European Journal of Political Economy 94 (2026) 102853

Contents lists available at ScienceDirect

European Journal of Political Economy

journal homepage: www.elsevier.com/locate/ejpe

Microtargeting and voters' unawareness: Experimental results☆

Freek van Gils[a], Wieland Müller[b,c,d,*], Jens Prüfer[c,e]

[a] The Netherlands Authority for Consumers and Markets (ACM), PO Box 16326, 2500 BH The Hague, the Netherlands
[b] Department of Economics, VCEE, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria
[c] Department of Economics, CentER, TILEC, Tilburg University, PO Box 90153, 5000 LE Tilburg, the Netherlands
[d] Corvinus Institute for Advanced Studies (CIAS), Corvinus University of Budapest, 1093 Budapest, Fővám tér 8, Hungary
[e] School of Economics and Centre for Competition Policy, University of East Anglia, NR4 7TJ Norwich, United Kingdom

ARTICLE INFO

JEL classification:
C92
D72
D82
D83

Keywords:
Disinformation
Social media
Voting
Microtargeting
Disclosure of interests
Laboratory experiments

ABSTRACT

Due to technological innovation, political interest groups sending messages via news platforms have the ability to (i) microtarget news based on individual-level voter data and (ii) obfuscate their identities, which can be exploited to spread disinformation. We experimentally study the implementation of two proposed interventions in the laboratory, aiming to prevent election manipulation by disinformation in various media environments. We find that mandatory disclosure of interests, with or without a microtargeting ban, increases the efficiency of aggregate voter decision-making. However, only the combination of disclosure of interests and a microtargeting ban mitigates sender influence in this stylized voting environment. The implementation of a microtargeting ban without disclosure requirements has adverse effects.

1. Introduction

In democracies, voters require accurate information about political events (*political news*). Social media platforms have rapidly become major sources of political news, with over half of U.S. adults—and a majority of users on platforms like X and Truth Social—now consuming news via these channels. This trend is mirrored globally, with social media's reach shaping public perceptions.[1] The scale of social media is immense: in 2023, 4.76 billion people—60% of the world's population—used these platforms.[2] The alleged role of TikTok in the 2024 Romanian Presidential election and the subsequent annulment of the election result by the Romanian

☆ This paper previously circulated under the title "Social Media and Democracy: Experimental Results." We are grateful to an editor, an anonymous reviewer, Tobias Klein, Boris van Leeuwen, Jan Potters, Jared Rubin, Martin Salm, Francesco Sobbrio, Sigrid Suetens, Mariya Teteryatnikova and seminar participants in Tilburg, Frankfurt, Passau, Lucerne, at the University of East Anglia, OFCOM, the TILEC Workshop on Economic Governance and Legitimacy, and the SIOE Conference (Toronto) for their comments. We thank Anna Osipenko for research assistance. Wieland Müller and Jens Prüfer gratefully acknowledge financial support from a research grant awarded by Data Science Center Tilburg. We have no competing interests to declare. The data of this study are available from the corresponding author upon request. The experiment in this study was reviewed and approved by TiSEM Institutional Review Board at Tilburg University (IRB EXE 2020-002).

* Corresponding author at: Department of Economics, VCEE, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria.
Email addresses: freek.van.gils@acm.nl (F. van Gils), wieland.mueller@univie.ac.at (W. Müller), J.Prufer@tilburguniversity.edu (J. Prüfer).

[1] See https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News-Report_2022.pdf. The European Parliament writes: "Political advertising is central to influencing how people vote, and may affect citizens' perceptions of the legitimacy of their own political system, particularly when published in the run-up to elections. Rules governing political advertising are therefore key to guaranteeing citizens' fundamental rights and the integrity of democratic processes. "(https://www.europarl.europa.eu/RegData/etudes/BRIE/2022/733592/EPRS_BRI(2022)733592_EN.pdf).

[2] https://datareportal.com/reports/digital-2023-global-overview-report.

https://doi.org/10.1016/j.ejpoleco.2026.102853
Received 23 August 2025; Received in revised form 18 March 2026; Accepted 10 May 2026
Available online 19 May 2026

Constitutional Court offer just one prominent example of the high relevance of new technologies affecting voters' perception of political news.[3]

The technologies underpinning social media (large-scale data collection and AI-driven algorithms) enable platforms to generate detailed information about users' political preferences (Kosinski et al., 2013) and characteristics: so-called *user information* (Argenton and Prüfer, 2012). They can also deliver *microtargeted* messages, allowing political interest groups (or advertisers) to differentiate their news reports and influence voters' beliefs in their favor across subgroups of the electorate.[4] Additionally, algorithmic curation often *obfuscates the original source* of news, making it harder for voters to identify who is behind the messages they see. Studies show that nearly half of the digital news consumers were unable to identify the source when accessing news via social media (Kalogeropoulos et al. (2019)).

This paper examines how microtargeting and sender obfuscation—two key features of social media—affect voter perceptions and behavior. Building on the theoretical framework of van Gils et al. (2025), which models interest group communication and voter types along a political spectrum, we move beyond theory to empirically test these dynamics using simplified experimental games. It is difficult to make valid causal claims about the impact of (dis)information on voting behavior based on the often highly aggregated social media data available to researchers (Guess and Lyons, 2020), which complicates evaluations of the effectiveness of changes with observational data.[5] Therefore, we conducted experiments in the lab.

We design four experimental games where a politically motivated actor ("interest group"), who is informed about a binary state of the world sends a cheap talk message concerning that state (the truth or disinformation) to uninformed voters, who subsequently make a voting decision.[6] The interest group and voters are assigned one of two types (Majority or Minority), which determines their preferred policy outcome and thus their incentives. Our key innovation is to test how voting behavior changes when messages are *microtargeted* versus *public*, and when the interest group's type is *disclosed* or *hidden*.

The first half of the experiment mimics real-world social media: the interest group can microtarget messages and its type is undisclosed. In the second half, we introduce three interventions: a microtargeting ban (requiring public messages), mandatory disclosure of the group's type, and a combination of both. This allows us to compare outcomes under each intervention to the baseline. A microtargeting ban can also speak to the importance of an electorate's shared beliefs about political events, as is created by traditional media, such as newspapers, radio or TV channels. As in the first half of the experiment, we let subjects interact repeatedly such that we can observe behavior that arises with experience.

Our results show that mandatory disclosure, with or without a microtargeting ban, improves collective voter decision-making: voters are more likely to trust truthful messages and are less likely to trust false messages due to disclosure, which ultimately leads to more efficient voting actions. However, only the combined intervention of disclosure and a microtargeting ban effectively prevents election manipulation (i.e., inducing voters to choose inefficient policies through misleading communication). A microtargeting ban alone can have unintended negative effects, reducing the efficiency of voter decisions when the interest group is likely to favor the efficient outcome.

Robustness checks using alternative treatment sequences and within-subject designs confirm that our main findings hold regardless of the order in which the games are played. Our laboratory approach offers clear advantages over observational studies: it allows precise control over information exposure, communication, and incentives, and enables direct measurement of voters' beliefs and actions, yielding internally valid insights into the impact of social media technologies and regulatory interventions on democratic processes.

On the flip side, lab experiments can never reach the level of external validity offered by observational data. We try to address this shortcoming in two ways. First, although the regulation of news platforms is an important policy issue,[7] this paper focuses on understanding the stylized effects of communication technologies on voters' beliefs and decisions rather than providing direct policy guidance. Second, by comparing four different technological configurations/games, we focus on the *changes* in voter behavior caused by specific features such as microtargeting and obfuscation, thereby zooming in on the relative effects of technological change while keeping the stylized nature of the setup fixed (akin to a diff-in-diff approach).

To enable such a clean identification, our design necessarily abstracts from key institutional features of real-world elections—such as strategic coordination, vote aggregation, and outcome-dependent payoffs. Instead, we adopt a simplified structure in which each voter makes an expressive decision based solely on their belief about the true state, independently of others' actions. This abstraction

[3] See https://edmo.eu/edmo-news/analysis-of-the-2024-romanian-presidential-elections-the-role-of-social-media-and-emerging-political-trends/ for an overview of facts and https://disa.org/the-influence-of-social-media-on-romanian-election-outcomes/ for a quick assessment of that election.

[4] Facebook's Custom Audience is a prominent example of a microtargeted advertising service. According to investigative journalism outlet ProPublica, Facebook offers a list of 29,000 user categories that ad buyers can use to determine their target audience (https://www.propublica.org/article/facebook-doesnt-tell-users-everything-it-really-knows-about-them).

[5] "It remains the case that the employees of the platforms are the only ones who really know the scale of the problems widely attributed to them. Those of us on the outside must make do with the glimpses provided through publicly available data, which may or may not paint an accurate picture of what is actually going on." Persily and Tucker (2020, p. xvii).

[6] We follow the definition of Tucker et al. (2018): "Disinformation […] is intended to be a broad category describing the types of information that one could encounter online that could possibly lead to misperceptions about the actual state of the world." For instance, by selectively reporting one-sided information (truthfully) an interest group produces disinformation without lying/fake news.

[7] See Allcott et al. (2022), Aridor et al. (2024), and Bursztyn et al. (2025) for recent academic papers offering evidence about all kinds of negative effects of social media. See the EU's Digital Services Act and the UK's Online Safety Bill for legislative initiatives trying to tackle some of these negative externalities.

allows us to isolate the informational effects of communication design. The model belongs to a broader class of multi-receiver cheap talk games with audience heterogeneity, following the logic of Farrell and Gibbons (1989). While we remain cautious in drawing direct electoral implications, our results identify mechanisms—such as how sender transparency and message differentiation affect belief formation—that plausibly extend to more complex voting environments.

The paper is organized as follows. Section 2 reviews related literature. Section 3 describes the theoretical framework as well as the experimental design and procedures. Section 4 presents the experimental results. Section 5 discusses key design choices, while Section 6 concludes. An (online) appendix contains a formal solution of our games and a comprehensive list to robustness checks, alongside experimental instructions and background data.

2. Related literature

We briefly discuss three strands of literature that are most closely related to our study.

The rise of social media has transformed political communication, allowing interest groups to exploit microtargeting and identity obfuscation to spread disinformation. These practices threaten electoral integrity by distorting voter decision-making, as voters increasingly depend on algorithmically curated information without knowing the sources. While van Gils et al. (2025) [henceforth referred to as "vGMP2025"] show that voter unawareness—not microtargeting alone—can alter election outcomes, empirical evidence on effective countermeasures is limited. Our study addresses this by experimentally testing policy interventions. Whereas vGMP2025 consider a continuum of voter and interest group types on a left-right spectrum, we simplify to two types for experimental tractability. In vGMP2025, ideal policy positions depend on a binary state of the world that objectively favors one side; in our framework, preferences are subjective, with interest groups and voters always favoring opposite actions. The effect is similar: different voter groups prefer different parties and are aware of it. While vGMP2025 derive nuanced theoretical results on persuasion and (de)mobilization effects, our framework sets the stage for experimentally grounded insights.

Our focus on voter unawareness connects to a growing empirical literature showing that users often do not know, or do not attend to, who is behind the political content they encounter online.[8] In distributed news environments, Kalogeropoulos et al. (2019) document that correct news-brand attribution is substantially lower when people access news via social media or search than via direct access, indicating that intermediated distribution systematically weakens source awareness. Studies of transparency and disclosure around online political advertising reach similar conclusions. Jansen and Krämer (2023) find that users are frequently unaware they are being targeted and that existing transparency tools do not fully correct this informational asymmetry. Experimental work on disclosure labels, such as Dobber et al. (2021)'s study on online political ad labels, shows that many users do not notice such labels and that their effects on comprehension and attitudes are modest. Carrella et al. (2025) report that warning people that they are being microtargeted does little to reduce the persuasive advantage of targeted ads. Survey evidence meanwhile points to widespread confusion and distrust around data-driven political advertising, with voters reporting that they feel misled by political ads and expressing strong opposition to data-driven targeting and to political ads on social media more generally (Gallup and Knight Foundation, 2020). Large-scale audits of platform ad delivery, such as recent analyses of German election campaigns, also highlight the opacity of algorithmic targeting and argue that current transparency tools are insufficient to evaluate how proprietary algorithms deliver political ads (Bär et al., 2024). This body of work underpins our assumption that many voters are effectively unaware of who is trying to influence them and with what informational advantages, and it motivates our experimental focus on environments with limited voter awareness of senders' identities and interests.

Experimental work on social media and voting is scarce.[9] Pogorelskiy and Shum (2019) report that media bias reduces collective decision-making efficiency in lab settings where voters share verifiable news. Unlike their setup, where voters' types are public and sharing is network-wide, we focus on strategic behavior by interest groups who can misinform, and on features unique to social media: microtargeted news and undisclosed ideological types.

Ziegler (2023) uses lab experiments to test the persuasiveness of public versus private messages, depending on the audience's strategic environment. Unlike his coordination-focused setting, our receivers' payoffs depend on the state, their action, and type, but not on others' actions. Notably, minority receivers in our games use public communication to assess message truthfulness when available; otherwise, they must guess.

Our results align with findings by Battaglini and Makarov (2014) and Drugov et al. (2017), who study information transmission in cheap talk games with public and private communication and transparent sender types. We extend this literature to cases with undisclosed sender types and observe similar patterns.

Our theoretical framework is related to the two-receiver cheap talk game by Farrell and Gibbons (1989), but we extend it by allowing abstention and by studying how public and private communication interact with sender disclosure. Related experiments with uncertain sender types show mixed effects of disclosure: Chung and Harbaugh (2019) find that disclosure helps receivers discount biased messages, while Cain et al. (2005) observe that disclosure can increase sender dishonesty, possibly due to moral licensing. Both study one-receiver settings, whereas we examine communication to multiple receivers, allowing us to analyze interactions between communication mode and disclosure.

[8] See also de Cornière and Sarvary (2023), who incorporate the "dilution" of news brands on social media in a theoretical model.

[9] Kartal and Tyran (2022) and Sun et al. (2021) study related questions but outside social media contexts.

Empirical studies on social media and voting behavior[10] show mixed effects: Bond et al. (2012) and Jones et al. (2017) find Facebook mobilization messages increased turnout; Liberini et al. (2025) and Beknazar-Yuzbashev and Stalinski (2022) report that microtargeted ads can both persuade and demobilize voters. Studies exploiting Twitter adoption (Fujiwara et al. (2024); Rotesi (2019)) find social media exposure can shift vote shares, while Allcott and Gentzkow (2017) and Guess et al. (2020) argue misinformation's impact on election outcomes is limited.

Focusing on the policy measures countering the negative effects of social media, recent research identifies disclosure and microtargeting bans as promising regulatory tools. Tucker et al. (2018) highlight how platform architectures enable disinformation, reinforcing our focus on identity obfuscation. Vaccari and Chadwick (2020) demonstrate that synthetic media heightens disinformation risks, further justifying our emphasis on identity disclosure.

3. Theory, experimental design and procedures

3.1. The four games

We consider a set-up with an electorate that consists of multiple voter groups with different sizes and interests (see also the discussion of vGMP2025 in Section 2). Voters can vote for one of two parties, *party X* or *party Y*, or *abstain* from voting. Voting implies a payoff that depends on the state of the world, which is unobservable to voters. Prior to voting, the voters receive a message about the state from an interest group that observes the true state.[11] The interest group shares the same interests as one of the voter groups. Its message is cheap talk (costless, non-binding and non-verifiable) and may be used to strategically inform or misinform voters. All these elements of the set-up are common knowledge.

For the sake of simplicity, our experimental games only contain two voter groups (*Majority* and *Minority*), one interest group (type *Majority* or *Minority*) and a binary state of the world (*One* or *Two*).

The presence of two voter groups with opposing interests is essential for comparing public and microtargeted communication. If all voters shared identical preferences, a public message and a set of targeted messages would be strategically equivalent. Heterogeneous voter interests create the trade-off that disciplines public communication, as a message tailored to one group may inadvertently inform the other. The Majority voter group captures the 'mainstream' of the electorate and comprises two voters. The Minority group represents a smaller, 'niche' voter group and consists of one voter.[12] In line with the sizes of the respective voter groups, the interest group has type Majority with probability $2/3$ and type Minority with probability $1/3$.[13] The payoffs of the interest group and voters are displayed in Table 1(a) and (b), respectively.

Due to this payoff structure, both voters and interest groups have state-dependent preferences. For voters, the optimal party choice depends on the true state of the world, which creates an incentive to learn the state and makes communication payoff-relevant. Likewise, the interest group's preferred outcome varies with the state. If preferences were state-independent (for example, if actors were simply aligned with one party regardless of the state) informative communication would not arise in equilibrium. As can be seen from Table 1(a), the lowest and intermediate payoffs for the interest group are, respectively, $1/3$ and $2/3$ of the highest interest group payoff.[14] For voters, the intermediate payoff is $2/3$ of the highest voter payoff (Table 1(b)). The lowest payoff is slightly less than $1/3$ of the highest payoff to induce voters who maximize their expected payoff to abstain in equilibria in which a message is uninformative (i.e., when a voter's posterior belief equals the prior belief).[15]

The payoff structure in Table 1 reflects two design choices. First, incentives are aligned by type: Majority interest groups and Majority voters both prefer party X in State One and party Y in State Two, whereas Minority interest groups and Minority voters prefer party Y in State One and party X in State Two. This generates natural alignment within groups (Majority-Majority, Minority-Minority) and misalignment across groups (Majority-Minority), which is crucial for studying strategic communication and potential manipulation. Second, abstention yields an intermediate payoff (20 for interest groups, 60 for voters) that does not depend on the state of the world. This ensures that abstention is the optimal choice when messages are uninformative, allowing us to distinguish equilibria with meaningful information transmission from babbling equilibria in which messages carry no information.

For interest groups, we set the highest payoff at 30, the intermediate payoff at 20, and the lowest payoff at 10. For voters, we use 90, 60, and 27. These values provide clear differences between preferred, intermediate, and worst outcomes while keeping calculations simple for subjects. The lowest voter payoff is deliberately set slightly below one-third of the highest payoff so that, under complete uncertainty, abstention is the unique payoff-maximizing action; if the lowest payoff equaled exactly one-third of the highest, all three actions would yield the same expected payoff in such cases, and abstention would not be uniquely optimal.

[10] See Zhuravskaya et al. (2020) and Persily and Tucker (2020) for reviews.

[11] Interest groups usually have access to expert knowledge and resources to discover the true state. See Shapiro (2016) and Kartal and Tremewan (2018) for a more detailed justification of the assumption that the interest group is perfectly informed about the true state of the world.

[12] Voting game experiments with three voters are quite common. See Großer (2020) for an overview.

[13] This assumption maintains internal consistency between group prevalence and group size and ensures that both types arise with meaningful frequency. It also reflects the fact that real-world interest groups range from narrow lobbies to broad-based movements. By making the Majority type more likely, we can study cases in which organized interests align with mainstream preferences. By assuming 1/3 probability for Minority interest groups, we capture cases in which they represent narrower objectives.

[14] Theoretically, any payoff structure in which the highest payoff is strictly greater than the intermediate payoff and the intermediate payoff is strictly greater than the lowest payoff generates the same predictions. We used payoff levels that are not too close together to make the differences clear and chose a positive amount for the lowest payoff to avoid introducing potential deterrent psychological effects from having a zero-payoff (Murphy et al., 2006).

[15] If the lowest payoff were equal to $1/3$ of the highest payoff, every voting action would be possible in an equilibrium with uninformative messages.

Table 1
Payoffs.

Type	Vote	State	
		One	Two
(a) Interest group payoffs (for each vote)			
Majority	X	30	10
	Y	10	30
	A	20	20
Minority	X	10	30
	Y	30	10
	A	20	20
(b) Voter payoffs			
Majority	X	90	27
	Y	27	90
	A	60	60
Minority	X	27	90
	Y	90	27
	A	60	60

Notes. The payoff that an interest group receives from the interaction with a voter depends on its type (Majority or Minority), the voting action (party X (X), party Y (Y) or abstain (A)) and the state of the world (One or Two). The payoff that voters receive depends on their type (Majority or Minority), voting action (party X (X), party Y (Y) or abstain (A)) and the state of the world (One or Two).

We analyze environments that differ in terms of the interest group-voter communication mode—messages can be the same for all voters (**P**ublic) or tailored to voter types (**M**icrotargeted)—and the transparency regime, which concerns the type of the interest group that can be **D**isclosed or **U**ndisclosed to voters. Combining the two dimensions, we have four different games—PD, PU, MD and MU.

The timing of each game is as follows:

Stage 0: Nature determines the state (One or Two, both with probability $1/2$) and the interest group type (Majority, with probability $2/3$, or Minority, with probability $1/3$). The interest group observes the state and its own type. Voters learn the interest group type in Disclosure games only.

Stage 1: The interest group selects one message (One or Two) for *all* voters in Public games or a separate message for *each* voter type in Microtargeting games.

Stage 2: Voters observe their own message and make a voting decision (party X, party Y or abstain (A)). All payoffs are realized as specified in Table 1(a) and (b). The party that receives a majority of the votes wins the election.[16] There is no election winner if neither party obtains a majority of the votes.

The solution concept for our games is Perfect Bayesian Equilibrium (Fudenberg and Tirole, 1991). As is conventional in the cheap talk literature, we focus on the most informative equilibrium, in which the most information is transmitted from the interest group to voters.[17] Before we informally describe the equilibria of the games, we first introduce some measures. The formal statement of the most informative equilibria of the four games and the corresponding proofs are presented in Appendices A.1–A.5.

3.2. Voting efficiency and communication measures

Our primary interest is the *efficiency* of voter decision-making, both for individual voters and for the electorate as a whole. On the individual level, we consider voting behavior *efficient* (*inefficient*) if it yields the maximal (minimal) voter payoff. Abstention from voting (voting action A) is neither an efficient nor an inefficient vote. On the aggregate level, we consider an election outcome *efficient* (*inefficient*) if the party favored by a majority of the voters receives the majority (minority) of the votes. Specifically, an election outcome is efficient (inefficient) if party X wins (loses) in State One or if party Y wins (loses) in State Two. A *tie* (both parties receive the same number of votes) is neither efficient nor inefficient.

In short, individual voter decision-making is efficient when voters vote in accordance with their own interests. Aggregate voter decision-making is efficient when all voters together vote in accordance with the interests of the majority of the voters. Elections in which all voters vote individually efficiently automatically also yield efficient aggregate voter decision-making. The reverse is not

[16] Note that the payoffs of the interest group and voters only depend on *individual* voting actions and not on the *aggregate* election outcome, which is in line with the expressive voting theory (e.g., Schuessler, 2000). Political parties are not active players and do not receive any payoffs.

[17] All cheap talk games also have a completely uninformative 'babbling' equilibrium.

Table 2
Equilibrium actions.

Game	Interest group	Voter	Probability truth	Probability trust
MD	Majority	Majority	1	1
		Minority	1/2	0
	Minority	Majority	1/2	0
		Minority	1	1
PD	Majority	Majority	1	1
		Minority	1	1
	Minority	Majority	1/2	0
		Minority	1/2	0
MU	Majority	Majority	1	1
		Minority	1/2	0
	Minority	Majority	0	1
		Minority	1/2	0
PU	Majority	Majority	1	1
		Minority	1	1
	Minority	Majority	0	1
		Minority	0	1

Notes. This table reports the equilibrium actions of the interest group and expected payoff-maximizing voters for each of the four games.

necessarily true. Efficient aggregate voter decision-making, which is in the interests of a majority of the electorate, may be the result of manipulating minority voters to vote against their own interests. Hence, from the point of view of democratic legitimacy, the efficiency of individual voter decision-making is most important.

The efficiency of voting behavior is the result of the communication between the interest group and the voters. Our communication measure for the interest group is *truth*, which takes the value one if a message is equal to the true state and the value zero if the interest group reports the opposite of the true state. The communication measure for voters, *trust*, is equal to one if a voting action yields the maximal voter payoff under the assumption that the message received is truthful, and zero otherwise. This measure only captures whether a voter believes that the message from the interest group is correct if voters choose their own expected payoff-maximizing action (i.e., if voters do not have strong other-regarding preferences). Our data from an additional task of the experiment suggest that this implicit assumption is valid. Elicited voter beliefs are consistent with their own payoff-maximizing actions in roughly 95% of the cases.

3.3. Equilibria

Let us first consider the equilibria of the Disclosure (D) games. As can be seen from Table 1(a) and (b) above, the incentives of a Majority interest group are aligned with the incentives of Majority voters and misaligned with the incentives of the Minority voter. For a Minority interest group, the opposite holds true. For the **MD-game**, in which there is Microtargeted communication (M), these payoff structures imply in equilibrium that a message from a Majority (Minority) interest group is truthful and induces trust on the side of a Majority (Minority) voter. If the types of the interest group and voter do not match, the message is unrelated to the true state and the voting action is unrelated to the message.

Table 2 displays equilibrium actions for interest groups and expected payoff-maximizing voters in each of the four games.[18] It shows that in the MD-game there is truthful communication *within* the same group (i.e., from the Majority (Minority) interest group to the Majority (Minority) voters) and that truthful communication is reciprocated by trust. There is no information transmission, meaning that messages are unrelated to the true state, *between* groups (i.e., from the Majority (Minority) interest group to the Minority (Majority) voters) and voters choose their voting actions independent of the message received.

In the **PD-game**, the Disclosure game with Public communication (P), the (mis)alignment between the incentives of the interest group and the Majority voters determines whether there is truthful communication and trust in equilibrium because the Majority interest group is *disciplined* to be truthful to a Minority voter and a Minority interest group can no longer be consistently truthful to a Minority voter due to the requirement that messages are the same for all voters.[19] Table 2 shows that in the PD-game a Majority interest group is always truthful and both Majority and Minority voters always trust a Majority interest group. The messages from a Minority interest group are unrelated to the true state and, therefore, not trusted by any voter.

Moving to the games without Disclosure (U), voters can no longer recognize the type of the interest group. In the equilibrium of the **MU-game**, Majority voters receive a truthful message from a Majority interest group type and a lie from a Minority interest group

[18] In the most informative equilibrium, an expected payoff-maximizing voter always abstains if a message is unrelated to the true state.
[19] Farrell and Gibbons (1989) call these cases, respectively, one-sided discipline and subversion.

type, as can be seen from Table 2. Majority voters trust their message because the interest group is most likely to have the Majority type. Information transmission to Minority voters breaks down: messages are unrelated to the true state and there is no trust.[20]

In the **PU-game**, a Majority interest group is *disciplined* to be truthful to all voters and a Minority interest group lies to all voters, as shown in Table 2. All voters trust their message because they are most likely facing an interest group with the Majority type.

3.4. Predictions

We study whether the efficiency of voting actions in a "social media" environment with microtargeted communication and undisclosed interest group types (MU-game) can be improved by implementing a microtargeting ban (PU-game), mandatory disclosure of interests (MD-game) or a combination of the two interventions (PD-game). Our predictions regarding the effects of these interventions follow directly from the most informative equilibria derived in the previous section.

Microtargeting ban. As the largest voter group, Majority voters are expected to always receive messages tailored to them, regardless of the communication technology in place. Consequently, a microtargeting ban should not affect the efficiency of their voting actions. Minority voters, on the other hand, are expected to benefit from a *discipline effect*: the ban forces an interest group to convey the same information to Minority voters as to Majority voters. Since it is more likely that an interest group is of type Majority, this should increase Minority voters' payoffs from voting: there are two effects affecting the Minority voter, which work in opposite directions. On the one hand, a Minority interest group will now consistently tell a lie to the Minority voter, which is an unwanted by-product of the lie that the Minority interest group tells to the Majority voters. On the other hand, a Majority interest group will now tell the truth to the Minority voter, which is the unwanted by-product of the truth that the Majority interest group tells to the Majority voters. Since it is more likely that the interest group is of type Majority, the message sent to a Minority voter is informative (in expectation) and should increase a Minority voter's payoff from voting.

Despite the fact that the microtargeting ban without disclosure of interests is expected to (weakly) improve the efficiency of voters on the individual level, we predict that this intervention has a negative effect on the aggregate level. Due to the discipline effect, the ban is expected to provide Minority voters with better information about the state of the world. Since Minority voters favor a party that is disliked by most voters, this makes it less likely that aggregate voter decision-making is efficient. Hypothesis 1 summarizes these predictions.

Hypothesis 1 (**Microtargeting ban**).

a. **Voter payoffs.** A microtargeting ban increases the payoffs from voting for Minority voters but has no effect on Majority voters.
b. **Election outcome.** A microtargeting ban decreases the efficiency of aggregate voter decision-making.

Mandatory disclosure of interests. This intervention is expected to be beneficial for both voter groups because it prevents them from giving weight to messages sent by interest groups with misaligned incentives and avoids the discounting of reliable messages. The aggregate effect is ambiguous. While disclosure is expected to stop the manipulation of Majority voters' beliefs by a Minority interest group, it is also expected to make information transmission to the Minority voter possible. Jointly, these two effects may enhance or decrease the efficiency of aggregate voter decision-making.

Hypothesis 2 (**Mandatory disclosure of interests**).

a. **Voter payoffs.** Mandatory disclosure of interests increases the payoffs from voting for Majority voters and Minority voters.
b. **Election outcome.** Mandatory disclosure of interests has an ambiguous effect on the efficiency of aggregate voter decision-making.

Microtargeting ban and mandatory disclosure of interests. Due to mandatory disclosure of interests, Majority voters can accurately assess whether a message is trustworthy or not. The microtargeting ban has no added value for them. Minority voters benefit from the combination of the two interventions: they receive more informative messages due to the microtargeting ban and are better able to recognize reliable messages due to the disclosure of interests.

According to our theory, only a combination of the two interventions under study has an unambiguously positive expected effect on the efficiency of aggregate voter decision-making. A Minority interest group is unable to convey any *false* information to Majority voters due to mandatory disclosure of interests and any (consistently) *truthful* information to a Minority voter due to the microtargeting ban, which mitigates the risk of electing the on aggregate inefficient political party. In contrast, a Majority interest group is in equilibrium able to resolve all uncertainty about the state of the world and can thereby ensure an efficient election outcome.

Hypothesis 3 (**Microtargeting ban and mandatory disclosure of interests**).

a. **Voter payoffs.** Mandatory disclosure of interests in combination with a microtargeting ban increases the payoffs from voting for Majority voters and Minority voters.

[20] Suppose that a Minority voter trusts a message with a probability higher than $1/2$. A Majority interest group would best respond by reporting the opposite of the true state. Since an interest group has the Majority type with probability $2/3$, the Minority voter would be worse off than by ignoring the message, which cannot be the case in equilibrium. For an analogous reason, no equilibrium exists in which a Minority voter trusts a message with a probability lower than $1/2$.

Table 3
Treatment overview.

Treatment	Part I	Part II	Description	Number of sessions	Number of subjects
1	MU	MD	Mandatory disclosure	3	36
2	MU	PU	Microtargeting ban	3	36
3	MU	PD	Mandatory disclosure and microtargeting ban	3	36
4	MU	MU	Status quo microtargeting and undisclosed interests	3	36
5	PD	PU	Obfuscation of interests	3	36
6	PD	MD	Microtargeting implementation	3	36
7	PD	MU	Obfuscation of interests and microtargeting implementation	3	36
8	PD	PD	Status quo public communication and disclosed interests	3	36
9	MD	MU		3	36
10	MD	PD		3	36
11	PU	MU		3	36
12	PU	PD		3	36
			Total	36	432

b. **Election outcome.** Mandatory disclosure of interests in combination with a microtargeting ban increases the efficiency of aggregate voter decision-making.

3.5. Experimental design

Before describing the details, we briefly summarize the basic structure of the experiment. In each round, an interest group sends a message about the state of the world, after which voters choose between two policies that affect the payoffs of the voter groups and the interest group. We implement the four games in a laboratory experiment using the parameter values specified in Section 3.1. Table 3 provides an overview of our treatments. Each session of each treatment employs twelve subjects and consists of two parts. Subjects first play forty rounds of one game (Part I) and then forty rounds of another game (Part II), as displayed in Table 3. The roles of all subjects are randomly determined for blocks of five consecutive rounds. After five rounds, new roles are assigned to all subjects that remain fixed for another block of five rounds. In each half of each part of the experiment, subjects act for five rounds in the role of a (Majority or Minority) interest group, ten rounds in the role of a Majority voter and five rounds as a Minority voter. We implemented role switching to enhance learning and reflect standard practice in signaling experiments (Brandts and Holt, 1992; Cooper et al., 1997a,b; Kübler et al., 2008). At the beginning of each round, the 12 subjects of each session are randomly assigned to groups of four, containing one interest group, two Majority voters and one Minority voter.[21, 22] In every first and fifth round of a block of five rounds, we directly elicit voter beliefs about the true state of the world. In these rounds we also ask subjects in the role of the interest group to predict the voting actions and subjects in the voter roles to predict the type of the interest group (if it was undisclosed). The prediction tasks are incentivized. After Part II, subjects complete a lying aversion task, a risk elicitation task and a survey. For details and justifications, see Appendix A.13.

The main goal is to study whether, starting from the **social media environment** (MU-game), the implementation of (i) a microtargeting ban, (ii) mandatory disclosure of interests and (iii) a combination of both interventions improves the efficiency of voter decision-making. For this purpose, we use the first four treatments listed in Table 3 and report their results in Section 4.2. Subjects are first locked into the social media environment (MU), after which one of the three interventions (MD, PU or PD) is implemented or—for clean comparison—the status quo is maintained (MU). This set of four treatments allows us to study the isolated and combined effects of two policy interventions on what we consider today's social-media environment (MU-game).

The second goal is to study behavior when we start from a **traditional media environment**, where microtargeting is not possible and voters know the interest group's objectives (PD-game), and then move to a game with microtargeting or obfuscated objectives or both or—again for clean comparison—the status quo is maintained (treatments 5–8 in Table 3). Note that this is not just the reverse of the exercise described in the previous paragraph. This second set of four treatments, allows us to mimic the "evolution" of the media landscape which, starting from what we call the traditional media environment (PD-game), saw over time the emergence of microtargeting and obfuscation of senders' identities.

We check the robustness of our results by studying our two main research goals both from a within-subject and a between-subject perspective. Additionally, we use treatments 9–12 in Table 3 to check for experimental order effects because the game played in Part I could set focal points or expectations and thereby influence subjects' behavior in Part II. To economize on space, we report the detailed results of the second research goal (treatments 5–8 in Table 3) and the order-effects analysis in Appendices A.6 to A.8, and only provide a short summary of the results regarding the second research goal in Section 4.3. The main findings are highly robust:

[21] In the experiment, we use neutral wording to avoid framing. For instance, an interest group is called an A-player and a voter is called a B-player.

[22] To ensure sufficient statistical power, we recruited 36 subjects per treatment, organized in three independent laboratory sessions with 12 participants each. Under conventional assumptions of 80% power and a 5% significance level, and taking into account the repeated-decision structure of the experiment, this design allows us to detect treatment effects on the order of approximately 0.3–0.4 standard deviations or larger. The experiment was not pre-registered.

independent of whether we use a within-subject or a between-subject design, and independent of the specific game we let subjects start with, the effects of microtargeting (or a ban thereof) and of obfuscation (or disclosure) of interest group types always have the same effects as reported in Section 4.2.

3.6. Experimental procedures

The experiment was programmed in zTree (Fischbacher, 2007) and was conducted in the laboratory of the Vienna Center for Experimental Economics (VCEE) at the University of Vienna between October 2020 and July 2021, with participants recruited from the center's subject pool, which primarily consists of university students from diverse academic backgrounds. We organized 36 sessions in which 432 subjects participated. Subjects were recruited through the online recruitment system ORSEE (Greiner, 2015) and took part in only one session. Upon arrival in the lab, subjects gave their consent to participate in the experiment. They were free to leave the lab at any point during the experiment without providing a reason.

Prior to the start of Part I of the experiment, subjects were informed that the experiment would consist of two parts, received written instructions for Part I and had to answer a series of tutorial questions correctly. After Part I, subjects received instructions for Part II and had to answer some additional tutorial questions correctly before Part II of the experiment started. At the end of the experiment, subjects were asked to complete a lying aversion task, a risk elicitation task and a survey. Detailed instructions for one of our treatments (MU-PU) can be found in Appendix A.13. Each session lasted about two and a half hours and subject earnings were 42.15 euros on average.[23]

4. Experimental analysis

4.1. Empirical strategy

Our aim is to analyze the effects of the three interventions (a microtargeting ban, mandatory disclosure of interests or a combination of the two measures) compared to the status quo (MU-game) *after* subjects have been locked into the "social media" environment. That is why we here only use data from the first four treatments listed in Table 3, excluding all treatments that do not start with an MU-game from our analysis. We restrict our analysis to rounds 21–40 of each game to purge behavior driven by early-round learning effects. This is a standard approach in experimental signaling and cheap-talk games, where subjects often require some time to learn the strategic environment before reaching more stable behavior (see, e.g., Sánchez-Páges and Vorsatz, 2007).[24] We estimate a random effects model for each of the measures explained in Section 3.2. This approach accounts for the dependency in our data which arises because we have repeated observations on each of 144 subjects.[25] The estimated models have the following form.[26]

$$y_{ith} = \alpha + \beta_1 PU + \beta_2 MD + \beta_3 PD + \beta_4 Part + u_i + \epsilon_{ith}, \quad (1)$$
$$i = 1 \ldots, 144, \quad t = 21, \ldots, 40, \quad h = 1, 2.$$

In (1), y_{ith} is the outcome of one of our voting efficiency and communication measures in round t of part h for subject i.[27] The intercept α is the aggregate outcome rate in the MU-game in Part II of the experiment. The regressors PU, MD and PD represent the three interventions and are equal to one if an observation comes from the respective game and zero otherwise. The variable $Part$ takes on value one if an observation comes from Part I of the experiment and zero otherwise. The subject-specific random effect u_i takes into account that subjects (i) make repeated decisions in the periods (t) of the parts (h) of the experiment. Lastly, the error term is represented by ϵ_{ith}.

4.2. Experimental results

Microtargeting and undisclosed interests. Tables 4 and 5 report the coefficient estimates of α, β_1, β_2, and β_3 in (1) for all our outcome measures.

[23] The frequent switching of player roles and the adequate monetary incentives should have kept subjects' attention levels high during the experiment.

[24] Qualitatively, however, our core results are the same if we include all rounds.

[25] The subject-level random effects models do not (fully) account for potential session effects. We have, however, aimed to minimize session effects by the design of our experiment. To limit static session effects, we have randomized the order of our sessions and made sessions as homogeneous as possible. To minimize the risk of dynamic session effects, we have implemented random matching. Subjects were extensively informed about random matching and could only proceed to the experiment if they answered a test question about random matching correctly. Addressing potential session effects in the design instead of in the model estimation is advocated by Kim (2025), who studies experimental designs in which subjects make repeated decisions within the same session. As a further precaution to limit the risk of dynamic session effects distorting our results, we have restricted our analysis to the second block of twenty rounds, in which behavior is settled more than in the first twenty rounds. In the post-experimental questionnaire, we have asked respondents about their strategies in each of the experimental roles. An overwhelming majority of subjects describe a 'fixed' decision rule. There were only a few subjects who stated they followed a dynamic decision rule. Hence, we have no reason to believe that there were large dynamic session effects. This observation seems to be in line with experimental literature. In a study of static and dynamic session effects in laboratory experiments, Fréchette (2012) concludes that it is difficult to completely rule out dynamic session effects but that it is also hard to find many situations in which dynamic session effects seem to be enormous.

[26] For ease of interpretation, we present a linear probability model. We have also estimated subject-level random effects (ordered) probit models, which produce very similar estimates and give rise to the same qualitative results.

[27] We consider truth and the efficiency of election outcomes (efficient, inefficient, tie) for subjects in the role of interest group and trust and the efficiency of vote choices (efficient, inefficient, abstention) for subjects in the role of voter. In addition, we also look at voter payoffs.

Table 4
Marginal effects of a microtargeting ban and disclosure of interests on truth-telling and trust.

	Measure	Interest group	Voters	Interaction						Intervention							
				Status quo		Microtargeting ban				Disclosure				Combination			
				Coeff.	S.E.	Coeff.	S.E.	Sig.	Pred.	Coeff.	S.E.	Sig.	Pred.	Coeff.	S.E.	Sig.	Pred.
1	Truth	Majority	Majority	0.88	0.05	−0.05	0.07		=	0.07	0.05		=	0.07	0.05		=
2			Minority	0.60	0.07	0.25	0.11	**	+	−0.06	0.12		=	0.34	0.09	***	+
3		Minority	Majority	0.42	0.10	−0.02	0.14		=	0.23	0.14		+	0.30	0.13	**	+
4			Minority	0.70	0.10	−0.30	0.14	**	−	0.23	0.11	**	+	0.02	0.13		=
5	Trust	Majority	Majority	0.74	0.05	−0.09	0.06		=	0.17	0.06	***	=	0.11	0.06	*	=
6			Minority	0.41	0.07	0.14	0.09		+	0.00	0.10		=	0.44	0.09	***	+
7		Minority	Majority	0.77	0.05	−0.17	0.08	**	=	−0.35	0.09	***	−	−0.45	0.08	***	−
8			Minority	0.38	0.08	0.25	0.11	**	+	0.43	0.12	***	+	0.16	0.12		=

Notes. This table reports the estimated outcome rates of the status quo as well as the marginal effects of the interventions microtargeting ban, disclosure of interests and a combination of these two interventions. The marginal effects are estimated using model (1) with robust standard errors, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The symbols '+', '−','=', and '?' in the column Pred. denote that the predicted marginal effect is, respectively, positive, negative, equal to zero, and ambiguous. **How to read:** By means of illustration, in the first row it can be seen that the estimated probability of a Majority interest group being truthful to the Majority voters is 0.88, with a standard error of 0.05, in the status quo (with microtargeting and undisclosed interests). The estimated probability that a Majority interest group is truthful to Majority voters is 0.05 lower (i.e., $0.88 - 0.05 = 0.83$) in the treatment with the microtargeting ban. This difference is, however, not statistically significant at the 10%-level, which is in line with our theoretical prediction (=). Similarly, the estimated probabilities are 0.07 higher in the treatments with disclosure of interests and a combination of a microtargeting ban and disclosure of interests. These differences are also not statistically significant at the 10%-level.

Table 5
Marginal effects of a microtargeting ban and disclosure of interests on voting behavior.

	Voter	Interest group	Vote choice	Intervention													
				Status quo		Microtargeting ban				Disclosure				Combination			
				Coeff.	S.E.	Coeff.	S.E.	Sig.	Pred.	Coeff.	S.E.	Sig.	Pred.	Coeff.	S.E.	Sig.	Pred.
1	Majority	Majority	Efficient	0.64	0.05	−0.08	0.07		=	0.24	0.07	***	+	0.17	0.06	***	+
2			Inefficient	0.14	0.02	0.00	0.03		=	−0.08	0.03	***	−	−0.10	0.03	***	−
3			Abstention	0.24	0.05	0.05	0.07		=	−0.17	0.06	***	−	−0.10	0.06	*	−
4		Minority	Efficient	0.32	0.06	0.00	0.08		=	0.00	0.08		=	−0.03	0.08		=
5			Inefficient	0.43	0.06	0.00	0.08		=	−0.22	0.08	***	−	−0.17	0.08	**	−
6			Abstention	0.20	0.05	0.07	0.07		=	0.28	0.09	***	+	0.27	0.08	***	+
7	Minority	Majority	Efficient	0.33	0.06	0.16	0.09	*	+	−0.07	0.08		=	0.51	0.07	***	+
8			Inefficient	0.32	0.04	−0.13	0.06	**	−	−0.07	0.07		=	−0.28	0.05	***	−
9			Abstention	0.33	0.05	0.02	0.06		−	0.17	0.08	**	=	−0.21	0.08	***	−
10		Minority	Efficient	0.37	0.09	−0.05	0.11		=	0.41	0.14	***	+	0.01	0.12		=
11			Inefficient	0.16	0.06	0.23	0.10	**	+	−0.07	0.07		=	0.12	0.08		=
12			Abstention	0.49	0.07	−0.19	0.10	*	−	−0.34	0.11	***	−	−0.13	0.10		=

Notes. This table reports the estimated outcome rates of the status quo (coefficients and standard errors) as well as the marginal effects (coefficients, standard errors and significance levels) of the interventions microtargeting ban, disclosure of interests and a combination of these two interventions. The marginal effects are estimated using model (1) with robust standard errors, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The symbols '+', '−','=', and '?' in the column Pred. denote that the predicted marginal effect is, respectively, positive, negative, equal to zero, and ambiguous.

In the benchmark, MU-game, as expected, we find higher truthfulness and trust if the sender and receiver have the same type: a Majority interest group is more truthful to a Majority voter than to a Minority voter (estimated coefficients of 0.88 and 0.60, respectively; rows 1 and 2 of Table 4).[28] For a Minority interest group, the reverse holds true (rows 3 and 4 of Table 4). As predicted by the most informative equilibrium, trust is lower for Minority voters than for Majority voters (rows 5 to 8 in Table 4).

In line with our theoretical predictions, a Majority voter is *more* likely to choose an efficient voting action, *less* likely to make an inefficient vote choice and *less* likely to abstain than a Minority voter if the interest group has type Majority (see rows 1 to 3 and 7 to 9 of Table 5). Consequently, a Majority voter obtains higher payoffs from voting than a Minority voter (rows 1 and 4 of Table 6). The outcomes with a Minority interest group are analogous to this case (see rows 4 to 6 and 10 to 12 of Table 5 and rows 2 and 5 of Table 6).

Aggregate voter behavior is also as expected: election outcomes are less efficient with a Minority interest group than with a Majority interest group (Table 7).

Microtargeting ban. Due to the disciplinary effect of public communication technology, a Majority interest group becomes *more* truthful and a Minority interest group *less* truthful to Minority voters, who display a higher level of trust in the messages that they receive. All these effects, except the higher trust of Minority voters in Majority interest groups, are statistically significant (rows 2, 4, 6 and 8 of Table 4).

[28] In this section, we solely analyze behavior and outcomes across games and do not provide statistical tests for our within-game comparative static predictions. An analysis of the four individual games can be found in Appendix A.8.

Table 6
Marginal effects of a microtargeting ban and disclosure of interests on voter payoffs.

Voter		Interest group	Intervention													
			Status quo		Microtargeting ban				Disclosure				Combination			
			Coeff.	S.E.	Coeff.	S.E.	Sig.	Pred.	Coeff.	S.E.	Sig.	Pred.	Coeff.	S.E.	Sig.	Pred.
1	Majority	Majority	74.63	1.87	−2.47	2.58		=	10.08	2.64	***	+	8.48	2.36	***	+
2		Minority	55.35	3.34	−0.78	4.31		=	7.19	4.14	*	+	4.97	4.02		+
3		All	67.68	1.59	−1.24	2.28		=	9.58	2.37	***	+	8.20	2.02	***	+
4	Minority	Majority	59.10	2.55	9.43	3.97	**	+	0.65	3.74		=	24.76	3.01	***	+
5		Minority	65.57	3.93	−9.58	5.71	*	−	14.37	5.50	***	+	−2.73	5.37		=
6		All	61.52	2.34	2.82	3.26		+	5.10	3.39		+	15.66	3.01	***	+

Notes. This table reports the estimated outcome rates of the status quo (coefficients and standard errors) as well as the marginal effects (coefficients, standard errors and significance levels) of the interventions microtargeting ban, disclosure of interests and a combination of these two interventions on voter payoffs. The marginal effects are estimated using model (1) with robust standard errors, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The symbols '+', '−','=', and '?' in the column Pred. denote that the predicted marginal effect is, respectively, positive, negative, equal to zero, and ambiguous.

Table 7
Marginal effects of a microtargeting ban and disclosure of interests on election outcomes.

Interest group		Election outcome	Intervention													
			Status quo		Microtargeting ban				Disclosure				Combination			
			Coeff.	S.E.	Coeff.	S.E.	Sig.	Pred.	Coeff.	S.E.	Sig.	Pred.	Coeff.	S.E.	Sig.	Pred.
1	Majority	Efficient	0.72	0.05	−0.20	0.07	***	−	0.16	0.07	**	+	0.08	0.07		+
2		Inefficient	0.17	0.05	0.02	0.06		=	−0.09	0.06		−	−0.11	0.05	**	−
3		Tie	0.13	0.04	0.15	0.05	***	+	−0.08	0.04	*	−	0.03	0.05		−
4	Minority	Efficient	0.32	0.10	0.05	0.12		=	−0.08	0.14		=	−0.10	0.11		=
5		Inefficient	0.53	0.08	−0.07	0.11		−	−0.03	0.12		?	−0.23	0.11	**	−
6		Tie	0.15	0.04	0.02	0.05		+	0.12	0.08		?	0.33	0.06	***	+

Notes. This table reports the estimated outcome rates of the status quo (coefficients and standard errors) as well as the marginal effects (coefficients, standard errors and significance levels) of the interventions microtargeting ban, disclosure of interests and a combination of these two interventions. The marginal effects are estimated using model (1) with robust standard errors, * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The symbols '+', '−','=', and '?' in the column Pred. denote that the predicted marginal effect is, respectively, positive, negative, equal to zero, and ambiguous.

As a result, the voting behavior of Minority voters becomes more efficient when the interest group has type Majority, but more inefficient when it has type Minority (rows 7 to 12 of Table 5). In contrast to our prediction, the first effect is not sizeable enough to outweigh the latter effect. On average, the microtargeting ban neither has a significant impact on efficient and inefficient voting actions nor on abstention.[29] As a result, a Minority voter's average payoffs are not significantly affected by the ban (row 6 of Table 6).

In line with our theory, we do not observe any significant differences in the truthfulness of messages to Majority voters (rows 1 and 3 of Table 4). At odds with this reporting behavior, Majority voters display a lower level of trust after the introduction of the microtargeting ban (rows 5 and 7 of Table 4). This drop in trust, however, does not lead to significant change in the efficiency of voting behavior of a Majority voters: the microtargeting ban has no significant effect on vote efficiency, vote inefficiency and abstention (rows 1 to 3 of Table 4). Similarly, a Majority voter's payoffs do not significantly change due to the ban (rows 1 to 3 of Table 6).

In line with our prediction (Hypothesis 1(b)), the implementation of the microtargeting ban without mandatory disclosure of interests decreases the efficiency of aggregate voter decision-making. The ban takes away an instrument of the interest group to influence the election outcome in its favor. This results in less efficient election outcomes (Table 7).[30]

Our main findings are summarized in the following result.

Result 1.

a. **Voter payoffs.** A microtargeting ban has no statistically significant effect on the payoffs from voting for Majority and Minority voters.
b. **Election outcome.** A microtargeting ban decreases the efficiency of aggregate voter decision-making significantly.

Mandatory disclosure of interests. As expected, mandatory disclosure of interests only affects the truthfulness of messages sent by a Minority interest group (Table 4). Although our theory predicts a positive effect for messages to both Majority and Minority voters, only messages directed at the latter group become significantly more truthful. All our predictions regarding trust are borne

[29] The estimated coefficients (with standard errors in parentheses) are, respectively 0.08 (0.07), −0.01 (0.05) and −0.04 (0.06) for efficient votes, inefficient votes and abstention.

[30] Note that we only observe a significant effect for election outcomes with a Majority interest group.

out by the data (Table 4): a Majority voter trusts messages from a Majority (Minority) interest group more (less) than before the intervention. A Minority voter has more trust in messages from a Minority interest group than before.

The changes in interest group and voter behavior lead to improvements in voter decision-making for all voters (Table 5 and rows 1 to 3 and 5 of Table 6). On aggregate, we find that mandatory disclosure of interests leads to a significant increase in efficient election outcomes (Table 7). This improvement is driven by outcomes in elections with a Majority interest group (rows 1 to 3 of Table 7).

Summarizing our analysis, we state the following result.

Result 2.

a. **Voting behavior.** Mandatory disclosure of interests significantly increases the payoffs from voting for Majority voters and Minority voters.
b. **Election outcome.** Mandatory disclosure of interests significantly increases the efficiency of aggregate voter decision-making.

Microtargeting ban and mandatory disclosure of interests. Our predictions about the effects of this intervention on interest group behavior are all borne out by the data. The disciplining effect of the public communication technology causes the Majority interest group to be more truthful to a Minority voter (row 2 of Table 4) but has no effect on its reporting behavior to Majority voters (row 1). Due to mandatory disclosure of interests, both voter groups are able to recognize the Majority interest group. As a result, we see that trust increases for all voters (rows 5 and 6).

According to our theory, a Minority interest group can no longer deceive Majority voters by consistently reporting false messages due to the disclosure of interests. Its equilibrium reporting strategy does not change with respect to a Minority voter. Both of these predictions are confirmed (rows 3 and 4 of Table 4). As expected, the intervention reduces the trust of a Majority voter while having no effect on the trust of a Minority voter regarding messages from a Minority interest group (rows 7 and 8).

The changes in interest group and voter behavior due to the intervention unambiguously improve the efficiency of voter decision-making for (i) Majority voters, (ii) Minority voters and (iii) the electorate as a whole. On the individual level, voters choose more efficient and fewer inefficient voting actions (Table 5), which gives rise to higher payoffs for both voter groups (Table 6). On the aggregate level, we do not observe a significant rise in efficient election outcomes (rows 1 and 4 of Table 7). However, the probability of having an inefficient election outcome decreases significantly (rows 2 and 5 of Table 7).

While Table 6 shows that average payoffs for majority voters are slightly lower in the combined treatment (PD) than under disclosure alone (MD), and that minority voters benefit mainly when the interest group is majority-aligned, we interpret the combined policy as beneficial overall. Table 7 shows that aggregate efficiency is at least as high as in the disclosure-only treatment, and sometimes higher. More importantly, the combination of disclosure and a microtargeting ban prevents interest groups from tailoring messages to different subgroups, thereby limiting the potential for strategic manipulation of beliefs. This aligns with our theoretical prediction that only the combined treatment removes both the incentive and the ability to deceive selectively.

Result 3.

a. **Voter payoffs.** Mandatory disclosure of interests in combination with a microtargeting ban significantly increases the payoffs from voting for Majority voters and Minority voters.
b. **Election outcome.** Mandatory disclosure of interests in combination with a microtargeting ban significantly increases the efficiency of aggregate voter decision-making.

4.3. Robustness: interventions in the traditional media environment

In Appendix A.6, we test the robustness of our main findings by reversing the policy trajectory (treatments 5–8 in Table 3). Instead of starting from a "social media" environment with microtargeting and hidden interests, subjects in these treatments begin in a "traditional media" environment where voters receive a public message and know the sender's identity (PD). From this benchmark, we sequentially introduce: (i) microtargeting while keeping disclosure (MD), (ii) obfuscation of sender identity while keeping public messages (PU), or (iii) both features simultaneously (MU). This within-subject design allows us to observe how the same individuals adjust their behavior in response to the incremental introduction of features that characterize modern social media.

The empirical results strongly support the predictions of our theoretical model. Introducing microtargeting lowers the probability that Minority voters choose the alternative corresponding to the true state of the world, while Majority voters are largely unaffected. Obfuscating sender identity reduces the accuracy of choices for both voter groups. These changes also lead to a significant increase in inefficient election outcomes, where the majority vote selects the incorrect alternative. These comparative statics mirror the patterns reported in Section 4.2: both microtargeting and obfuscation reduce the probability that voters choose the alternative corresponding to the true state of the world, and increase the likelihood of inefficient election outcomes in which the majority vote selects the incorrect alternative. Only the joint removal of these two features—public messaging combined with disclosure of the sender's type, as in the PD treatment analyzed in Section 4.2—restores high rates of correct individual choices and collectively more efficient election results.

4.4. Economic mechanisms and real-world illustrations

Our results show that a microtargeting ban and mandatory disclosure of interests are economic complements: each addresses a different failure in political communication, and only their combination clearly improves both individual and aggregate efficiency.

In the baseline MU environment, the interest group can selectively misrepresent the state to different voter types while hiding its alignment, leading to substantial individual inefficiencies and frequent manipulation of Minority voters against their own interests. A microtargeting ban without disclosure (PU) removes differentiated delivery and thus disciplines the sender in some cases. However, consistent with Hypothesis 1, it can reduce aggregate efficiency when the interest group is likely to favor the Majority. Disclosure without a microtargeting ban (MD) allows voters to condition trust on the sender's type and raises individual efficiency as Hypothesis 2 predicts, yet still permits selective deception across audiences. Only in PD, where messages are public and the sender's type is disclosed, do we observe both the highest levels of efficient individual voting and the strongest reduction in sender influence (Hypothesis 3).

The mechanism is straightforward. A microtargeting ban restricts the technology of communication: the sender must commit to one message for all voters, so any lie aimed at one group also affects others. This couples payoffs across audiences and limits profitable patterns of selective deception. Disclosure changes the information structure: voters can identify whether the sender's incentives are aligned with their own and adjust trust accordingly. If either element is missing, some manipulation opportunities remain—either through targeted tailoring (without a ban) or through anonymous persuasion (without disclosure). When combined, the policies jointly shut down both selective deception (facilitated by targeting) and broad deception (facilitated by anonymity).

Two policy domains illustrate how these mechanisms can matter in practice. For energy policy, a fossil-fuel-aligned interest group could, with microtargeting and obscured identity, present a carbon tax as moderate and climate-friendly to environmentally minded users, while emphasizing exaggerated cost and job-loss narratives to cost-sensitive voters, without either group realizing that both messages come from the same interested actor. A microtargeting ban would force a single public narrative, limiting contradictory tailoring. Mandatory disclosure would additionally reveal that the source is fossil-fuel-aligned, enabling voters to discount biased claims. Together, these measures move the environment closer to PD, where our data show higher individual efficiency and reduced manipulation.

A similar logic applies to immigration policy: an anti-immigration group can currently send fear-based messages to some segments and economic-threat frames to others under neutral-sounding branding. A pure microtargeting ban constrains this tailoring but does not tell users who is speaking. Pure disclosure allows users to infer alignment but still permits divergent messages across audiences. The joint regime, mirroring PD in our experiment, makes it harder to exploit informational asymmetries across voter types and mitigates the risk that either group is nudged into supporting policies at odds with its underlying preferences.

5. Model discussion

Competition among senders. Our experiment features a single political interest group sending one message to voters before they decide. This reduced-form approach mimics the algorithmic selection of one message from many, potentially tailored to each voter. In reality, voters encounter multiple messages from various senders. While our study does not empirically address this, theoretical insights suggest that competition among senders affects each game differently.[31] In the *MD-game*, we found that interest groups will always send a truthful message to voters of the same type, and that voters will trust this message due to the disclosed identity of the interest group. Consequently, all uncertainty about the state of the world is resolved as soon as a voter encounters a message from an interest group with a matching type (regardless of how many messages are sent). Relatedly, in the *PD-game* all uncertainty about the state of the world is resolved once a voter encounters an interest group's message with type Majority (regardless of how many messages are sent).

In games without disclosed sender types, decision-making is more complex. In the *MU-game*, Majority voters' beliefs converge to the truth as messages accumulate, but Minority voters remain uninformed regardless of message volume. In the *PU-game*, all voters' beliefs become more accurate with more messages, though at a diminishing rate. Thus, without disclosed sender types, competition among interest groups can improve voter information even when senders have divergent goals.

Voter group size. We focused on scenarios where the Majority group is significantly larger than the Minority, though many real elections are closely contested.[32] What if both groups are roughly equal in size? Theoretical results indicate that in the *MD-game*, only type alignment matters, not group size. In the *PD-game*, voter group size matters because messages target the largest group, and multiple equilibria arise if groups are equal. The main findings remain robust, though, as long as one group is slightly larger. In the *PU-* and *MU-games*, group sizes also matter only under complete symmetry: then messages become uninformative. In the PU-game, communication can break down if the largest interest group aligns with the smallest voter group, but such symmetry is rare in practice.

Microtargeting. Real-world microtargeting has at least two distinct components: (i) inferring granular user characteristics, and (ii) delivering differentiated messages based on those inferences. Our experiment abstracts from (i) by assuming that the sender already knows voter types and instead focuses on (ii), comparing environments with and without the possibility to tailor messages while holding sender information fixed. This reflects both tractability and policy relevance: Kosinski et al. (2013) show that inferring political preferences from digital traces is already highly accurate, making a full ban on inference difficult in practice, whereas recent regulation (e.g., the EU's 2024 Regulation on Transparency and Targeting of Political Advertising) primarily constrains how such

[31] vGMP2025 study these dynamics in a more general model.

[32] For instance, the general elections in Brazil in October 2022 ended with 50.9% of the votes for the challenging Presidential candidate and 49.1% for the incumbent (https://en.wikipedia.org/wiki/2022_Brazilian_general_election).

inferred information can be used for targeting and mandates more public, standardized formats for political ads.[33] Our design is intended to speak directly to this second dimension by identifying how restrictions on differentiated delivery affect belief formation and manipulation once granular information is available.

6. Conclusion

We have theoretically and experimentally examined two policy measures aimed at preventing election manipulation in the digital era: banning microtargeted political messages and mandating disclosure of interest group identities, within a stylized social media context. Our results show that mandatory disclosure, with or without a microtargeting ban, enhances aggregate voter decision-making. However, only the combination of disclosure and a microtargeting ban effectively counters election manipulation. Notably, a microtargeting ban alone can reduce the likelihood that the majority-backed party wins, as seen in our experiments.

At the individual level, mandatory disclosure benefits both Majority and Minority voters, regardless of the microtargeting policy. A microtargeting ban without disclosure does not significantly help individual voting choices. Our model explains this: microtargeting disadvantages Minority voters, who receive no credible messages, while Majority voters can trust messages aligned with their preferences, especially when the sender type is known. Consequently, banning microtargeting benefits Minority voters by making messages more trustworthy, but if non-mainstream voters' objectives diverge from the majority, banning microtargeting alone can harm overall election efficiency. Robustness checks confirm these findings (see Appendices A.6–A.8).

These insights are timely for legislative initiatives, such as the EU's Digital Services Act and Digital Markets Act, which increase transparency (DSA Art. 24, DMA Art. 6(g)) and restrict microtargeting, especially before elections (DMA Art. 6(aa)), and the US Platform Accountability and Transparency Act, which mandates greater disclosure by social media platforms.[34] Our findings suggest that disclosure requirements, rather than microtargeting bans alone, are more effective in safeguarding electoral integrity.

Implementing these policies remains challenging. While microtargeting bans may be easier to enforce technologically, disclosure rules are more complex in practice. Our experiment simplifies reality by using two clear-cut interest group types, whereas real-world settings are more nuanced and multidimensional. While acknowledging these loopholes and open issues regarding policy implementation, we hold that the clarity, causality, and robustness of the experimental results reported here can offer some guidance (and warning) to policymakers facing the challenge of regulating media in times of technological transformation.

Our experiment draws inspiration from the challenges posed by political communication on social media. However, it deliberately abstracts from institutional features of real-world elections (such as strategic voting, turnout considerations, or collective outcomes. This simplification allows us to isolate core informational mechanisms in a controlled setting. Our contribution does not lie in predicting the effects of specific regulatory policies, but in identifying how communication design (through microtargeting and transparency) shapes belief formation and individual decision-making. These insights speak to general forces at play in information environments with heterogeneous audiences.

CRediT authorship contribution statement

Freek van Gils: Writing – review & editing, Writing – original draft, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Wieland Müller:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Jens Prüfer:** Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data for this article can be found online at doi:10.1016/j.ejpoleco.2026.102853.

Data availability

Data will be made available on request.

References

Allcott, H., Gentzkow, M., 2017. Social media and fake news in the 2016 election. J. Econ. Perspect. 31 (2), 211–236.
Allcott, H., Gentzkow, M., Song, L., 2022. Digital addiction. Am. Econ. Rev. 112 (7), 2424–2463.
Argenton, C., Prüfer, J., 2012. Search engine competition with network externalities. J. Compet. Law Econ. 8 (1), 73–105.
Aridor, G., Jiménez-Durán, R., Levy, R., Song, L., 2024. The economics of social media. J. Econ. Lit. 62 (4), 1422–1474.

[33] https://eur-lex.europa.eu/eli/reg/2024/900/oj/eng

[34] https://www.brennancenter.org/our-work/research-reports/law-requiring-social-media-transparency-would-break-new-ground.

Bär, D., Pierri, F., Morales, G.D.F., Feuerriegel, S., 2024. Systematic discrepancies in the delivery of political ads on facebook and instagram. PNAS Nexus 3 (7), pgae247.
Battaglini, M., Makarov, U., 2014. Cheap talk with multiple audiences: an experimental analysis. Games Econ. Behav. 83, 147–164.
Beknazar-Yuzbashev, G., Stalinski, M., 2022. Do social media ads matter for political behavior? A field experiment. J. Public Econ. 214, 104735.
Bond, R.M., Fariss, C.J., Jones, J.J., Kramer, A.D.I., Marlow, C., Settle, J.E., Fowler, J.H., 2012. A 61-million-person experiment in social influence and political mobilization. Nature 489, 295–298.
Brandts, J., Holt, C.A., 1992. An experimental test of equilibrium dominance in signaling games. Am. Econ. Rev. 82 (5), 1350–1365.
Bursztyn, L., Handel, B., Jimenez-Duran, R., Roth, C., 2025. When product markets become collective traps: the case of social media. Am. Econ. Rev. 115 (12), 4105–4136.
Cain, D.M., Loewenstein, G., Moore, D.A., 2005. The dirt on coming clean: perverse effects of disclosing conflicts of interest. J. Leg. Stud. 34 (1), 1–25.
Carrella, F., Simchon, A., Edwards, M., Lewandowsky, S., 2025. Warning people that they are being microtargeted fails to eliminate persuasive advantage. Commun. Psychol. 3, 15.
Chung, W., Harbaugh, R., 2019. Biased recommendations from biased and unbiased experts. J. Econ. Manag. Strategy 28 (3), 520–540.
Cooper, D.J., Garvin, S., Kagel, J.H., 1997a. Adaptive learning vs. equilibrium refinements in an entry limit pricing game. Econ. J. 107 (442), 553–575.
Cooper, D.J., Garvin, S., Kagel, J.H., 1997b. Signaling and adaptive learning in an entry limit pricing game. RAND J. Econ. 28 (4), 662–683.
de Cornière, A., Sarvary, M., 2023. Social media and news: content bundling and news quality. Manag. Sci. 69 (1), 162–178.
Dobber, T., Kruikemeier, S., Goodman, E.P., Helberger, N., Minihold, S., 2021. Effectiveness of online political ad disclosure labels: empirical findings. Information, Communication and the Data Society (ICDS), Summary report available at, University of Amsterdam. https://riipl.rutgers.edu/files/2021/03/Summary-transparency-discloures-experiment_update.pdf.
Drugov, M., Gonzalez, R.H., Kujal, P., Troya-Martinez, M., 2017. Cheap talk with two audiences: an experiment. Working paper, Available at https://doi.org/10.2139/ssrn.3086405
Farrell, J., Gibbons, R., 1989. Cheap talk with two audiences. Am. Econ. Rev. 79 (5), 1214–1223.
Fischbacher, U., 2007. z-Tree: zurich toolbox for ready-made economic experiments. Exp. Econ. 10 (2), 171–178.
Fréchette, G., 2012. Session-effects in the laboratory. Exp. Econ. 15 (3), 485–498.
Fudenberg, D., Tirole, J., 1991. Perfect Bayesian equilibrium and sequential equilibrium. J. Econ. Theory 53 (2), 236–260.
Fujiwara, T., Müller, K., Schwarz, C., 2024. The effect of social media on elections: evidence from the United States. J. Eur. Econ. Assoc. 22 (3), 1495–1539.
Gallup and Knight Foundation, 2020. In U.S. most oppose micro-targeting in online political ads. https://news.gallup.com/opinion/gallup/286490/oppose-micro-targeting-online-political-ads.aspx (Accessed 7 March 2026).
Greiner, B., 2015. Subject pool recruitment procedures: organizing experiments with ORSEE. J. Econ. Sci. Assoc. 1 (1), 114–125.
Großer, J., 2020. Voting game experiments with incomplete information: a survey. In: Capra, C.M., Croson, R.T.A., Rigdon, M.L., Rosenblat, T.S. (Eds.), Handbook of Experimental Game Theory. Edward Elgar Publishing, pp. 376–398.
Guess, A.M., Lyons, B.A., 2020. Misinformation, disinformation, and online propaganda. In: Persily, N., Tucker, J.A. (Eds.), Social Media and Democracy: The State of the Field, Prospects for Reform. Cambridge University Press, pp. 10–33.
Guess, A.M., Nyhan, B., Reifler, J., 2020. Exposure to untrustworthy websites in the 2016 US election. Nat. Hum. Behav. 4 (5), 472–480.
Jansen, M.-P., Krämer, N.C., 2023. Balancing perceptions of targeting: an investigation of political microtargeting transparency through a calculus approach. PLOS ONE 18 (12), e0295329.
Jones, J.J., Bond, R.M., Bakshy, E., Eckles, D., Fowler, J.H., 2017. Social influence and political mobilization: further evidence from a randomized experiment in the 2012 US presidential election. PLOS ONE 12 (4), 1–9.
Kalogeropoulos, A., Fletcher, R., Nielsen, R.K., 2019. News brand attribution in distributed environments: do people know where they get their news? New Media Soc. 21 (3), 583–601.
Kartal, M., Tremewan, J., 2018. An offer you can refuse: the effect of transparency with endogenous conflict of interest. J. Public Econ. 161, 44–55.
Kartal, M., Tyran, J.-R., 2022. Fake news, voter overconfidence, and the quality of democratic choice. Am. Econ. Rev. 112 (10), 3367–3397.
Kim, D., 2025. Clustering standard errors at the session level. J. Econ. Theory Econom. 36 (2), 1–29.
Kosinski, M., Stillwell, D., Graepel, T., 2013. Private traits and attributes are predictable from digital records of human behavior. In: Proceedings of the National Academy of Sciences, vol. 110 (15). pp. 5802–5805.
Kübler, D., Müller, W., Normann, H.-T., 2008. Job-market signaling and screening: an experimental comparison. Games Econ. Behav. 64 (1), 219–236.
Liberini, F., Redoano, M., Russo, A., Cuevas, A., Cuevas, R., 2025. Politics in the facebook era. Evidence from the 2016 US presidential elections. Eur. J. Polit. Econ. 87, 102641.
Murphy, R.O., Rapoport, A., Parco, J.E., 2006. The breakdown of cooperation in iterative real-time trust dilemmas. Exp. Econ. 9, 147–166.
Persily, N., Tucker, J.A., 2020. Social Media and Democracy: The State of the Field, Prospects for Reform. Cambridge University Press.
Pogorelskiy, K., Shum, M., 2019. News we like to share: how news sharing on social networks influences voting outcomes. Working paper, Available at https://doi.org/10.2139/ssrn.2972231
Rotesi, T., 2019. The impact of twitter on political participation. Working paper, Available at http://trotesi.github.io/trotesi/Rotesi_Twitter_Participation.pdf.
Sánchez-Páges, S., Vorsatz, M., 2007. An experimental study of truth-telling in a sender–receiver game. Games Econ. Behav. 61 (1), 86–112.
Schuessler, A.A., 2000. A Logic of Expressive Choice. Princeton University Press.
Shapiro, J.M., 2016. Special interests and the media: theory and an application to climate change. J. Public Econ. 144, 91–108.
Sun, J., Schram, A., Sloof, R., 2021. Elections under biased candidate endorsements — an experimental study. Games Econ. Behav. 125, 141–158.
Tucker, J., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., Nyhan, B., 2018. Social media, political polarization, and political disinformation: a review of the scientific literature. https://ssrn.com/abstract=3144139.
Vaccari, C., Chadwick, A., 2020. Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Soc. Media Soc. 6 (1).
van Gils, F., Müller, W., Prüfer, J., 2025. Microtargeting, voters' unawareness, and democracy. J. Law Econ. Organ. 41 (2), 634–653.
Zhuravskaya, E., Petrova, M., Enikolopov, R., 2020. Political effects of the internet and social media. Annu. Rev. Econ. 12, 415–438.
Ziegler, A., 2023. Persuading an audience: testing information design in the laboratory. Working Paper, Available at https://www.creedexperiment.nl/creed/pdffiles/aziegler_jmp.pdf.