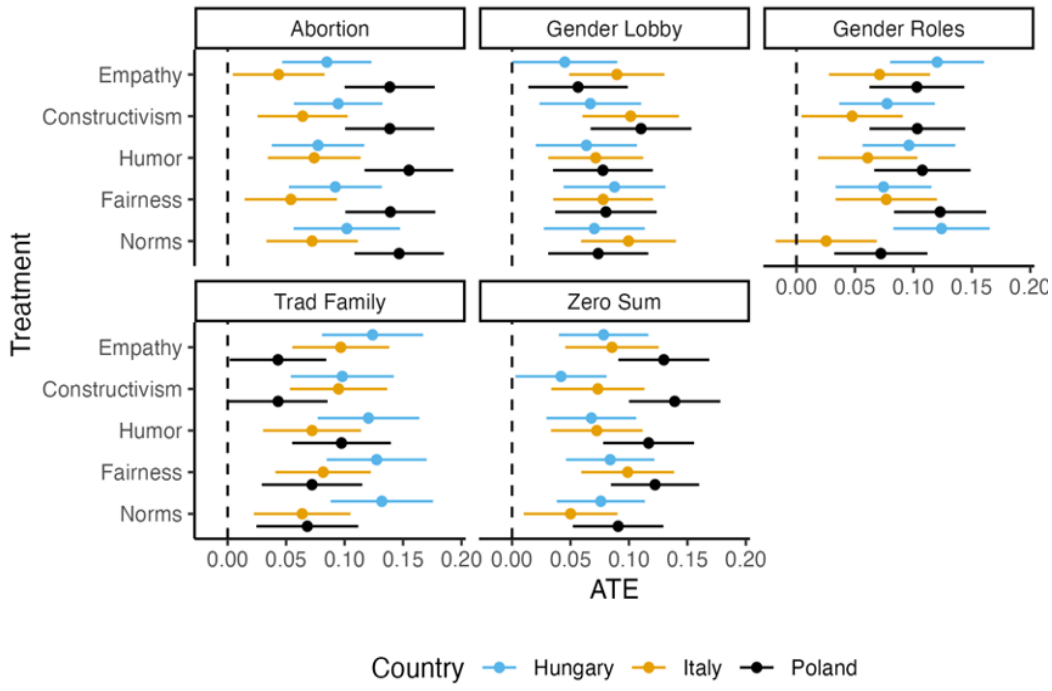


Figure 23. Effects of treatment on perceived normativity by post and strategy
The figure presents average treatment effects (ATEs) of each counter-argument strategy on respondents' perceptions of how widely supported the anti-gender posts are in their country, disaggregated by the five posts (Abortion, Gender Lobby, Gender Roles, Traditional Family, Zero-Sum). Dots represent point estimates and horizontal lines indicate 95 percent confidence intervals. Results are shown separately for Hungary, Italy and Poland. Perceived normativity is measured using a 0–100 scale asking respondents what percentage of people in their country they believe would agree with the post. Across posts and strategies, most point estimates are positive, indicating that counter-arguments tend to increase perceptions of how many people in the country agree with the anti-gender content relative to the control condition.

6.3 Implications

The analyses presented in this section speak directly to Objective O6.2 of Work Package 6: whether counter-arguments can mitigate the influence of anti-gender discourse. Taken together, the results suggest a nuanced picture. Counter-arguments can be effective in shaping how anti-gender messages are received, but its effects are conditional, issue-specific, and not uniformly positive. On one side, counter-arguments reduced agreement with anti-gender content and, in many cases, heightened recognition of its sexist intent. These are central indicators of counter-argument effectiveness: they show that exposure to well-crafted counter-arguments makes respondents more likely to view anti-gender communication as harmful and less likely to endorse it. Such effects were particularly evident in Poland, where counter-arguments consistently reduced agreement and increased detection of sexism on posts dealing with abortion and traditional family roles. Some effects were also observed in Italy, especially for abortion-related content, and in Hungary in isolated cases such as humour-based responses to traditional family frames.

At the same time, the effects were not uniform. In many instances, counter-arguments had no measurable impact, particularly in Hungary and, to a lesser extent, Italy. Moreover, we observed occasional backfire effects: for example, constructivist counters reduced sexism detection on the gender lobby post in both Hungary and Italy. These findings underline that not all strategies are equally effective across contexts, and that the same strategy can resonate in one setting while failing or even backfiring in another.

The most striking and unexpected finding concerns perceived normativity. Across countries and strategies, exposure to counter-arguments consistently increased the perceived share of the population that would agree with the anti-gender post. This suggests that counter-arguments, while reducing personal agreement, may simultaneously convey the impression that anti-gender discourse is widespread. This dynamic was especially pronounced among respondents who themselves rejected sexist content: for them, the very presence of counter-arguments appeared to signal that such discourse is socially prevalent and therefore socially powerful.

From the perspective of Objective O6.2, these results highlight both the promise and the complexity of counter-argument strategies. Counter-argument do influence how people evaluate anti-gender content: they make sexism more visible and reduce endorsement. Yet, they may also shape perceptions of the social environment in ways that may be counterproductive, reinforcing the impression that anti-gender narratives enjoy broad support. This duality suggests that counter-arguments can change the reception of anti-gender discourse at the micro-level of message evaluation, but whether these changes accumulate into broader societal shifts depends on how such interventions are framed and contextualised.

In sum, counter-arguments emerge as a tool with real but limited potential. They do not shift attitudes in the short run, but they do alter how individual messages are perceived in the moment of exposure. Their effects are strongest in contexts where anti-gender discourse is politically salient and contested, as in Poland. However, counter-arguments may also unintentionally contribute to the perceived normalisation of anti-gender views. Future work should focus on calibrating counter-argument strategies so that they not only disarm individual messages but also avoid reinforcing the illusion of broad societal consensus behind them.

7. Limitations, discussion, conclusion

Overview

- This section highlights the key conclusions from WP6 and summarises the main limitations.
- The findings show that gender-related attitudes are stable and do not shift in the short term in response communication, even when respondents are exposed to anti-gender messages or counter-arguments (Objective O6.1).
- Counter-arguments nevertheless influence how anti-gender communication is perceived, reducing agreement with anti-gender content and increasing recognition of sexist claims (Objective O6.2).
- The section highlights that AI can play a role in the design of counter-arguments, which is important for those who might want to get involved in a counter discourse.
- A key limitation of the study is that the experiment tests short-term reactions; longer-term effects on attitudes or behaviour remain outside the scope of this design.
- Overall, the results indicate that counter-arguments can shape immediate message reception, even when attitudes remain unchanged, offering valuable insights for communication strategies responding to anti-gender narratives.

Work Package 6 set out to answer two key questions: what are the causal effects of anti-gender egalitarian communication on citizens (Objective O6.1), and can counter-strategies mitigate these effects (Objective O6.2)? Our findings provide a nuanced picture that can be understood as operating across different processes: long-term change, short-term change, and immediate information-processing effects.

The baseline survey in eight countries (Wave-1) documented clear cross-national differences in attitudes toward gender equality, perceptions of threat, perspective-taking, and willingness to engage politically. These differences align with broader contextual factors: higher levels of misogyny, zero-sum thinking, and lower perspective-taking were observed in countries with greater far-right support and higher structural gender inequality. These patterns underscore the importance of context for understanding the reception of anti-gender communication. While our design cannot directly speak to long-term processes of change, the descriptive cross-national evidence highlights that political and structural environments shape the climate in which gender equality is supported or contested.

The panel design in Hungary, Italy, and Poland (Wave-1–Wave-2) allowed us to test whether exposure to unchallenged anti-gender communication or to counter-arguments produced measurable shifts in attitudes over a matter of weeks (Objective O6.1). The results show stability. Recognition of discrimination against women, used as a repeated measure, did not shift significantly between waves, irrespective of treatment assignment. Even in Poland, where Wave-2 coincided with a polarised presidential election campaign, there was little evidence of aggregate attitudinal change. This suggests that attitudes toward gender equality are not easily swayed in the short run: they function as dispositions rather than volatile opinions. However, we did find one important qualification: in Poland, counter-arguments prevented erosion among respondents who already recognised discrimination at baseline. Exposure to unchallenged anti-

gender posts reduced recognition of discrimination among these respondents, whereas exposure to counter-arguments "protected" their views. In this sense, counter-arguments may not move attitudes, but can serve as a shield against decline in politicised environments.

While attitudes proved stable, counter-arguments did alter how participants processed and evaluated individual messages, which speaks to Objective O6.2. Across our experimental treatments, counter-arguments reduced agreement with anti-gender communication and, in many cases, increased recognition of sexist intent. These effects were strongest in Poland, where multiple strategies (including fairness, empathy, and humour) reduced agreement and heightened detection of sexism, particularly on posts about abortion and traditional family roles. Some effects also appeared in Italy (especially for abortion-related content) and in Hungary in specific cases. Crucially, no single strategy consistently outperformed the others.

This study, therefore, offers a nuanced understanding of how different counter-argument strategies influence audience responses to anti-gender egalitarian communication. Drawing on preregistered experiments across three national contexts, we tested five distinct strategies (leveraging empathy, constructivism, humour, fairness, and norms) against salient anti-gender narratives. While counter-arguments invoking fairness, empathy, or humour often showed promise, their effects were not consistent across national contexts. In countries where one strategy was effective, other strategies typically were too; where one had no effect, none did. Exploratory analyses also showed that for a given topic, multiple strategies either proved effective together or had no effect. These patterns indicate that context matters more than the specific strategy in determining counter-argument effectiveness.

Importantly, the counter-arguments we tested were generated with the help of GenAI, then curated and adapted for the study. People might be able to come up with reactions to anti-gender egalitarian messages on their own. Our results show that there is a way to leverage GenAI in this context. We do not claim that what GenAI produces is better than what people might write themselves. In fact, experts and those with practice in writing strong reactions might write stronger text than GenAI. We only suggest that there is a way for GenAI to help those who might find that kind of support helpful. The opportunity to get this kind of support might lower the threshold for those who might otherwise not react to anti-gender egalitarian messages. Thus, the ease of generation content with GenAI may empower a broader segment of the public to respond to anti-gender rhetoric, even if final tailoring by humans remains essential.

At the same time, our findings reveal an unexpected paradox. Across countries and strategies, counter-arguments consistently increased participants' estimates of how many people in their society would agree with the anti-gender post. While counter-arguments reduced personal agreement and enhanced recognition of sexism, it simultaneously conveyed the impression that anti-gender discourse is socially widespread. This effect was especially pronounced among egalitarian respondents in Poland and Italy, suggesting that those most inclined to challenge sexist rhetoric were also most vulnerable to misinterpreting its social prevalence. This pattern resonates with broader work on corrective communication, where rebuttals can inadvertently signal that a contested claim is socially salient and widely endorsed. Our inclusion of a direct measure of perceived societal support allowed us to capture this counterintuitive outcome, which is often overlooked in other studies. Future work should incorporate such measures more

routinely and explore strategies to mitigate these unintended consequences, for example, by pairing counter-arguments with explicit cues that anti-gender positions are minority views.

Several limitations of our study must be acknowledged. First, our design presented posts and counter-arguments in an experimental survey rather than embedding them in real social media feeds. While necessary for causal identification, this abstraction reduced ecological validity. Second, we did not vary the identity of the speakers. In real-world communication, the persuasiveness of counter-arguments depends heavily on who delivers it: whether an ally, an opponent, or a trusted public figure. Third, while our focus on Hungary, Italy, and Poland maximises relevance in contexts where anti-gender narratives are highly salient, it leaves open how counter-arguments might function in settings where such narratives are less prominent or less politically mainstream. Extending this work to a broader range of national contexts remains essential.

Taken together, our results demonstrate both the promise and the limits of counter-arguments. Counter-arguments do not shift attitudes in the short run, but they do alter how anti-gender messages are received in the moment of exposure. They reduce acceptance and increase recognition of sexism. At the same time, they may unintentionally reinforce perceptions of the widespread acceptance of anti-gender views. The broader implication is that attitudes and message evaluations operate at different levels of malleability: attitudes are stable dispositions, while message evaluations are more flexible and constructed when processed. Countering should therefore be seen not as a universal remedy but as a contextual tool. When used with attention to unintended effects, it can help shape the public discourse around gender equality. Over time, repeated counterspeech interventions may accumulate into meaningful attitudinal change, though future research needs to test this.

References

Álvarez-Benjumea, A. (2024). Social Norms and the Expression of Prejudice: How the Norm Changes. *Current Opinion in Psychology*, 101974. https://doi.org/10.1016/j.copsyc.2024.101974

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221. https://doi.org/10.1073/pnas.1804840115 .

Balázs, Zs. (2018, August 10). Nemcsak a genderszakokat, a szabad oktatáshoz való alkotmányos jogot is veszélyezteti a magyar kormány terve. *Qubit*. https://qubit.hu/2018/08/10/nemcsak-a-genderszakokat-a-szabad-oktatashoz-valo-alkotmanyos-jogot-is-veszelyezteti-a-magyar-kormany-terve

Batson, C. D. (2009). These things called empathy: Eight related but distinct phenomena. In J. Decety & W. Ickes (Eds.), *The social neuroscience of empathy* (pp. 3–15). Boston Review. https://doi.org/10.7551/mitpress/9780262012973.003.0002

Batson, C. D., Polycarpou, M. P., Harmon-Jones, E., Imhoff, H. J., Mitchener, E. C., Bednar, L. L., Klein, T. R., & Highberger, L. (1997). Empathy and attitudes: Can feeling for a member of a stigmatized group improve feelings toward the group? *Journal of Personality and Social Psychology, 72*(1), 105–118. https://doi.org/10.1037/0022-3514.72.1.105

Bian, L., Leslie, S. J., & Cimpian, A. (2017). Gender stereotypes about intellectual ability emerge early and influence children's interests. *Science*, *355*(6323), 389-391. https://doi.org/10.1126/science.aah6524

Bicchieri, C., Dimant, E., Gachter, S., & Nosenzo, D. (2022) Social proximity and the erosion of norm compliance. *Games and Economic Behavior 132*, 59-72. https://doi.org/10.1016/j.geb.2021.11.012

Bilewicz, M., Tempska, P., Leliwa, G., Dowgiałło, M., Tańska, M., Urbaniak, R., & Wroczyński, M. (2021). Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment. *Aggressive Behavior, 47*(3), 260–266. https://doi.org/10.1002/ab.21948

Blum, A. (2015). Germany. In: E. Kovats & M. Poim (Eds.) *Gender as Symbolic Glue*. Foundation for European Progressive Studies, pp.40-61.

Boykoff, M. T., & Boykoff, J. M. (2004). Balance as bias: Global warming and the US prestige press. *Global environmental change*, *14*(2), 125-136. https://doi.org/10.1016/j.gloenvcha.2003.10.001

Broemer, P., & Grabowski, A. (2018). Nature vs. nurture and the flexibility of gender stereotypes: Counterstereotypical information can both diminish and enhance ingroup stereotyping. *Polish Psychological Bulletin*, *49*(2), 251-261. https://doi.org/10.24425/119493

Chan, M. P. S., Jones, C., & Albarracín, D. (2017). Countering false beliefs: An analysis of the evidence and recommendations of best practices for the retraction and correction of scientific misinformation. In *The Oxford Handbook of the Science of Science Communication* (pp. 341-350). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190497620.013.37

Chaney, K. E., Sanchez, D. T., Alt, N. P., & Shih, M. J. (2021). The breadth of confrontations as a prejudice reduction strategy. *Social Psychological and Personality Science, 12*(3), 314–322. https://doi.org/10.1177/1948550620919318

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology, 55*(1), 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of personality and social psychology*, *58*(6), 1015-1026. https://psycnet.apa.org/doi/10.1037/0022-3514.58.6.1015

Cohen-Chen, S., Dhensa-Kahlon, R. K., & Hameiri, B. (2024). Humorous responses to gender injustice: The contrasting effects of efficacy and emotions on women's collective action intentions. *Sex Roles, 90*(2), 336–351. https://doi.org/10.1007/s11199-024-01459-0

Constantino, S. M., Sparkman, G., Kraft-Todd, G. T., Bicchieri, C., Centola, D., Shell-Duncan, B., Vogt, S., & Weber, E. U. (2022). Scaling up change: A critical review and practical guide to harnessing social norms for climate action. *Psychological Science in the Public Interest, 23*(2), 50–97. https://doi.org/10.1177/15291006221105279

Crabtree, C., Holbein, J., Bosley, M., & Sevi, S. (2025). Can AI help reduce prejudice? Evaluating the effectiveness of ai-powered personalized persuasion on support for transgender rights. *SSRN*. https://doi.org/10.2139/ssrn.5229084

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. JSAS Catalog of Selected Documents in Psychology, 10, 85. https://doi.org/10.1037/t01093-000

Davis, M. H. (1983). The effects of dispositional empathy on emotional reactions and helping: A multidimensional approach. *Journal of Personality, 51*(2), 167–184. https://doi.org/10.1111/j.1467-6494.1983.tb00860.x

De Souza, L., & Schmader, T. (2022). The misjudgment of en: Does pluralistic ignorance inhibit allyship? *Journal of Personality and Social Psychology, 122*(2), 265. https://doi.org/10.1037/pspi0000362

Dobai, A., & Hopkins, N. (2020). Humour is serious: Minority group members' use of humour in their encounters with majority group members. *European Journal of Social Psychology, 50*(2), 448–462. https://doi.org/10.1002/ejsp.2612

Dömötör B., & Horváth-Kávai, A. (2023, May 04) Orbán at CPAC Hungary: Hungary has developed the antidote against the deadly progressive virus. *Telex*. https://telex.hu/english/2023/05/04/orban-at-cpac-hungary-hungary-has-developed-the-antidote-against-the-deadly-progressive-virus

Eagly, A. H., & Wood, W. (2013). Feminism and evolutionary psychology: moving forward. *Sex Roles, 69*(9–10), 549–556. https://doi.org/10.1007/s11199-013-0315-y

European Commission: Directorate-General for Justice and Consumers, Gender Equality 2017, European Commission, 2017, https://data.europa.eu/doi/10.2838/431877

European Policy Centre. (2025, June 2). *2025 Presidential Elections in Poland: What are the Implications for the country and the EU?* European Policy Centre. https://www.epc.eu/publication/2025-presidential-elections-in-poland-what-are-the-implications-for-the-country-and-the-eu/

Fasoli, F., Hopkins-Doyle, A., & Guizzo, F. (2025). Thank you for sharing: Body-positive and humour parody images increase women's body satisfaction and predict collective action intentions via gratitude. *Sex Roles, 91*(7), 48. https://doi.org/10.1007/s11199-025-01593-3

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, *114*(3), 817-868. https://doi.org/10.1162/003355399556151

Garaigordobil, M. (2014). Sexism and empathy: Differences as a function of sociodemographic variables and relations between both constructs. *Advances in psychology research*, *100*, 59-80.

Garbagnoli, S. (2016). Against the Heresy of Immanence: Vatican's 'Gender' as a New Rhetorical Device Against the Denaturalization of the Sexual Order. Religion and Gender, 6(2), pp.187–204. https://doi.org/10.18352/rg.10156

Garbagnoli, S., & Prearo, M. (2018). Il Movimento va al congresso. Reti di movimenti anti-gender tra dispositivi retorici, partecipazione dal basso, conoscenza e alleanze Politiche. *Polis, 32*(2), 323–338.

Gelfand, M. J., Gavrilets, S., & Nunn, N. (2024). Norm dynamics: Interdisciplinary perspectives on social norm emergence, persistence, and change. *Annual Review of Psychology, 75*, 341–378. https://doi.org/10.1146/annurev-psych-033020-013319

Gomez, E. M., & Kaiser, C. R. (2019). From pixels to protest: Using the internet to confront bias at the societal level. In R. K. Mallett & M. J. Monteith (Eds.), *Confronting prejudice and discrimination* (pp. 319–335). Academic Press. https://doi.org/10.1016/B978-0-12-814715-3.00011-4

Guess, A., McGregor, S., Pennycook, G., & Rand, D. (2024). Unbundling digital media literacy tips: Results from two experiments. *OSF*. https://doi.org/10.31234/osf.io/u34fp

Grabowska, M., (2022). From Soviet feminism to the European Union: Transnational women's movements between East and West. In: K. Fabian, J. E. Johnson & M. Lazda (Eds.), *The Routledge Handbook of Gender in Central-Eastern Europe and Eurasia*. Routledge, pp.145-153.

Graff, A & Korolczuk, E., 2022. *Anti-Gender Politics in the Populist Moment*. London/New York: Routledge.

Grzebalska, W., Kováts, E., & Pető, A. (2017). Gender as symbolic glue: How 'gender' became an umbrella term for the rejection of the (neo) liberal order. *Krytyka Polityczna*, 113.

Grzebalska, W., & Pető, A. (2018). The gendered modus operandi of the illiberal transformation in Hungary and Poland. *Women's Studies International Forum, 68*, 164–172. https://doi.org/10.1016/j.wsif.2017.12.001

Gutierrez, B., Kaatz, A., Chu, S., Ramirez, D., Samson-Samuel, C., & Carnes, M. (2014). "Fair Play": a videogame designed to address implicit race bias through active perspective taking. *Games for Health Journal, 3*(6), 371–378. https://doi.org/10.1089/g4h.2013.0071

Gwiazda, A. (2020). Right-wing populism and feminist politics: the case of Law and Justice in Poland. *International Political Science Review, 42*(5), 580–595. https://doi.org/10.1177/0192512120948917

Hampes, W. P. (2001). Relation between humor and empathic concern. *Psychological Reports, 88*(1), 241–244. https://doi.org/10.2466/pr0.2001.88.1.241

Hartman, R., Blakey, W., Womick, J., Bail, C., Finkel, E. J., Han, H., Sarrouf, J., Schroeder, J., Sheeran, P., Van Bavel, J. J., Willer, R., & Gray, K. (2022). Interventions to reduce partisan animosity. *Nature Human Behaviour, 6*(9), 1194–1205. https://doi.org/10.1038/s41562-022-01442-3

Haslam, N., Rothschild, L., & Ernst, D. (2002). Are essentialist beliefs associated with prejudice?. *British Journal of Social Psychology*, *41*(1), 87-100. https://doi.org/10.1348/014466602165072

Hylton, K. (2018). I'm not joking! The strategic use of humour in stories of racism. *Ethnicities, 18*(3), 327–343. https://doi.org/10.1177/1468796817743998

Ingram, K. M., & Campion, K. (2024). Of Heroes and Mothers: Locating Gender in Ideological Narratives of Salafi-Jihadist and Extreme Right Propaganda. *Studies in Conflict & Terrorism*, 1-27. https://doi.org/10.1080/1057610X.2024.2322758

Jost, J. T., & Kay, A. C. (2005). Exposure to benevolent sexism and complementary gender stereotypes: Consequences for specific and diffuse forms of system justification. *Journal of personality and social psychology*, *88*(3), 498-509.

Korolczuk, E. & Graff, A. (2017). "Worse than Communism and Nazism put together": War on Gender in Poland. In: R. Kuhar & D. Paternotte (Eds.), *Anti-Gender Campaigns in Europe Mobilizing Against Equality.* Rowman and Littlefield International Ltd, pp.175-194. DOI:10.5040/9798881809683.ch-010

Koszałkowska, K., & Wróbel, M. (2024). Traditionally identified women are amused with sexist jokes about men through benevolent but not hostile sexism. *HUMOR*, *37*(4), 551-576. https://doi.org/10.1515/humor-2024-0012

Kozyreva, A., Philipp Lorenz-Spreen, Herzog, S. M., Ullrich, Lewandowsky, S., Hertwig, R., Ali, A., Bak-Coleman, J., Sarit Barzilai, Basol, M., Berinsky, A. J., Betsch, C., Cook, J., Fazio, L. K., Geers, M., Guess, A. M., Huang, H., Larreguy, H., Maertens, R., & Folco Panizza. (2024). Toolbox of individual-level interventions against online misinformation. *Nature Human Behaviour*, *8*(6), 1044–1052. https://doi.org/10.1038/s41562-024-01881-0

Kuchynka, S. L., Bosson, J. K., Vandello, J. A., & Puryear, C. (2018). Zero-sum thinking and the masculinity contest: Perceived intergroup competition and workplace gender bias. *Journal of Social Issues*, *74*(3), 529-550. https://doi.org/10.1111/josi.12281

Kuhar, R., & Smrdelj, R. (2025). The rise of anti-gender mobilizations. In R. Smrdelj & R. Kuhar (Eds.), *Anti-gender mobilizations in Europe and the feminist response: Productive Resistance* (pp. 1–24). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-87693-6_1

Lavizzari, A. (2025). The anti-gender movement and the populist radical right in Italy: A symbiotic relationship. *European Journal of Politics and Gender, 8*(1), 207–233. https://doi.org/10.1332/25151088Y2024D000000047

Lavizzari, A., & Prearo, M. (2019). The anti-gender movement in Italy: Catholic participation between electoral and protest politics. *European Societies, 21*(3), 422–442. https://doi.org/10.1080/14616696.2018.1536801

Lee, H., & Rojas, H. (2025). Media or political affiliations? Political transition, government trust, and affective polarization in South Korea. *Asian Journal of Comparative Politics*. https://doi.org/10.1177/20578911251355739

Leshin, R. A., Leslie, S. J., & Rhodes, M. (2021). Does it matter how we speak about social kinds? A large, preregistered, online experimental study of how language shapes the development of essentialist beliefs. *Child Development, 92*(4), e531–e547. https://doi.org/10.1111/cdev.13527

Lewis, T., & Yoshimura, S. M. (2017). Politeness strategies in confrontations of prejudice. *Atlantic Journal of Communication, 25*(1), 1-16. https://doi.org/10.1080/15456870.2017.1251198

Li, A. H., Noland, E. S., & Monteith, M. J. (2025). Following prejudiced behavior, confrontation restores local anti-bias social norms. Personality and Social Psychology Bulletin, 51(12), 2389-2404., https://journals.sagepub.com/doi/abs/10.1177/01461672241229006

Lombardo, E., & Meier, P. (2022). Challenging boundaries to expand frontiers in gender and policy studies. *Policy & Politics*, *50*(1), 99-115. https://doi.org/10.1332/030557321X16309516650101

Mallett, R. K., Ford, T. E., & Woodzicka, J. A. (2016). What did he mean by that? Humor decreases attributions of sexism and confrontation of sexist jokes. *Sex Roles, 75*(5–6), 272–284. https://doi.org/10.1007/s11199-016-0605-2

Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications, 5*(1), 47. https://doi.org/10.1186/s41235-020-00252-3

Mayer, S & Sauer, B., 2017. "Gender Ideology" in Austria: Coalitions around an empty signifier. In: R. Kuhar & D. Paternotte (Eds.), *Anti-Gender Campaigns in Europe Mobilizing Against Equality*. Rowman and Littlefield International Ltd, pp.23-40.

Meyer, J. C. (2000). Humor as a double-edged sword: Four functions of humor in communication. *Communication theory*, *10*(3), 310-331. https://doi.org/10.1111/j.1468-2885.2000.tb00194.x

Montada, L., & Schneider, A. (1989). Justice and emotional reactions to the disadvantaged. *Social Justice Research, 3*(4), 313–344. https://doi.org/10.1007/BF01048081

Mortensen, C. R., Neel, R., Cialdini, R. B., Jaeger, C. M., Jacobson, R. P., & Ringel, M. M. (2019). Trending norms: A lever for encouraging behaviors performed by the minority. *Social Psychological and Personality Science*, *10*(2), 201-210. https://doi.org/10.1177/1948550617734615

Nicholas, L., & Agius, C. (2018). *The persistence of global masculinism: Discourse, gender and neo-colonial re-articulations of violence*. Palgrave MacMillan. https://doi.org/10.1007/978-3-319-68360-7

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior, 32*(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2

Paluck, E. L., Porat, R., Clark, C. S., & Green, D. P. (2021). Prejudice reduction: Progress and challenges. *Annual Review of Psychology, 72*(1), 533–560. https://doi.org/10.1146/annurev-psych-071620-030619

Paternotte, D. (2023). Victor Frankenstein and his creature: The many lives of 'gender ideology'. *International Review of Sociology, 33*(1), 80–104. https://doi.org/10.1080/03906701.2023.2187833

Paternotte, D & Kuhar, R., (2017). "Gender Ideology" in movement: Introduction. In: R. Kuhar & D. Paternotte (Eds.), *Anti-Gender Campaigns in Europe Mobilizing Against Equality.* Rowman and Littlefield International Ltd, pp.1-22.

Paternotte, D., & Kuhar, R. (2018). Disentangling and locating the "global right": Anti-gender campaigns in Europe. *Politics and Governance*, *6*(3), 6-19. https://doi.org/10.17645/pag.v6i3.1557

Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences, 116*(7), 2521–2526. https://doi.org/10.1073/pnas.1806781116

Pető, A. (2022). Reproductive Rights as Battlefield in the New Cold War: A Historical Comparison of Illiberal Gender Politics Regarding Reproductive Rights in Hungary. In A. Scheele, J. Roth, & H. Winkel (Eds.), Global Contestations of Gender Rights (pp. 227-244). Bielefeld: Bielefeld University Press. https://nbn-resolving.org/urn:nbn:de:0168- ssoar-78732-5.

Pettigrew T. F., & Tropp L. R. (2008). How does intergroup contact reduce prejudice? Meta-analytic tests of three mediators. *European Journal of Social Psychology, 38*(6), 922–934. https://doi.org/10.1002/ejsp.504

Powell, A. A., Branscombe, N. R., & Schmitt, M. T. (2005). Inequality as ingroup privilege or outgroup disadvantage: The impact of group focus on collective guilt and interracial attitudes. *Personality and Social Psychology Bulletin, 31*(4), 508–521. https://doi.org/10.1177/0146167204271713

Pratto, F., Çidam, A., Stewart, A. L., Zeineddine, F. B., Aranda, M., Aiello, A., ... & Henkel, K. E. (2013). Social dominance in context and in individuals: Contextual moderation of robust effects of social dominance orientation in 15 languages and 20 countries. Social

Psychological and Personality Science, 4(5), 587-599. https://doi.org/10.1177/1948550612473663

Prearo, M. (2020). *L'ipotesi neocattolica: Politologia dei movimenti anti-gender*. Mimesis. https://www.mimesisedizioni.it/libro/9788857570167

Renger, D., Eschert, S., Teichgräber, M. L., & Renger, S. (2020). Internalized equality and protest against injustice: The role of disadvantaged group members' self-respect in collective action tendencies. *European Journal of Social Psychology, 50*(3), 547–560. https://doi.org/10.1002/ejsp.2637

Rhodes, M., Leslie, S. J., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, *109*(34). https://doi.org/10.1073/pnas.1208951109

Riquelme, A. R., Carretero-Dios, H., Megías, J. L., & Romero-Sánchez, M. (2021). Joking for gender equality: Subversive humor against sexism motivates collective action in men and women with weaker feminist identity. *Sex Roles, 84*(1–2), 1–13. https://doi.org/10.1007/s11199-020-01154-w

Rooduijn, Matthijs, Andrea L.P. Pirro, Daphne Halikiopoulou, Caterina Froio, Stijn van Kessel, Sarah L. de Lange, Cas Mudde, and Paul Taggart (2023). The PopuList: A Database of Populist, Far-Left, and Far-Right Parties Using Expert-Informed Qualitative Comparative Classification (EiQCC). *British Journal of Political Science*, 1–10. https://doi.org/10.1017/S0007123423000431.

Rottweiler, B. & Gill, P. (2021). Measuring Individuals' Misogynistic Attitudes: Development and Validation of the Misogyny Scale. (Preprint) https://psyarxiv.com/6f829.

Samson, A. C., & Gross, J. J. (2012). Humour as emotion regulation: The differential consequences of negative versus positive humour. *Cognition & emotion*, *26*(2), 375-384. https://doi.org/10.1080/02699931.2011.585069

Sargent, R. H., & Newman, L. S. (2021). Pluralistic Ignorance Research in Psychology: A Scoping Review of Topic and Method Variation and Directions for Future Research. *Review of General Psychology*, *25*, 163-184. https://doi.org/10.1177/1089268021995168

Schuster, C., Sparkman, G., Walton, G. M., Alles, A., & Loschelder, D. D. (2023). Egalitarian norm messaging increases human resources professionals' salary offers to women. *Journal of Applied Psychology, 108*(4), 541–552. https://doi.org/10.1037/apl0001033

Selvanathan, H. P., Techakesari, P., Tropp, L. R., & Barlow, F. K. (2017). Whites for racial justice: How contact with Black Americans predicts support for collective action among White Americans. *Group Processes & Intergroup Relations, 21*(6), 893–912. https://doi.org/10.1177/1368430217690908

Sinclair, V. M., & Saklofske, D. H. (2019). Is there a place in politics for compassion? The role of compassion in predicting hierarchy-legitimizing views. *Self and Identity, 18*(4), 425–442. https://doi.org/10.1080/15298868.2018.1468354

Simonovits, G., Kézdi, G., & Kardos, P. (2018). Seeing the world through the other's eye: An online intervention reducing ethnic prejudice. *American Political Science Review, 112*(1), 186–193. https://doi.org/10.1017/S0003055417000478

Skewes, L., Fine, C., & Haslam, N. (2018). Beyond Mars and Venus: The role of gender essentialism in support for gender inequality and backlash. *PLOS ONE, 13*(7), e0200921. https://doi.org/10.1371/journal.pone.0200921

Sharpe, S., & Hynes, M. (2016). Black-faced, red faces: The potentials of humour for anti-racist action. *Ethnic and Racial Studies, 39*(1), 87–104. https://doi.org/10.1080/01419870.2016.1096405

Strehovec, T. (2013). Zakonska zveza v perspektivi drugega vatikanskega koncila. *Bogoslovni vestnik*, *73*(2), 233-240.

Suls, J., & Wan, C. K. (1987). In search of the false-uniqueness phenomenon: Fear and estimates of social consensus. *Journal of Personality and Social Psychology, 52*(1), 211–217. https://doi.org/10.1037/0022-3514.52.1.211

Swim, J. K., Aikin, K. J., Hall, W. S., & Hunter, B. A. (1995). Sexism and racism: Old-fashioned and modern prejudices. *Journal of personality and social psychology*, *68*(2), 199-214. https://doi.org/10.1037/0022-3514.68.2.199

Szekeres, H., Lantos, N. A., Faragó, L., Nyúl, B., & Kende, A. (2024). When the shoe does not fit: The role of perspective-taking orientation in a perspective-taking prejudice reduction intervention. *Journal of Community & Applied Social Psychology, 34*(2), e2785. https://doi.org/10.1002/casp.2785

Szóstakowski, D., & Besta, T. (2024). Perceived threat, injustice appraisal and willingness to join progressive collective action. *Current Psychology: A Journal for Diverse Perspectives on*

Diverse Psychological Issues, 43(8), 7399–7410. https://doi.org/10.1007/s12144-023-04926-1

Tankard, M. E., & Paluck, E. L. (2016). Norm perception as a vehicle for social change. *Social issues and policy review*, *10*(1), 181-211. https://doi.org/10.1111/sipr.12022

Taylor, M. G., Rhodes, M., & Gelman, S. A. (2009). Boys will be boys; Cows will be cows: Children's essentialist reasoning about gender categories and animal species. *Child Development, 80*(2), 461–481. https://doi.org/10.1111/j.1467-8624.2009.01272.x

Thomas, E. F., & McGarty, C. (2018). *Giving versus acting: Using latent profile analysis to distinguish between benevolent and activist support for global poverty reduction*. *British Journal of Social Psychology, 57*(1), 189–209. https://doi.org/10.1111/bjso.12228

Tirion, A. S. C., Mulder, L. B., Kurz, T., Koudenburg, N., Prosser, A. M. B., Bain, P. G., & Bolderdijk, J. W. (2024). The sound of silence: The importance of bystander support for confronters in the prevention of norm erosion. *British Journal of Social Psychology, 63*(2), 909–935. https://doi.org/10.1111/bjso.12709

Turner, R. N., & Cameron, L. (2016). Confidence in contact: A new perspective on promoting cross-group friendship among children and adolescents. *Social Issues and Policy Review, 10*(1), 212–246. https://doi.org/10.1111/sipr.12023

Tyler, T. R., & Smith, H. J. (1998). Social justice and social movements. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The Handbook of Social Psychology* (4th ed., Vol. 2, pp. 595–629. New York: McGraw-Hill.

United Nations Development Programme. (2025). *Gender Inequality Index (GII)*. Human Development Reports Data Center. https://hdr.undp.org/data-center/thematic-composite-indices/gender-inequality-index

Van der Graaff, J., Branje, S., De Wied, M., Hawk, S., Van Lier, P., & Meeus, W. (2014). Perspective taking and empathic concern in adolescence: Gender differences in developmental changes. *Developmental Psychology, 50*(3), 881–888. https://doi.org/10.1037/a0034325

Villa, P.-I., 2017. "Anti-genderismus": German angst?. In: R. Kuhar & D. Paternotte (Eds.), *Anti-Gender Campaigns in Europe Mobilizing against Equality.* Rowman and Littlefield International Ltd, pp.99-116.

Vizcaíno-Cuenca, R., Riquelme, A. R., Romero-Sánchez, M., Megías, J. L., & Carretero-Dios, H. (2024). Exposure to feminist humor and the proclivity to collective action for gender equality: The role of message format and feminist identification. *Sex Roles: A Journal of Research, 90*(1), 186–201. https://doi.org/10.1007/s11199-023-01430-5

Wardle, C., & Derakhshan, H. (2023). *Information Disorder: Toward an interdisciplinary framework for research and policymaking.* Council of Europe, Strasbourg.

Walter, N., & Tukachinsky, R. (2020). A meta-analytic examination of the continued influence of misinformation in the face of correction: How powerful is it, why does it happen, and how to stop it? *Communication Research, 47*(2), 155–177. https://doi.org/10.1177/0093650219854600

Wedell, E., & Chaney, K. E. (2025). Flexible confronters, informative confronters, and low stakes prodders: A person-centered approach to prejudice confrontation styles. *Personality and Individual Differences*, *233*, 112904. https://doi.org/10.1016/j.paid.2024.112904

Wenzel, M., & Woodyatt, L. (2024). The Power and Pitfalls of Social Norms. *Annual Review of Psychology*, 76, 583–606. https://doi.org/10.1146/annurev-psych-020124-120310

Zhou, Y., & Shen, L. (2024). Processing of misinformation as motivational and cognitive biases. *Frontiers in Psychology, 15,* 1430953. https://doi.org/10.3389/fpsyg.2024.1430953

Appendix A: Measures

- Misogyny (Rottweiler & Gill, 2021)

Response scale: 1 = Strongly disagree … 5 = Strongly agree
"Women seek to gain power by getting control over men."
"Women use their sexuality to manipulate men."
"Women exploit men for their own agendas."
"If things don't go their way, women will play the victim."
"It is generally safer not to trust women too much."
"When it comes down to it a lot of women are deceitful."
"Most women would lie just to get ahead."
"Sometimes women bother me by just being around."
"I feel uncomfortable when a woman dominates the conversation."

- Threat perceptions (Zero-sum thinking, Kuchynka et al., 2018)

Response scale: 1 = Strongly disagree … 5 = Strongly agree
"As women have gained status in the workplace, men have lost status in the workplace."
"Women in the workforce take jobs away from men."
"The more money women earn in the workplace, the less money men earn."

- Gender stereotypes (Special Eurobarometer on Gender Equality, 2017)

Response scale: 1 = Strongly disagree … 5 = Strongly agree
"Women do not have the necessary qualities and skills to fill positions of responsibility in politics."

- Modern sexism (one item from Swim et al., 1995)

Response scale: 1 = Strongly disagree … 5 = Strongly agree
"Discrimination against women is no longer a problem in COUNTRY."

- Willingness to get involved

Response scale: 1 = Very unlikely … 5 = Very likely

"Imagine that the [COUNTRY] parliament is discussing gender quotas. Some parties propose that all party lists should include equal numbers of women and men candidates. Other parties are against this proposal. The proposal could pass if citizens signal their preferences to the parliament. How likely is it that you would join an online petition *in support* of the gender quota?"

- Perspective taking (Davis, 1980)

Response scale: 1 = Strongly disagree … 5 = Strongly agree

"I sometimes find it difficult to see things from the point of view of others."
"I try to look at everybody's side of a disagreement before I make a decision."
"I sometimes try to understand my friends better by imagining how things look from their perspective."
"If I'm sure I'm right about something, I don't waste much time listening to other people's arguments."
"When I'm upset at someone, I usually try to put myself in their shoes for a while."
"Before criticizing somebody, I try to imagine how I would feel if I were in their place."

- Gender system justification (Jost and Kay, 2005)

Response scale: 1 = Strongly disagree … 5 = Strongly agree

"In general, relations between men and women are fair."
"The division of labor in families generally operates as it should."
"Gender roles need to be radically restructured."
"Sexism in society is getting worse every year."
"Society is set up so that men and women usually get what they deserve."
"Discrimination against women is no longer a problem in COUNTRY."

- Attitudes towards gender equality (reworded Pratto et al., 2013)

Response scale: 1 = Strongly disagree … 5 = Strongly agree

"In setting priorities, we must consider men and women alike."
"We should not push for equality between men and women."
"The equality between men and women should be our goal."
"Men should dominate over women."

Appendix B: Additional Results

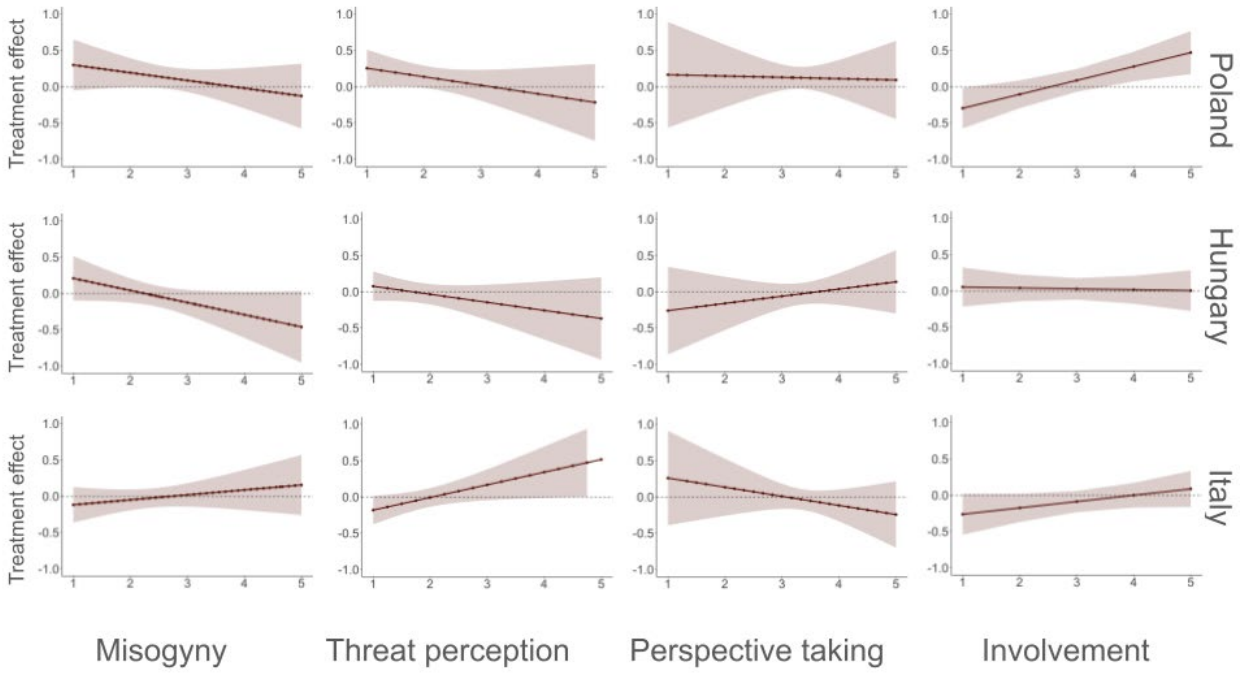


Figure A 1. Marginal effect plots of treatment on agreement with statement "Discrimination against women is no longer a problem in a country"

The figure displays the marginal effect of receiving counter-arguments (treatment track) relative to viewing unchallenged anti-gender content (control track) on Wave-2 scores of the reverse-coded modern sexism item "Discrimination against women is no longer a problem in [COUNTRY]," where higher values indicate greater recognition of discrimination. Each plot shows how the estimated treatment effect varies across baseline levels of four Wave-1 moderators: misogyny, threat perception, perspective taking and willingness to get involved. The solid line represents the marginal treatment effect at each moderator value, while the shaded band indicates the 95 percent confidence interval. The top row presents results for Poland, the middle row for Hungary and the bottom row for Italy. Across all three countries, the interaction analyses reveal no statistically significant treatment heterogeneity for misogyny, threat perception or perspective taking. In Hungary and Italy, the marginal effects remain close to zero across all moderator values. In Poland, however, the marginal effect becomes positive and statistically significant at higher levels of involvement, indicating that for respondents who were more willing to engage with gender-equality causes at baseline, exposure to counter-arguments increases recognition of discrimination relative to the control condition.

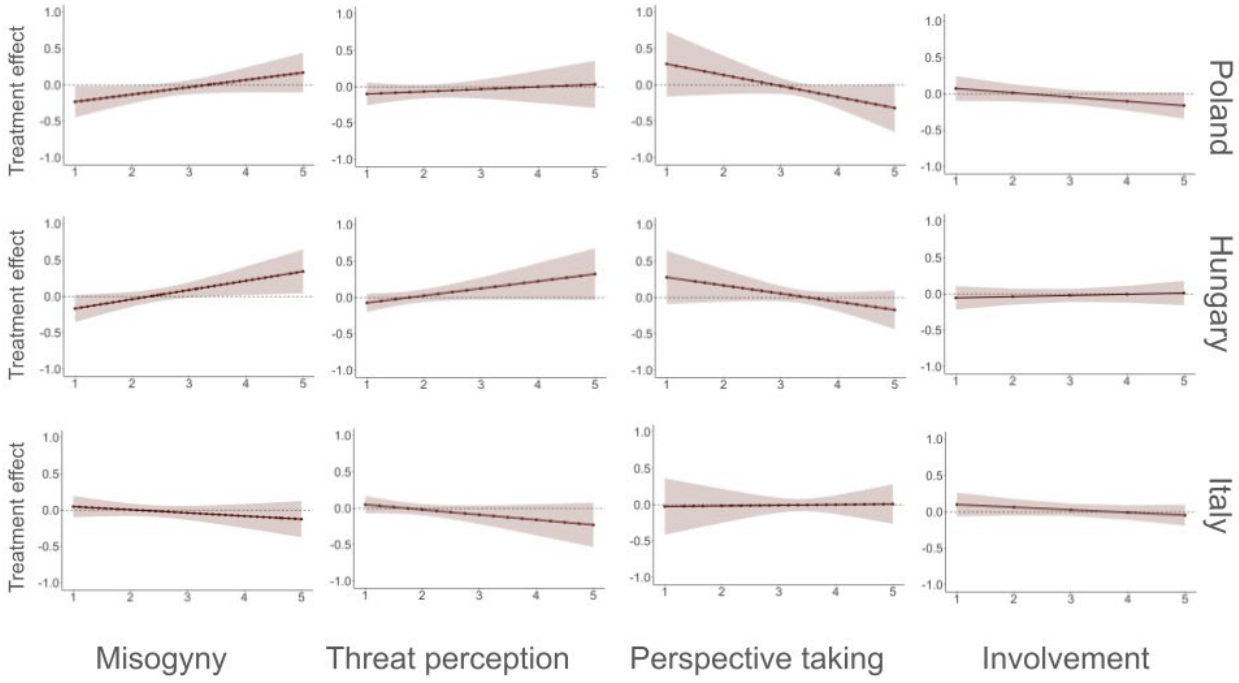


Figure A 2. Marginal effect plots of treatment on gender system justification battery as outcome

The figure displays the marginal effect of receiving counter-arguments (treatment track) relative to viewing unchallenged anti-gender content (control track) on Wave-2 scores of the gender system justification battery. The outcome is the average score across six items capturing perceived fairness and legitimacy of gender relations, with higher values indicating stronger system justification. Each plot shows how the estimated treatment effect varies across baseline values of four Wave-1 moderators: misogyny, threat perception, perspective taking and willingness to get involved. The solid line represents the marginal treatment effect at each level of the moderator, while the shaded band indicates the 95 percent confidence interval. The top row shows results for Poland, the middle row for Hungary and the bottom row for Italy. Across countries, the interaction analyses reveal no statistically significant heterogeneity in treatment effects for any of the moderators. In Poland, the marginal effect becomes slightly more positive at high levels of involvement, but confidence intervals overlap zero throughout, indicating that this apparent pattern is not statistically meaningful. Overall, the results do not provide evidence that treatment effects on gender system justification vary systematically by respondents' prior levels of misogyny, threat perception, perspective taking or involvement.

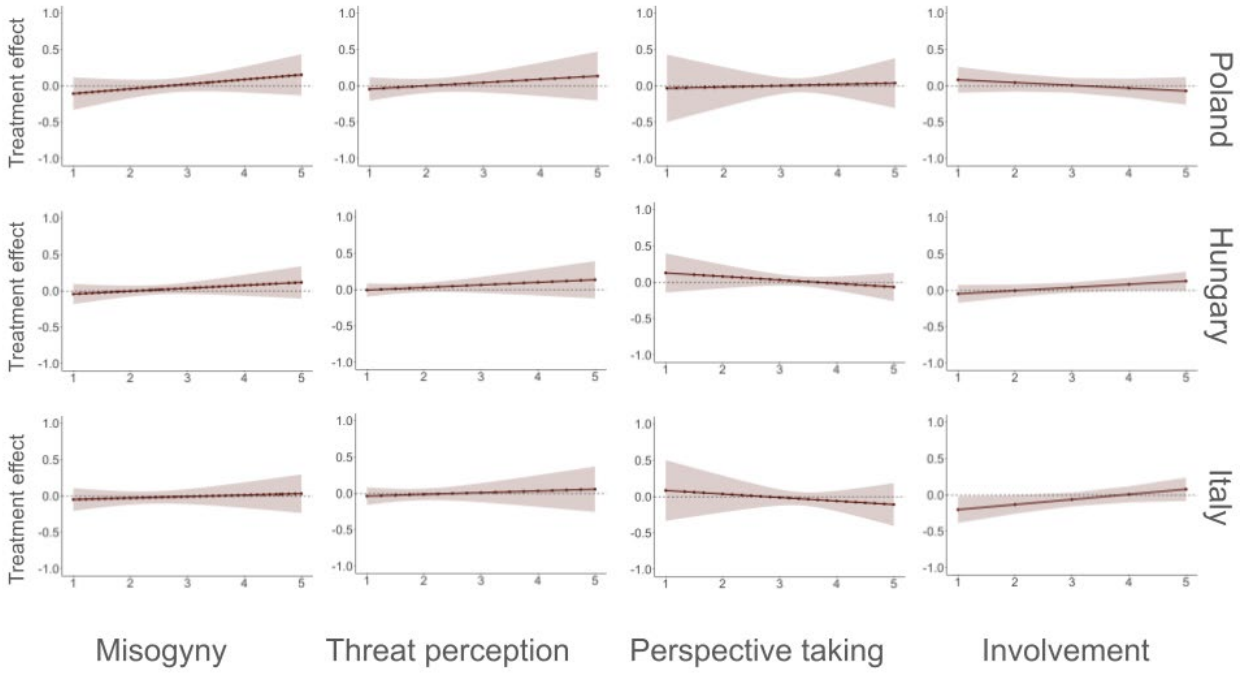


Figure A 3. Marginal effects plots of treatment on attitude towards gender equality battery

The figure displays the marginal effect of receiving counter-arguments (treatment track) relative to viewing unchallenged anti-gender content (control track) on Wave-2 scores of the four-item gender equality battery. Higher values on this battery indicate stronger support for gender equality. Each subplot shows how the estimated treatment effect varies across baseline levels of four Wave-1 moderators: misogyny, threat perception, perspective taking and willingness to get involved. The solid line represents the marginal treatment effect at each moderator value, the shaded band indicates the 95 percent confidence interval and the horizontal zero line marks the threshold indicating no effect. The top row shows results for Poland, the middle row for Hungary and the bottom row for Italy. Across all three countries and all moderators, the marginal effects remain close to zero and confidence intervals consistently overlap zero, indicating no evidence that the impact of treatment varies by respondents' prior attitudes. In Poland, the treatment effect becomes slightly more positive at higher levels of involvement, but confidence intervals still overlap zero throughout, so this pattern is not statistically meaningful. Overall, the results suggest that the treatment does not systematically influence gender-equality attitudes for any subgroup defined by these baseline characteristics.

Table A 1. OLS estimates for the outcome variable "Discrimination against women is no longer a problem in a country" - Hungary

	Dependent variable:					
	W2 Measure				W2 - W1	
Treatment	-0.021	-0.010	-0.015	0.060	-0.004	-0.003
	(0.076)	(0.061)	(0.061)	(0.187)	(0.070)	(0.070)
W1 Measure		0.578***	0.550***	0.595***		
		(0.019)	(0.020)	(0.047)		
Gender: Male			-0.270***			-0.021
			(0.048)			(0.054)
University Education			0.092			-0.008
			(0.050)			(0.057)
Age			-0.002			-0.001
			(0.002)			(0.002)
Treatment * W1 Measure				-0.021		
				(0.051)		
Constant	3.477***	1.491***	1.761***	1.433***	0.033	0.115
	(0.069)	(0.087)	(0.121)	(0.170)	(0.063)	(0.112)
Observations	1,574	1,542	1,542	1,542	1,542	1,542
Adjusted R^2	-0.001	0.363	0.376	0.363	-0.001	-0.002
Note:				*p<0.05; **p<0.01; ***p<0.001		

The table reports OLS regression estimates for Hungary using the reverse-coded modern sexism item as the outcome, where higher values indicate greater recognition of discrimination. The key

independent variable, Treatment, indicates assignment to the treatment track in Wave-2, where respondents viewed anti-gender posts accompanied by counter-arguments. The control group viewed the same posts without counter-arguments. The Wave-1 measure refers to respondents' baseline score on the same modern sexism item, measured several weeks earlier, and captures their initial level of discrimination recognition prior to exposure to the Wave-2 experimental content. Columns 1–3 model Wave-2 levels of the outcome, while Columns 4–6 examine changes from Wave-1 to Wave-2 by using the Wave-2 minus Wave-1 difference as the dependent variable. Across all specifications, the treatment indicator is small and not statistically significant, indicating that exposure to counter-arguments did not produce detectable changes in recognition of discrimination for Hungarian respondents. In contrast, the Wave-1 measure is a strong and highly significant predictor of Wave-2 scores wherever included, reflecting substantial temporal stability in the outcome.

Table A 2. OLS estimates for the outcome variable "Discrimination against women is no longer a problem in a country" - Poland

	Dependent variable:					
	W2 Measure				W2 - W1	
Treatment	0.119	0.130	0.077	-0.227	0.228*	0.171
	(0.077)	(0.078)	(0.074)	(0.221)	(0.111)	(0.108)
W1 Measure		-0.001	0.004	-0.100		
		(0.025)	(0.024)	(0.063)		
Gender: Male			-0.643***			-0.700***
			(0.055)			(0.080)
University Education			0.061			0.119
			(0.056)			(0.081)
Age			-0.009***			-0.007**
			(0.002)			(0.003)
Treatment * W1 Measure				0.118		
				(0.068)		
Constant	2.865***	2.861***	3.590***	3.161***	-0.187	0.489**
	(0.070)	(0.104)	(0.137)	(0.203)	(0.101)	(0.168)
Observations	1,587	1,546	1,546	1,546	1,546	1,546
Adjusted R^2	0.001	0.001	0.100	0.002	0.002	0.056
Note:						*p<0.05; **p<0.01; ***p<0.001

The table reports OLS regression estimates for Poland using the reverse-coded modern sexism item as the outcome, where higher values indicate greater recognition of discrimination. The key independent variable, Treatment, indicates assignment to the Wave-2 treatment track, in which respondents viewed anti-gender posts accompanied by counter-arguments; the control group viewed the same posts without counter-arguments. The Wave-1 measure represents respondents' baseline score on the same modern sexism item, capturing their initial level of recognition of discrimination before exposure to the Wave-2 experimental content. Columns 1–3 model Wave-2 levels of the outcome, and Columns 4–6 examine changes from Wave-1 to Wave-2 by using the Wave-2 minus Wave-1 difference as the dependent variable.

Table A 3. OLS estimates for the outcome variable "Discrimination against women is no longer a problem in a country"– Italy

	Dependent variable:					
	W2 Measure				W2 - W1	
Treatment	-0.041	-0.050	-0.053	0.010	-0.064	-0.070
	(0.072)	(0.062)	(0.061)	(0.220)	(0.071)	(0.071)
W1 Measure		0.528***	0.493***	0.542***		
		(0.021)	(0.021)	(0.052)		
Gender: Male			-0.270***			-0.0005
			(0.047)			(0.053)
University Education			0.020			-0.062
			(0.057)			(0.066)
Age			-0.005**			-0.003*
			(0.001)			(0.002)
Treatment * W1 Measure				-0.017		
				(0.057)		
Constant	3.717***	1.774***	2.280***	1.723***	0.038	0.234*
	(0.066)	(0.096)	(0.132)	(0.201)	(0.065)	(0.116)
Observations	1,621	1,584	1,584	1,584	1,584	1,584
Adjusted R^2	-0.0004	0.286	0.305	0.285	-0.0001	0.001
Note:						*p<0.05; **p<0.01; ***p<0.001

The table reports OLS regression estimates for Italy using the reverse-coded modern sexism item as the outcome, where higher values indicate greater recognition of discrimination. The key independent variable, Treatment, indicates assignment to the Wave-2 treatment track, in which respondents viewed anti-gender posts accompanied by counter-arguments; the control group viewed the same posts without counter-arguments. The Wave-1 measure represents respondents' baseline score on the same modern sexism item, capturing their initial level of recognition of discrimination before exposure to the Wave-2 experimental content. Columns 1–3 model Wave-2 levels of the outcome, and Columns 4–6 examine changes from Wave-1 to Wave-2 by using the Wave-2 minus Wave-1 difference as the dependent variable.

Table A 4. OLS estimates for the outcome variable "Discrimination against women is no longer a problem in a country" - Hungary

	Dependent variable:			
	Wave-2 Measure			
Treatment	0.374	0.187	-0.360	0.064
	(0.244)	(0.179)	(0.433)	(0.189)
Misogyny	-0.222*			
	(0.086)			
Treatment * Misogyny	-0.167			
	(0.095)			
Threat Perception		-0.455***		
		(0.081)		
Treatment * Threat Perception		-0.111		
		(0.089)		
Perspective			0.060	
			(0.115)	
Treatment * Perspective			0.099	
			(0.127)	
Involvement				0.279***
				(0.053)
Treatment * Involvement				-0.012
				(0.059)
Constant	4.024***	4.318***	3.276***	2.644***
	(0.223)	(0.163)	(0.389)	(0.172)
Observations	1,574	1,574	1,574	1,452
Adjusted R^2	0.063	0.139	0.004	0.086
Note:		*p<0.05; **p<0.01; ***p<0.001		

The table reports OLS regression estimates for Hungary using the reverse-coded modern sexism item as the outcome, where higher values indicate greater recognition of discrimination. The key independent variable, Treatment, indicates assignment to the Wave-2 treatment track, in which respondents viewed anti-gender posts accompanied by counter-arguments; the control group viewed the same posts without counter-arguments. The models examine whether the effect of treatment varies by baseline attitudes measured at Wave-1. Each specification interacts Treatment with one of four moderators collected in Wave-1: misogyny, threat perception, perspective taking and willingness to get involved. The Wave-2 modern-sexism score serves as

the dependent variable in all specifications. Across the four models, treatment coefficients are small and not statistically significant, indicating no detectable average effect of counter-arguments on recognition of discrimination in Hungary.

Table A 5. OLS estimates for the outcome variable "Discrimination against women is no longer a problem in a country" - Poland

	Dependent variable:			
	Wave-2 Measure			
Treatment	0.404	0.370	0.181	-0.490*
	(0.268)	(0.211)	(0.527)	(0.196)
Misogyny	0.141			
	(0.086)			
Treatment * Misogyny	-0.107			
	(0.094)			
Threat Perception		0.143		
		(0.083)		
Treatment * Threat Perception		-0.117		
		(0.091)		
Perspective			-0.043	
			(0.145)	
Treatment * Perspective			-0.018	
			(0.157)	
Involvement				-0.152**
				(0.056)
Treatment * Involvement				0.192**
				(0.061)
Constant	2.485***	2.558***	3.007***	3.350***
	(0.244)	(0.192)	(0.485)	(0.180)
Observations	1,587	1,587	1,587	1,432
Adjusted R^2	0.002	0.002	0.0003	0.005
Note:				*p<0.05; **p<0.01; ***p<0.001

The table reports OLS regression estimates for Poland using the reverse-coded modern sexism item as the outcome, where higher values indicate greater recognition of discrimination. The key independent variable, Treatment, indicates assignment to the Wave-2 treatment track, in which respondents viewed anti-gender posts accompanied by counter-arguments; the control group viewed the same posts without counter-arguments. The models examine whether the effect of treatment varies by baseline attitudes measured at Wave-1. Each specification interacts Treatment with one of four moderators collected in Wave-1: misogyny, threat perception,

perspective taking and willingness to get involved. The Wave-2 modern-sexism score serves as the dependent variable in all specifications. Across the four models, treatment coefficients are generally positive but imprecisely estimated, with the exception of the involvement model, where treatment becomes statistically significant when interacted with baseline involvement. As in the main analyses, the results provide only limited evidence of heterogeneity, with no consistent pattern indicating that the effect of counter-arguments varies systematically by respondents' prior attitudes in Poland.

Table A 6. OLS estimates for the outcome variable "Discrimination against women is no longer a problem in a country" - Italy

	Dependent variable:			
	Wave-2 Measure			
Treatment	-0.188	-0.357*	0.385	-0.351
	(0.194)	(0.170)	(0.466)	(0.195)
Misogyny	-0.507***			
	(0.070)			
Treatment * Misogyny	0.068			
	(0.077)			
Threat Perception		-0.720***		
		(0.074)		
Treatment * Threat Perception		0.174*		
		(0.081)		
Perspective			0.266*	
			(0.125)	
Treatment * Perspective			-0.126	
			(0.137)	
Involvement				0.134*
				(0.052)
Treatment * Involvement				0.087
				(0.057)
Constant	4.919***	5.104***	2.818***	3.292***
	(0.177)	(0.155)	(0.428)	(0.177)
Observations	1,621	1,621	1,621	1,518
Adjusted R^2	0.131	0.183	0.005	0.061
Note:				*p<0.05; **p<0.01; ***p<0.001

The table reports OLS regression estimates for Italy using the reverse-coded modern sexism item as the outcome, where higher values indicate greater recognition of discrimination. The key independent variable, Treatment, indicates assignment to the Wave-2 treatment track, in which respondents viewed anti-gender posts accompanied by counter-arguments; the control group viewed the same posts without counter-arguments. The models examine whether the effect of treatment varies by baseline attitudes measured at Wave-1. Each specification interacts Treatment with one of four moderators collected in Wave-1: misogyny, threat perception,

perspective taking and willingness to get involved. The Wave-2 modern-sexism score serves as the dependent variable in all specifications.

Table A 7. OLS for gender system justification battery as the outcome - Hungary

	Dependent variable:			
	System Justification			
Treatment	-0.295*	-0.174	0.388	-0.072
	(0.150)	(0.112)	(0.264)	(0.113)
Misogyny	0.076			
	(0.053)			
Treatment * Misogyny	0.128*			
	(0.058)			
Threat Perception		0.192***		
		(0.050)		
Treatment * Threat Perception		0.099		
		(0.056)		
Perspective			0.027	
			(0.070)	
Treatment * Perspective			-0.112	
			(0.077)	
Involvement				-0.201***
				(0.032)
Treatment * Involvement				0.016
				(0.035)
Constant	2.599***	2.431***	2.696***	3.388***
	(0.137)	(0.101)	(0.237)	(0.103)
Observations	1,574	1,574	1,574	1,452
Adjusted R^2	0.044	0.094	0.002	0.113
Note:				*p<0.05; **p<0.01; ***p<0.001

The table reports OLS regression estimates for Hungary using the gender system justification battery as the outcome, where higher scores indicate stronger endorsement of existing gender relations as fair and legitimate. The key independent variable, Treatment, indicates assignment to the Wave-2 treatment track, in which respondents viewed anti-gender posts accompanied by counter-arguments; the control group viewed the same posts without counter-arguments. The models examine whether the effect of treatment varies by baseline attitudes measured at Wave-1. Each specification interacts Treatment with one of four moderators collected in Wave-1: misogyny, threat perception, perspective taking and willingness to get involved. The Wave-2 system justification score serves as the dependent variable in all specifications.

Table A 8. OLS for gender system justification battery as the outcome - Poland

	Dependent variable:			
	System Justification			
Treatment	-0.334*	-0.133	0.436	0.130
	(0.164)	(0.129)	(0.323)	(0.120)
Misogyny	-0.108*			
	(0.053)			
Treatment * Misogyny	0.100			
	(0.058)			
Threat Perception		-0.057		
		(0.051)		
Treatment * Threat Perception		0.033		
		(0.056)		
Perspective			0.164	
			(0.089)	
Treatment * Perspective			-0.151	
			(0.097)	
Involvement				0.032
				(0.034)
Treatment * Involvement				-0.059
				(0.038)
Constant	3.392***	3.224***	2.558***	2.986***
	(0.150)	(0.118)	(0.298)	(0.110)
Observations	1,588	1,588	1,588	1,433
Adjusted R^2	0.002	0.001	0.002	0.001
Note:				*p<0.05; **p<0.01; ***p<0.001

The table reports OLS regression estimates for Poland using the gender system justification battery as the outcome, where higher scores indicate stronger endorsement of existing gender relations as fair and legitimate. The key independent variable, Treatment, indicates assignment to the Wave-2 treatment track, in which respondents viewed anti-gender posts accompanied by counter-arguments; the control group viewed the same posts without counter-arguments. The models examine whether the effect of treatment varies by baseline attitudes measured at Wave-1. Each specification interacts Treatment with one of four moderators collected in Wave-1: misogyny, threat perception, perspective taking and willingness to get involved. The Wave-2 system justification score serves as the dependent variable in all specifications.

Table A 9. Pooled OLS for gender system justification battery as the outcome - Italy

	Dependent variable:			
	System Justification			
Treatment	0.090	0.118	-0.034	0.136
	(0.117)	(0.103)	(0.278)	(0.114)
Misogyny	0.277***			
	(0.042)			
Treatment * Misogyny	-0.043			
	(0.046)			
Threat Perception		0.375***		
		(0.045)		
Treatment * Threat Perception		-0.070		
		(0.049)		
Perspective			-0.083	
			(0.075)	
Treatment * Perspective			0.009	
			(0.081)	
Involvement				-0.116***
				(0.030)
Treatment * Involvement				-0.036
				(0.033)
Constant	2.022***	1.959***	2.961***	3.043***
	(0.106)	(0.094)	(0.255)	(0.103)
Observations	1,621	1,621	1,621	1,518
Adjusted R^2	0.106	0.156	0.002	0.086
Note:		*p<0.05; **p<0.01; ***p<0.001		

The table reports OLS regression estimates for Italy using the gender system justification battery as the outcome, where higher scores indicate stronger endorsement of existing gender relations as fair and legitimate. The key independent variable, Treatment, indicates assignment to the Wave-2 treatment track, in which respondents viewed anti-gender posts accompanied by counter-arguments; the control group viewed the same posts without counter-arguments. The models examine whether the effect of treatment varies by baseline attitudes measured at Wave-1. Each specification interacts Treatment with one of four moderators collected at Wave-1: misogyny, threat perception, perspective taking and willingness to get involved. The Wave-2 system justification score serves as the dependent variable in all specifications.

Table A 10. Pooled OLS for the attitude towards gender equality as the outcome - Hungary

	Dependent variable:			
	Gender Equality			
Treatment	-0.084	-0.042	0.176	-0.093
	(0.110)	(0.081)	(0.193)	(0.087)
Misogyny	-0.200***			
	(0.039)			
Treatment * Misogyny	0.040			
	(0.043)			
Threat Perception		-0.279***		
		(0.037)		
Treatment * Threat Perception		0.035		
		(0.040)		
Perspective			0.212***	
			(0.051)	
Treatment * Perspective			-0.048	
			(0.057)	
Involvement				0.006
				(0.024)
Treatment * Involvement				0.044
				(0.027)
Constant	4.316***	4.340***	3.115***	3.809***
	(0.101)	(0.074)	(0.174)	(0.079)
Observations	1,573	1,573	1,573	1,451
Adjusted R^2	0.064	0.141	0.036	0.011
Note:				*p<0.05; **p<0.01; ***p<0.001

The table reports OLS regression estimates for Hungary using the gender equality battery as the outcome, where higher scores indicate stronger support for gender equality. The key independent variable, Treatment, indicates assignment to the Wave-2 treatment track, in which respondents viewed anti-gender posts accompanied by counter-arguments; the control group viewed the same posts without counter-arguments. The models examine whether the effect of treatment varies by baseline attitudes measured at Wave-1. Each specification interacts Treatment with one of four moderators collected in Wave-1: misogyny, threat perception, perspective taking and willingness to get involved. The Wave-2 gender-equality score serves as the dependent variable in all specifications.

Table A 11. OLS, attitude towards gender equality as the outcome - Poland

	Dependent variable:			
	Gender Equality			
Treatment	-0.174	-0.092	-0.054	0.119
	(0.170)	(0.134)	(0.336)	(0.125)
Misogyny	-0.017			
	(0.055)			
Treatment * Misogyny	0.065			
	(0.060)			
Threat Perception		-0.011		
		(0.053)		
Treatment * Threat Perception		0.045		
		(0.058)		
Perspective			0.038	
			(0.092)	
Treatment * Perspective			0.018	
			(0.100)	
Involvement				0.039
				(0.036)
Treatment * Involvement				-0.038
				(0.039)
Constant	3.993***	3.970***	3.822***	3.828***
	(0.155)	(0.122)	(0.309)	(0.115)
Observations	1,588	1,588	1,588	1,433
Adjusted R^2	0.001	-0.001	-0.0005	-0.001
Note:				*p<0.05; **p<0.01; ***p<0.001

The table reports OLS regression estimates for Poland using the gender equality battery as the outcome, where higher scores indicate stronger support for gender equality. The key independent variable, Treatment, indicates assignment to the Wave-2 treatment track, in which respondents viewed anti-gender posts accompanied by counter-arguments; the control group viewed the same posts without counter-arguments. The models examine whether the effect of treatment varies by baseline attitudes measured at Wave-1. Each specification interacts Treatment with one of four moderators collected in Wave-1: misogyny, threat perception, perspective taking and willingness to get involved. The Wave-2 gender-equality score serves as the dependent variable in all specifications.

Table A 12. OLS, attitude towards gender equality as outcome - Italy

	Dependent variable:			
	Gender Equality Variable			
Treatment	-0.071	-0.063	0.133	-0.274*
	(0.124)	(0.105)	(0.300)	(0.126)
Misogyny	-0.351***			
	(0.045)			
Treatment * Misogyny	0.020			
	(0.049)			
Threat Perception		-0.480***		
		(0.046)		
Treatment * Threat Perception		0.024		
		(0.050)		
Perspective			0.302***	
			(0.081)	
Treatment * Perspective			-0.048	
			(0.088)	
Involvement				0.084*
				(0.034)
Treatment * Involvement				0.070
				(0.037)
Constant	5.072***	5.163***	3.219***	3.975***
	(0.113)	(0.096)	(0.275)	(0.115)
Observations	1,621	1,621	1,621	1,518
Adjusted R^2	0.170	0.272	0.038	0.069
Note:				*p<0.05; **p<0.01; ***p<0.001

The table reports OLS regression estimates for Italy using the gender equality battery as the outcome, where higher scores indicate stronger support for gender equality. The key independent variable, Treatment, indicates assignment to the Wave-2 treatment track, in which respondents viewed anti-gender posts accompanied by counter-arguments; the control group viewed the same posts without counter-arguments. The models examine whether the effect of treatment varies by baseline attitudes measured at Wave-1. Each specification interacts Treatment with one of four moderators collected in Wave-1: misogyny, threat perception, perspective taking and willingness to get involved. The Wave-2 gender-equality score serves as the dependent variable in all specifications.

Appendix C: Survey Questionnaire

Section 1

1.2.
In a typical week, how often do you visit or use the following sources:

Facebook
Twitter (X)
Instagram
YouTube
TikTok

1= Not at all
2= Less than once a week or once in a while
3= A few days per week (e.g., 2–3 times)
4= Most days (e.g., 4–6 times per week)
5= Once per day
6= a few times per day
7= More than 5 times a day

1.7. You will now read statements that one might come across on social media. To what extent do you agree or disagree with each of the following statements?

Response options:
1 = Strongly disagree
2 = Disagree
3 = Somewhat disagree
4 = Somewhat agree
5 = Agree
6 = Strongly agree

1. Society should remind girls that having children is more important than pursuing a career.
2. Abortion should be more strictly regulated, as it involves making a decision about another life.
3. Tourism harms local communities and therefore needs stricter regulation.
4. The traditional family model must be protected from the influence of the international gender lobby.
5. Fast fashion exploits workers and harms the planet.
6. Policies aimed at promoting gender equality can unfairly disadvantage men.
7. EU member states should not be forced to adopt liberal gender norms.

Section 2
2. Secondary schools could engage with the topic of social media to help students navigate in this complex world. In the following section, we would like to learn your opinion about various types of content. You will see social media posts followed by related questions. We are interested in what you think about the contents themselves, and whether you believe these contents would be suitable for discussion among final-year secondary school students.

2.1 Filler 1: Tourism

Post:
Post: “Tourists, go home!” — This graffiti can be seen all over downtown Barcelona, scrawled by angry locals who feel that excessive tourism has made the city unlivable. It’s not just about the trash left behind by millions of visitors or the suffocating crowds: the multiplication of hotels and short-term rentals has made life unaffordable for residents.

Comment:
Comment: Unaffordable, really? Sure. But let’s be honest — if the tourists actually stopped coming, we’d just end up complaining that the last café in the neighborhood had to shut down because there weren’t enough “local customers.” It’s kind of like when you can’t stand your relatives — but still miss them when they don’t show up for Christmas.

Please place yourself on this scale.

- I completely disagree with the **post**– I completely agree with the **post**
  Slider: 1–6
- I completely disagree with the **comment** – I completely agree with the **comment**
  Slider: 1–6
- Schools **should not address** this topic at all – Schools **definitely should address** this topic
  Slider: 1–6

The next few questions refer only to the post at the top — not the comment below the post.

To what extent do you agree with the following statements about the original post?

1 = Strongly disagree
2 = Disagree
3 = Somewhat disagree

4 = Somewhat agree
5 = Agree
6 = Strongly agree

1. If I saw a post like this, I would click the “like” button on social media.
2. The post is well-intentioned.
3. The post helps address social problems.
4. The post creates division between groups.

When thinking about people living in [Country’s name], what percentage of them do you think agree with the original post?

Slider: 0–100%
0% = No one agrees
100% = Everyone agrees

2.2. Gender_roles

Post:

We must remind our daughters that their primary purpose is to become mothers, because there is a risk they may forget the importance of having children in pursuit of professional advancement.

Comments:

e) Comment: Every woman has the right to choose her own path in life — whether that means pursuing a career, motherhood, or something entirely different. Let's remind our daughters that their worth does not depend on meeting others' expectations, but on following their own dreams and passions.

c) Comment: Motherhood is an important role, but it does not define a woman's value. Statements like this reinforce the outdated stereotype that a woman's main purpose is to have children. This not only underestimates women's professional abilities but also undermines efforts toward gender equality.

h) Comment: Interesting — I've never seen a post reminding boys that their greatest ambition should be to become fathers. Maybe someone forgot to send them the memo?

f) Comment: Having children is an important decision, but the responsibility should not fall solely on women. Reminding our daughters that they must become mothers places a burden on them and implies that having children is primarily their responsibility, rather than a shared decision. A fair society requires that both genders take equal part in this issue.

n) Comment: We live in a society that values equality, where individuals — regardless of gender — should have the opportunity to decide what matters most to them. Teaching our daughters that career and motherhood are mutually exclusive undermines a core social value: that both paths are equally valid and worthy of respect.

- I completely disagree with the **post** – I completely agree with the **post**
  Slider: 1–6
- I completely disagree with the **comment** – I completely agree with the **comment**
  Slider: 1–6
- Schools **should not address** this topic at all – Schools **definitely should address** this topic
  Slider: 1–6

The next few questions refer only to the post at the top — not the comment below the post.

To what extent do you agree with the following statements about the original post?

1 = Strongly disagree
2 = Disagree
3 = Somewhat disagree
4 = Somewhat agree
5 = Agree
6 = Strongly agree

1. If I saw a post like this, I would click the “like” button on social media.
2. The post is well-intentioned.
3. The post poses a threat to gender equality.
4. The post helps address social problems.
5. The post promotes gender equality.
6. The post creates division between groups.
7. The post promotes fairness.

When thinking about people living in [Country’s name], what percentage of them do you think agree with the original post?

Slider: 0–100%
0% = No one agrees
100% = Everyone agrees

2.3. Filler 2: Fashion

Post: At first, it might come as a surprise that the fashion industry has a massive ecological footprint. Yet it's actually one of the most polluting industries in the world. With the rise of brands operating under the fast fashion model, clothing production has nearly doubled compared to previous levels, now exceeding 100 million tonnes annually worldwide.

Comment:
Comment: It feels like just existing is already too much for the planet. I drink water — the springs dry up. I eat — my footprint grows. I buy clothes — I'm basically a criminal. If I really wanted to be eco-friendly, I'd have to disappear. But as long as I'm still here, I might as well wear something cute.

- I completely disagree with the **post** – I completely agree with the **post**
  Slider: 1–6
- I completely disagree with the **comment** – I completely agree with the **comment**
  Slider: 1–6
- Schools **should not address** this topic at all – Schools **definitely should address** this topic
  Slider: 1–6

The next few questions refer only to the post at the top — not the comment below the post.

To what extent do you agree with the following statements about the original post?
1 = Strongly disagree
2 = Disagree
3 = Somewhat disagree
4 = Somewhat agree
5 = Agree
6 = Strongly agree

1) If I saw a post like this, I would click the "like" button on social media.
2) The post is well-intentioned.
3) The post helps address social problems.
4) The post creates division between groups.

When thinking about people living in [Country's name] , what percentage of them do you think agree with the original post?

Slider: 0–100%
0% = No one agrees
100% = Everyone agrees

2.4. Abortion
Post:
It doesn't matter whether a fetus has already been born or not — they should have rights, and their life is equal in value to that of their mother, and to all of us.
Comments:
e) Comment: While protecting the rights of unborn fetuses is important, we must not forget that women's rights are at least equally important. Claiming that the rights of the unborn override all else ultimately questions a woman's right to autonomy. Women's rights should not be subordinated to the rights of potential life.
c) Comment: When women are defined solely by their ability to give birth, we promote a standard that ignores their individual ambitions, societal contributions, and personal identities. This essentialist view reduces women to this single role, regardless of their other skills, goals, or desires. Treating a fetus as fully equal to its mother reinforces the stereotype that a woman's primary value lies in motherhood — and this undermines the progress made toward gender equality.
h) Comment: Equal rights for all! But let's be honest: only one person is carrying around a watermelon-sized "roommate" 24/7, dealing with nausea, and tolerating comments like "You're glowing!" Maybe we could give her a bit of an edge in making the final call?
f) Comment: According to the principle of equality, every woman should have the right to determine her own future — including when and how she becomes a mother. Treating the rights of unborn life as equal to those of the mother puts women in a position where they lose control over their own lives. A woman's life and decisions are just as valuable as anyone else's, and she should have the final say.
n) Comment: Personal situations are unique. The circumstances leading to pregnancy-related decisions can vary greatly. Every woman's story and context is different, and it is unfair to impose rigid rules on them that fail to account for these complex realities.

- I completely disagree with the **post** – I completely agree with the **post**
  Slider: 1–6
- I completely disagree with the **comment** – I completely agree with the **comment**
  Slider: 1–6
- Schools **should not address** this topic at all – Schools **definitely should address** this topic
  Slider: 1–6

The next few questions refer only to the post at the top — not the comment below the post.

To what extent do you agree with the following statements about the original post?
1 = Strongly disagree
2 = Disagree
3 = Somewhat disagree
4 = Somewhat agree
5 = Agree

6 = Strongly agree

1) If I saw a post like this, I would click the “like” button on social media.
2) The post is well-intentioned.
3) The post poses a threat to gender equality.
4) The post helps address social problems.
5) The post promotes gender equality.
6) The post creates division between groups.
7) The post promotes fairness.

When thinking about people living in [Countrys name] , what percentage of them do you think agree with the original post?

Slider: 0–100%
0% = No one agrees
100% = Everyone agrees

2.5 Gender_lobby
Post:
Meanwhile, the international gender lobby has launched a campaign of conquest — and this conquest threatens our families as well. The goal of gender lobby's conquest is to sow confusion. The aim is to make us uncertain about even the most fundamental things — like whether a father is a man and a mother is a woman.

Comments:
e) Comment: I believe that true family values aren't confined to strict gender roles but are rooted in acceptance and love. People build families in many different ways, and all of them deserve to live in a caring environment. Instead of calling it a 'campaign of conquest,' we should focus on celebrating diversity.
c) Comment: Gender theory isn't about dismantling family structures, but about creating gender equality. Recognizing that 'gender' is more than just a biological label allows us to pursue social justice and eliminate discrimination. Respecting gender identity helps individuals find their place in society and enriches families, rather than threatening them.
h) Comment: Oh, so the real global threat isn't climate change, unemployment, or war… it's that someone might question whether a father must always be a man and a mother always a woman. Got it — time to update my list of priorities!
f) Comment: The fight for gender equality isn't about who can be a father or mother — it's about ensuring that everyone can live freely, regardless of their gender or identity. This doesn't destroy society; it creates the opportunity for a more just world where all people are valued and respected.
n) Comment: The diversity of gender identities and roles has long been part of evolution of societies. Understanding and acceptance help us recognize the wide range of human experiences and identities, which ultimately enriches our communities. Instead of fearing confusion, it is important to foster dialogue and support one another.

- I completely disagree with the **post** – I completely agree with the **post**
  Slider: 1–6
- I completely disagree with the **comment** – I completely agree with the **comment**
  Slider: 1–6
- Schools **should not address** this topic at all – Schools **definitely should address** this topic
  Slider: 1–6

The next few questions refer only to the post at the top — not the comment below the post.
To what extent do you agree with the following statements about the original post?
1 = Strongly disagree
2 = Disagree
3 = Somewhat disagree
4 = Somewhat agree

5 = Agree
6 = Strongly agree

1) If I saw a post like this, I would click the “like” button on social media.
2) The post is well-intentioned.
3) The post poses a threat to gender equality.
4) The post helps address social problems.
5) The post promotes gender equality.
6) The post creates division between groups.
7) The post promotes fairness.

When thinking about people living in [Country’s name], what percentage of them do you think agree with the original post?

Slider: 0–100%
0% = No one agrees
100% = Everyone agrees

2.6 Zero_sum
Post:
What do gender equality enforcement measures lead to? They will result in men being pushed out by force, and positions being filled based not on competence, but on gender. As a consequence, competent men will be replaced by less competent women.

Comments:
e) Comment: It's important to highlight that gender equality initiatives are not about forcing anyone out, but about giving everyone a fair chance. It means that positions should still be filled based on competence, but removing discriminatory barriers helps ensure that everyone truly starts from an equal footing.
c) Comment: Gender equality doesn't mean 'excluding' men—it means ensuring that women also have access to positions that were previously less accessible to them. The goal is for promotions to be based on competence, not gender, and data shows that teams with more gender diversity can be more effective and innovative.
h) Comment: So when a woman gets promoted, it's automatically because she's a woman? But when a man is promoted, that's just the natural order of the universe? Maybe it's time to rewrite that universe to make it fairer.
f) Comment: Measures aimed at gender equality are not about forcing anyone out, but about creating equal opportunities for everyone. Women often face fewer opportunities due to stereotypes, biases, and systemic barriers. Equality doesn't mean ignoring competence—it means ensuring that competent women get a fair shot, too.
n) Comment: We can all agree that competence matters most. But gender equality doesn't mean changing the criteria. This isn't a battle between men and women, it's about fair treatment and access to opportunity. Instead of fearing that women will 'displace' men, we should work together toward equal opportunities, which leads to a better society for all.

- I completely disagree with the **post** – I completely agree with the **post**
  Slider: 1–6
- I completely disagree with the **comment** – I completely agree with the **comment**
  Slider: 1–6
- Schools **should not address** this topic at all – Schools **definitely should address** this topic
  Slider: 1–6

The next few questions refer only to the post at the top — not the comment below the post.

To what extent do you agree with the following statements about the original post?
1 = Strongly disagree
2 = Disagree
3 = Somewhat disagree
4 = Somewhat agree

5 = Agree
6 = Strongly agree

1) If I saw a post like this, I would click the “like” button on social media.
2) The post is well-intentioned.
3) The post poses a threat to gender equality.
4) The post helps address social problems.
5) The post promotes gender equality.
6) The post creates division between groups.
7) The post promotes fairness.

When thinking about people living in [Country’s name], what percentage of them do you think agree with the original post?

Slider: 0–100%
0% = No one agrees
100% = Everyone agrees

2.7 Trad_family
Post:
We don’t need to abandon our traditions in the name of progress. People often say that if we want to live in Europe, we must live like Europeans — with two mothers or two fathers. But that’s not how it’s going to be. We stand for the traditional family. The family is the foundation of everything. We will do everything we can to preserve it in its traditional form: one woman, one man, and children.

Comments:
e) I understand that the traditional family model feels safe and familiar to many. However, the most important thing for children is growing up in a loving and stable environment — regardless of their parents’ gender. Love and support family members give one another matters more than composion of the family. Equality and acceptance can strengthen families, not weaken them.
c) Family is not a rigid concept. Same-sex couples or single parents can provide just as much love, stability, and care for children. Embracing the diversity of family structures enriches society rather than threatening it.
h) Oh, so the "traditional family" only means one kind of structure? Then I guess my single mom — who works two jobs and still manages to raise us with love — must’ve missed that memo. Sorry, Mom — I guess you’re not “traditional” enough. Maybe we can try again when you’re not busy being incredible.
f) A family’s worth and well-being aren’t defined by its structure but by love and support. Why would it be fair to exclude couples who raise their children with care and affection? Equality means everyone has the right to build a family that reflects their life and values.
n) Family isn't a fixed formula — it’s built on love, support, and trust. The world is changing, and so are the ways we form families. What children need most is love and stability, not one “correct” form. Rejecting diverse family models holds back our progress toward equality, which benefits everyone in society.

- I completely disagree with the **post** – I completely agree with the **post**
  Slider: 1–6
- I completely disagree with the **comment** – I completely agree with the **comment**
  Slider: 1–6
- Schools **should not address** this topic at all – Schools **definitely should address** this topic
  Slider: 1–6

The next few questions refer only to the post at the top — not the comment below the post.

To what extent do you agree with the following statements about the original post?

1 = Strongly disagree
2 = Disagree

3 = Somewhat disagree
4 = Somewhat agree
5 = Agree
6 = Strongly agree

1) If I saw a post like this, I would click the “like” button on social media.
2) The post is well-intentioned.
3) The post poses a threat to gender equality.
4) The post helps address social problems.
5) The post promotes gender equality.
6) The post creates division between groups.
7) The post promotes fairness.

When thinking about people living in [Country’s name], what percentage of them do you think agree with the original post?

Slider: 0–100%
0% = No one agrees
100% = Everyone agrees

Section 3.

3.2. How much do you agree with the following statements?

1 = Strongly disagree
2 = Disagree
3 = Neither disagree, nor agree
4 = Agree
5 = Strongly agree

(1) In general, relations between men and women are fair
(2) The division of labor in families generally operates as it should
(3) Gender roles need to be radically restructured
(4) Sexism in society is getting worse every year
(5) Society is set up so that men and women usually get what they deserve
(6) Discrimination against women is no longer a problem in COUNTRY

3.4. How much do you agree with the following statements?

1 = Strongly disagree
2 = Disagree
3 = Neither disagree, nor agree
4 = Agree
5 = Strongly agree

1) In setting priorities, we must consider men and women alike
2) We should not push for equality between men and women
3) The equality between men and women should be our goal
4) Men should dominate over women

Thank you for answering this survey. All questions were part of an academic experiment and the social media posts you saw were not real. Please click the button below to complete the survey.

Appendix D: Pre-registration

Study Information

Hypotheses

1. Main Hypothesis (H1) Counterarguments increase the extent to which individuals recognise that a communication challenges gender equality.

2. Specific Hypotheses by Counterargument Type

H1.1 (Empathy): Empathy-based counterarguments increase recognition of gender equality challenges. Empathy is the ability to understand mental states (e.g cognitive and emotional states) of another person and being able to „feel with" them (feel distress based on their feelings (Batson, 2009). Importantly, empathy is related to different kinds of prosocial behavior. Empathy-driven interventions can reduce prejudice and foster perspective-taking, potentially making individuals more attuned to challenges in social issues, including gender equality. Empathy-based messages can bridge divides and foster greater understanding of marginalized perspectives. For instance, empathy predicts collective action tendencies for disadvantaged groups (Montada & Schneider, 1989; Thomas & McGarty, 2018, Fernando et al., 2014; Selvanathan et al., 2017). Empathy mediates the relationship between outgroup-contact and (lower) level of prejudice toward an outgroup (Pettigrew & Tropp, 2008). In context related to gender equality, empathy predicts a lower level of hierarchy-legitimizing beliefs regarding gender and sexual minorities (Sinclair & Saklofske, 2019).

H1.2 (Fairness): Fairness-based counterarguments increase recognition of gender equality challenges. Perceived injustice can trigger a range of intergroup emotions, such as existential guilt (stemming from awareness of one's own unfair advantage), sympathy (for the oppressed or disadvantaged group), and moral outrage (directed at privileged or oppressor groups). These emotions are powerful drivers of collective action, as they motivate individuals to seek social change (Montada & Schneider, 1989; Thomas & McGarty, 2018; Selvanathan et al., 2017). Existential guilt may prompt individuals to support efforts to reduce inequality as a way to alleviate discomfort associated with unfair privilege. Research has demonstrated that reminding privileged groups of their unfair advantages can reduce their prejudice toward disadvantaged outgroups (Powell et al., 2005). Further, the cooperative (non-violent) demand for gender equality among women has been predicted by perceived injustice and self-respect—the recognition of oneself as having equal rights (Renger et al., 2020). Additionally, perceived injustice has been shown to mediate the relationship between perceived outgroup threat and the tendency to engage in collective actions for women's rights (Szóstakowski & Besta, 2024).

H1.3 (Essentialism): Counterarguments that challenge essentialist beliefs increase recognition of gender equality challenges. Gender essentialism—the belief that gender categories are fixed, naturally rooted in biological sex, and capable of fully predicting behavior (Bastian & Haslam, 2006; Haslam et al., 2000; Rothbart & Taylor, 1992)—has been shown to reinforce social inequalities (e.g., Yzerbyt et al., 1997; Morton et al., 2009). This belief can be used by privileged groups to maintain the status quo, especially when they perceive it as threatened. Previous research (Skewes et al., 2018) found that gender essentialism is negatively correlated with

support for gender-role egalitarianism and positively correlated with the acceptance of gender discrimination and denial of gender inequalities. Challenging essentialist views with constructivist arguments (i.e., highlighting the socially constructed nature of gender differences) could therefore be an effective tool in counteracting anti gender-egalitarian claims.

H1.4 (Humour): Humour-based counterarguments increase recognition of gender equality challenges. Humor has been shown to diffuse oppositional views and create a more receptive audience for complex or controversial topics. Humor-based interventions can lower tension and increase engagement. Subversive, humor-based arguments can be more effective in fostering support for collective action toward gender equality than serious discourse (Vizcaíno-Cuenca et al., 2024). Vraga et al. (2019) highlight two key mechanisms for why humor-based counter arguments may be effective. First, humor serves as a gateway to introduce audiences to engage individuals who may otherwise be disengaged from an issue. For example, satirical messages have been shown to increase the acceptance of climate change, particularly among people with low prior interest in the subject (Brewer & McKnight, 2015, 2017), and to enhance young people's intentions to engage in activism (Bore & Reid, 2014; Skurka et al., 2018). Second, humor can affect how individuals process information. It can enhance the credibility of the messenger, particularly when the audience is unfamiliar with a source (Vraga et al., 2019)

H1.5 (Consensus): Consensus-based counterarguments increase recognition of gender equality challenges. The use of consensus-based messages, which emphasize widespread agreement or shared values, has been found to reduce polarization and strengthen alignment with social norms. Egalitarian norm signaling (Li et al., 2024) not only reinforces prosocial consensus, but also discourages further norm violation (Gelfand, et al., 2024). Studies show that confronting backlash statements aids in inferring gender egalitarian norms as prevalent (Tirion et al., 2023). Even when the perception of consensus is subjective, norm violating messages are more likely to be considered unacceptable to bystanders (Bicchieri, et al., 2022, Tirion et al., 2023). Discussing social norm violations and establishing consensus is crucial as observing unchallenged norm violations can lead others to imitate deviant behavior and norm erosion as a result (Li et al., 2024). H3. Counterarguments decrease the extent to which respondents agree with a post that challenges gender equality.

H4. Counterarguments reduce the estimated share of others in society who agree with the original post. Exploratory Research Question Does exposure to anti-gender egalitarian communication — when unchallenged — influence broader attitudes toward gender equality, such as fairness perceptions, endorsement of traditional roles, and support for equality?

Design Plan

Study type
Experiment - A researcher randomly assigns treatments to study subjects, this includes field or lab experiments. This is also known as an intervention experiment and includes randomized controlled trials.

Blinding
For studies that involve human subjects, they will not know the treatment group to which they have been assigned.

Is there any additional blinding in this study?
No.

Study design
Respondents will first read the following introductory message (translated into the official language of the respective country) before being presented with a series of social media posts: "Secondary schools could engage with the topic of social media to help students navigate in this complex world. In the following section, we would like to learn your opinion about various types of content. You will see social media posts followed by related questions. We are interested in what you think about the contents themselves, and whether you believe these contents would be suitable for discussion among final-year secondary school students." Participants are randomly assigned to a control track or a treatment track.
•Control track: Participants see the original social media posts only, with no counterargument.
•Treatment track: Participants see the same posts accompanied by a counterargument comment. Each counterargument is randomly drawn from one of five predefined categories: empathy-based, fairness-based, essentialism-challenging, humour-based, or consensus-based.

All respondents view a total of seven posts: five posts that challenge gender equality and two filler posts on unrelated topics (over-tourism and fast fashion), included to obscure the study's focus on gender-related content and reduce demand effects. The five key posts reflect recurring narrative strategies often used to challenge gender equality. Each post is based on one of the following themes:
1. Gender Equality as a Threat to the Traditional Family This narrative frames gender equality as undermining the "natural" or "nuclear" family structure, particularly by challenging traditional gender roles and heterosexual norms. It suggests that equality endangers children's well-being and societal cohesion (e.g., Villa, 2017; Blum, 2015).
2. Gender Equality as a Liberal Project This frame presents gender equality as an elitist, liberal agenda that is incompatible with conservative or traditional values. It casts equality efforts as externally imposed rather than grassroots-driven, and exploits ideological divides (e.g., Graff & Korolczuk, 2017; Grabowska, 2022).
3. Gender Equality as a Threat to National Identity Here, gender equality is depicted as a foreign or Western ideology imposed on national traditions, weakening cultural sovereignty and threatening local values (e.g., Gwiazda, 2020; Graff & Korolczuk, 2022).
4. Gender Equality as Oppression of Men This narrative argues that gender equality marginalises men and unfairly limits their rights. It promotes the idea that masculinity is under threat and that current equality efforts are imbalanced (e.g., Graff & Korolczuk, 2022).
5. Reproductive Rights as Dangerous or Immoral This theme portrays abortion rights and related reproductive freedoms as immoral or harmful to traditional values, suggesting they undermine the sanctity of motherhood and family integrity (e.g., Paternotte & Kuhar, 2017; Mayer & Sauer). All materials are presented in the national language of the respective country. Each participant sees one version of each post (with or without a counterargument), depending on their randomly assigned condition.
No files selected

Randomization

Respondents will be randomly assigned to either a control condition or one of several treatment conditions. Those assigned to the control condition will only see social media posts without counterarguments throughout the experiment. Thus, we have variation at the level of posts, where respondents are either seeing posts in the control condition or posts in the treatment condition, which allows us to analyze the effectiveness of counter arguments. By having a group of respondents that only sees unchallenged anti-gender egalitarian communication, we can also test how the exposure to (unchallenged) anti-gender egalitarian messages affects individuals. Respondents in the treatment conditions will see a version of a social media post that includes one of the various counter arguments. For those in the treatment conditions, the specific counter argument they see will vary between posts due to random assignment, exposing them to different counter argument strategies. Due to the randomization, respondents may see a counter argument that uses humour in the first post and one that uses fairness in the next. Each respondent will view a total of five social media posts that we include in the analysis (additionally, there are 2 filler posts on unrelated content). After viewing each social media post, respondents will answer a set of questions designed to measure the extent to which they perceive the message as challenging gender equality. The social media material we use reflects key framing strategies employed by actors challenging gender equality in Italy, Poland, and Hungary. These materials are real-world social media content that represent frames currently present in public discourse in these countries. The literal wording of the counter arguments is generated in the following way: ChatGPT is used to generate a sample of counter arguments that follow the strategies above (e.g. using fairness, humour, etc.), with a subsequent selection of appropriate material being conducted by humans.

Sampling Plan

Existing Data
Registration prior to creation of data

Explanation of existing data
No data

Data collection procedures
Respondents will be recruited online by the survey company YouGov in Poland, Hungary, and Italy. YouGov handles remuneration according its standard rates.
No files selected

Sample size
The study will include approximately 1,300 respondents in each country. Participants will be randomly assigned to one of six groups: five treatment conditions (each featuring a different counterargument strategy) and one control group. Each participant will evaluate five communications that challenge gender equality, resulting in five repeated measures per respondent.

Sample size rationale
This sample size provides a statistical power of 0.80 to detect a small effect size (Cohen's f = 0.1) in a research design with six experimental groups (five treatment groups and one control group) and five repeated measures.

Stopping rule
No data
Variables
Manipulated variables
The key manipulated variable is the presence or absence of a counterargument following the original communication that challenges gender equality. Participants are randomly assigned to one of six experimental conditions:
•Control group: sees only the original post without any counterargument.
•Treatment groups (5 total): each sees the same original post followed by a counterargument drawn from one of five types:
• Empathy-based
• Fairness-based
• Essentialism-challenging
• Humour-based
• Consensus-based Each participant is exposed to five different original posts that challenge gender equality. In the treatment arms, each post is followed by a counterargument; the type of counterargument varies randomly across participants and posts, ensuring between-subject variation in exposure. The manipulated variable is thus the presence and type of counterargument, with six between-subject conditions.

Measured variables
Main Outcome Measures Participants will evaluate each post using the following items:
Agreement with Post Content

•"I completely disagree with the post – I completely agree with the post" (Continuous slider scale ranging from 1 to 6)

Perceptions of the Post Participants will be asked: "To what extent do you agree with the following statements about the original post?" Each item is rated on a 6-point Likert scale (1 = Strongly disagree, 6 = Strongly agree):
1. If I saw a post like this, I would click the "like" button on social media.
2. The post is well-intentioned.
3. The post poses a threat to gender equality.
4. The post helps address social problems.
5. The post promotes gender equality.
6. The post creates division between groups.
7. The post promotes fairness.

Perceived Social Consensus

•"When thinking about people living in [country], what percentage of them do you think agree with the original post?" (Continuous slider scale ranging from 0% = No one agrees to 100% = Everyone agrees) Secondary Outcome

Measures: Broader Gender Attitudes Respondents are asked to indicate their agreement with the following statements on a 5-point Likert scale (1 = Strongly disagree, 5 = Strongly agree):

1. In general, relations between men and women are fair.
2. The division of labor in families generally operates as it should.
3. Gender roles need to be radically restructured.
4. Sexism in society is getting worse every year.
5. Society is set up so that men and women usually get what they deserve.
6. Discrimination against women is no longer a problem in [Country].

Respondents are also asked to indicate agreement with the following statements, using the same 5-point scale:
1. In setting priorities, we must consider men and women alike.
2. We should not push for equality between men and women.
3. The equality between men and women should be our goal.
4. Men should dominate over women.

No files selected
Indices
No data
No files selected
Analysis Plan
Statistical models
We will conduct both post-level and person-level analyses to assess the effectiveness of the counterarguments and the broader impact of exposure to anti-gender egalitarian communication.
Post-Level Analyses: Effectiveness of Counterarguments
•The unit of analysis is individual post evaluations (i.e., repeated observations nested within individuals).
•We use a between-subjects design, where each respondent is randomly assigned to one of six conditions: five treatment groups (each corresponding to a type of counterargument) or a control group with no counterargument.
•Each respondent rates five communications that challenge gender equality, along with two unrelated filler posts. Treatment condition is randomly assigned for each respondent and applies consistently across the five relevant posts.
•Post-level outcomes include:
•Agreement with the post
•Perceptions of the post (e.g., fairness, intention, threat to gender equality)
•Estimated percentage of societal agreement (perceived consensus)
•To analyse treatment effects, we will estimate linear regression models with standard errors clustered at the individual level. The outcome will be regressed on dummy variables representing the five treatment conditions, with the control group as the reference category.

Transformations
No data

Inference criteria
We will use a two-tailed significance level of $\alpha = .05$ to evaluate all hypothesis tests. An effect will be considered statistically significant if the p-value is less than or equal to .05.

Data exclusion

Data quality procedures, including exclusion of low-quality responses (e.g., speeders, straight-liners, or respondents failing attention checks), will be handled by YouGov according to their standard quality assurance protocols.

Missing data

We will exclude missing data from the analysis. No data imputation will be performed.

Exploratory analysis

Person-Level Analyses: Broader Attitudes Toward Gender Equality

•To assess whether exposure to challenged vs. unchallenged anti-gender egalitarian communication influences broader gender-related attitudes, we conduct analyses at the respondent level.

•These outcomes include agreement with general statements about gender roles, fairness, equality (see measurement section).

•We compare average scores between the treatment group (exposed to counterarguments) and the control group (exposed to unchallenged messages).

•We will use independent-samples t-tests or linear regression models to assess whether there are statistically significant differences across conditions.