Testing the Causal Impact of Social Media Reduction Around the Globe

Steve Rathje[1]*†, Nejla Asimovic[2,5]†, Tiago Ventura[2,5]†, Sarah Mughal[1], Claire E. Robertson[1], Christopher Barrie[3,5], The Global Social Media Experiment Team[8], & Joshua A. Tucker[4,5], Jay J. Van Bavel[1,6,7]*,

[1] New York University Psychology Department

[2] Georgetown University McCourt School of Public Policy

[3] New York University Department of Sociology

[4] New York University Wilf Family Department of Political Science

[5] New York University Center for Social Media and Politics

[6] New York University Center for Neural Science

[7] Norwegian School of Economics

[8] Detailed authorship contributions of the Global Social Media Experiment team are shown in ***Supplementary Appendix F.4***

* Corresponding authors: Steven Rathje (sr6276@nyu.edu); Jay J. Van Bavel (jay.vanbavel@nyu.edu).

†These authors contributed equally to this work.

Abstract

More than half of the world's population uses social media. There is widespread debate among the public, politicians, and academics about social media's impact on important outcomes, such as intergroup conflict and well-being. However, most prior research on the impact of social media relies on samples from the United States and Western Europe, despite emerging evidence suggesting that the impact of social media is likely to differ across the globe. Building on the results of pilot experiments from three countries (n = 894), we plan to conduct a global field experiment to measure the causal impact of reducing social media usage for two weeks across 23 countries (projected n > 8,000). We will then test how social media reduction influences four main outcomes: news knowledge, exposure to online hostility, intergroup attitudes, and well-being. We will also explore how the effects of social media reduction vary across world regions, focusing on three theoretically-informed country-level moderators: income level, inequality, and democratic strength. This large-scale, high-powered field experiment, and the global dataset resulting from it, will offer rare causal evidence to inform ongoing debates about the impact of social media and how it varies around the world.

Introduction

Social media has become integral to contemporary life, with more than half of the global population actively using social media[1]. On average, people spend over two hours per day on these platforms[2], and social media has become the main way people access news worldwide[3]. Despite initial enthusiasm about social media's potential to break down barriers, foster connections, and spread information, there is growing concern among academics, policymakers, and the public that social media may facilitate the spread of misinformation[4–6], fuel polarization and intergroup conflict[7–10], and harm mental health and well-being[11–14].

As governments and other stakeholders consider appropriate responses to these challenges, there is an urgent need for more evidence about how social media is impacting individuals and society. Yet, most prior evidence on this topic is correlational and has relied primarily on samples from the United States and other Western, Educated, Industrialized, Rich, and Democratic (WEIRD) countries[15]. This is problematic given that the vast majority of social media users reside in countries outside of these contexts[16,17], and that the impact of social media is likely to differ across nations. For example, one review argued that while social media may lead to detrimental outcomes in established democracies, it could lead to more beneficial outcomes in emerging democracies[18]. While some studies find an association between depression and social media use in the Global North, this association has not been found in the Global South[15,16]. A cross-national survey found that most people in 19 developed nations view social media as beneficial for democracy, but the United States was a major outlier, with 64% considering it harmful for democracy[19]. These papers suggest that the results from studies about social media conducted with U.S. or other WEIRD samples may not generalize to other contexts. To address this issue, we plan to test the impact of social media across the world.

Many prior studies that have tried to test the *causal* impact of social media usage have employed deactivation or reduction designs in which people are instructed to temporarily reduce their social media usage or deactivate their social media accounts entirely. For example, prior experiments have had participants temporarily deactivate their Facebook accounts[20–23], reduce their screen time[24–26], or change aspects of their WhatsApp experience[27] to help infer the causal impact of social media or smartphone usage on key outcomes. However, these studies have led to conflicting results – particularly in different cultural contexts. For instance, one large-scale experiment in the US found that deactivating Facebook for one month reduced political polarization[20]. In contrast, a similar deactivation

experiment in Bosnia and Herzegovina found that a week-long Facebook deactivation conducted during the week of conflict commemoration *increased* ethnic polarization for users with ethnically homogenous offline networks[21]. Other deactivation studies, conducted in the United States[28], Cyprus[22], and France[29], have found null effects on affective polarization. While many studies have more consistently found that taking a break from social media improves subjective well-being[20,21,23,24,26] and reduces news knowledge[20,21,27], recent meta-analyses of deactivation studies have revealed mixed results[30–33]. These mixed findings may stem from variation across studies – in geographic regions, design decisions (such as platforms tested, outcomes measured, and time period), changes in platform user composition or design features (e.g., recommendation algorithms) over time, or simply from broader societal changes that occur with time. It is therefore critical to implement a consistent design across multiple contexts to systematically assess how the impact of social media varies around the globe.

We plan to conduct a global field experiment to provide a high-powered test of the causal effects of social media reduction around the world. In this experiment, over 8,000 active social media users from 23 countries will be incentivized to substantially reduce their social media usage for two weeks (see ***Supplementary Appendix F.2*** for the full list and expected sample sizes). Specifically, participants will be instructed to almost completely eliminate their usage of four widely used social media apps on their smartphones: Facebook, Instagram, TikTok, and X, each with between 500 million and 3.3 billion global users.[34] Following the approach of prior international collaborations[35–38], we have assembled a team of more than 200 researchers from around the world to help conduct this large-scale, global field experiment. Unlike previous cross-cultural studies, which are usually cross-sectional surveys[38] or survey experiments[39], this study is a multi-wave field experiment incorporating both survey and behavioral measures of social media usage. Our large sample size provides us with high statistical power to detect small effects, which are crucial to detect given that even small effects will be amplified across the more than 5 billion people globally who use social media platforms for hours each day. Moreover, our global sample allows us to make more generalizable conclusions about the causal impact of social media outside of WEIRD contexts.

We plan to test the impact of this social media reduction on four pre-registered key outcomes: news knowledge, online hostility, intergroup attitudes, and well-being. Since social media has become the primary channel through which people access news[3], we expect that the treatment will (H1) decrease news knowledge, consistent with findings from multiple

prior studies[20–22]. Further, given past research showing that social media amplifies hostile content[40–42], we hypothesize that social media reduction will (H2) reduce exposure to online hostility. These shifts in information consumption are also likely to have downstream effects on polarization and well-being. Specifically, we predict that reducing social media usage will reduce affective polarization, or animosity toward country-relevant political and social out-groups (H3). Prior social media deactivation experiments have produced mixed results for affective polarization[20,22,21], suggesting that this outcome may vary significantly by country. However, based on theoretical work suggesting that social media platforms may exacerbate polarization by amplifying negative content about out-groups[7,42], facilitating online echo-chambers[43,44], and increasing political misperceptions[45], we predict an overall improvement in out-group attitudes from reduced social media usage. Finally, we hypothesize that (H4) social media reduction will improve subjective well-being, as documented in multiple prior deactivation experiments[20,21,23–26]. Social media usage is often theorized to reduce well-being by promoting harmful social comparisons or displacing well-being-boosting activities, such as in-person socializing.[11,46]

Critically, we expect the effects of social media reduction to vary across world regions. To examine this systematically, we will test for heterogeneous treatment effects on the basis of three theoretically informed country-level variables: income-level, democratic strength, and income inequality. We have four pre-registered secondary hypotheses regarding these three moderator variables. Specifically, (M1) we expect that the treatment's impact on subjective well-being will vary based on the income-level of a country, given past work suggesting that the association between depression and social media usage is not present in lower-income nations[15], as well as past theorizing that social media might have positive consequences in lower-income nations[16]. While many have noted a decline in youth mental health in recent years and attributed this decline to smartphones and social media, a similar decline has not been observed in many lower-income contexts, such as in African countries[47]. Further, anxiety disorders, which are often blamed on social media, are more prevalent in high-income nations than low-income nations[48].

We also predict that (M2) the treatment's effect on news knowledge will be moderated by the income-level of a country such that stronger declines in news knowledge will be found among participants from lower-income countries. Recent studies find that in lower-income countries, individuals rely more heavily on social media for news consumption as compared to high-income countries[49–51]. In contrast, individuals in high-income countries often have access to a wider and more diverse range of news sources, which may encourage more news

consumption compared to the incidental news exposure[52,53] that occurs frequently on social media platforms[54].

We hypothesize that the strength of a country's democracy will moderate the treatment's impact on out-party attitudes (M3) such that reductions in social media use will lead to larger declines in affective polarization in more robust democratic contexts. This hypothesis builds on findings from a systematic review which suggests that, while social media may lead to *detrimental* outcomes in established democracies (such as increasing polarization and declining trust), it might lead to more *beneficial* outcomes in emerging democracies and autocracies (such as facilitating political participation and enhancing access to information)[55]. Further, in less (as compared to more) democratic countries, restricted political discourse, tighter control over digital platforms, and higher levels of self-censorship[56] may limit the visibility of partisan hostility online. Thus, reducing social media use in those settings may not lead to the same decline in affective polarization.

Finally, we expect that (M4) country-level inequality will moderate the treatment's effect on exposure to hostility. A recent cross-cultural survey found that people in more economically unequal countries experience greater hostility online[57], potentially due to increased societal instability and heightened status-seeking behavior. If social media exacerbates these underlying tensions more strongly in unequal societies, then reducing exposure to it may be especially effective in lowering perceived online hostility in these settings. See ***Table 1*** for a detailed description of all hypotheses and analysis decisions.

Design Table 1.

Question	**Hypothesis**	**Sampling plan (e.g. power analysis)**	**Analysis Plan**	**Interpretation given to different outcomes**
		Primary Hypotheses		
RQ1: How does social media reduction affect news knowledge, online hostility, affective polarization, and subjective well-being?	Below are our primary hypotheses, which consider the effects across the entire sample: *H1.* Social media reduction will decrease news knowledge. *H2.* Social media reduction will reduce exposure to online hostility. *H3.* Social media reduction will reduce affective polarization. *H4.* Social media reduction will improve subjective well-being.	Our minimum expected sample size is 8,000. Assuming 95% of statistical power, our power analysis (which is informed by a large pilot sample (n = 894) across three countries) estimates that this sample size can detect a minimal effect size of approximately 0.10 standard deviations. We use two-sided tests with adjusted p-values < 0.05 as our measure of statistical significance for all hypotheses. Specifically, for H1 we estimate a minimum detectable effect size of 0.095 SD, for H2 of 0.096 SD, for H3 of 0.097 SD, and for H4 of 0.096 SD.	For hypotheses 1-4, we will run linear multilevel models that test for the effects of the social media reduction treatment on each of the four outcome variables (news knowledge, exposure to online hostility, affective polarization, and subjective well-being). For each outcome, we will add covariates selected via a separate lasso regression on pretreatment variables (e.g. social media usage, political attitudes, political interest, news consumption, pre-treatment outcomes) and demographic variables (e.g., age, gender, race, ethnicity). All models will	For hypotheses 1-3, if we find a significant *negative* effect of the treatment condition on outcome, we will reject the null and find evidence in favor of a causal effect of social media reduction on the specified outcome. For hypothesis 4, if we find a significant *positive* effect of the treatment condition on subjective well-being, we will reject the null and find evidence in favor of a causal effect of social media reduction on the specified outcome. For any effects with adjusted *p-values* greater than 0.05, we will fail to reject the null hypothesis. To evaluate the robustness of these null

			include a country-level random intercept to account for country-level variation in outcomes. See the ***Analysis*** section for full models. For all four primary hypotheses, we will apply the Benjamini-Hochberg False Discovery Rate (FDR) adjustment for the potential for false discovery with our four main hypothesis comparisons. For full transparency, we will present all statistical tests using unadjusted (*p*-values) and adjusted (*q*-values*)*. Our measure of statistical significance for all tests will take as a reference adjusted two-sided tests *p*-values < 0.05. See the ***inference*** subsection.	effects, we will compute one-sided default Bayes Factors (BF10) to quantify the relative evidence the data provides in favor of the null hypothesis compared to the alternative. We will interpret a Bayes Factor of at least 10 times in favor of the null hypothesis over the alternative hypotheses as strong evidence in favor of the null. See the ***inference*** subsection for more details on the Bayes factor models.
Secondary Hypotheses				

RQ2: How does the effect of social media reduction on main outcomes vary as a function of theoretically informed country-level moderators (i.e., income, inequality, and democracy)?	M1: The effect of the treatment on *subjective well-being* will be moderated by a country's income level such that social media reduction produces more substantial improvements in subjective well-being in higher-income countries than lower-income countries. M2: The effect of the treatment on *news knowledge* will be moderated by a country's income level such that social media reduction produces more substantial declines in news knowledge in lower-income countries than in higher-income countries. M3: The effect of the treatment on *affective polarization* will be moderated by a country's strength of democracy such that social media reduction produces more substantial declines in affective polarization in established	Our minimum expected sample size is 8,000, though we anticipate achieving a higher sample size. Assuming 95% of statistical power, our power analysis, using a large pilot sample (n = 894) across three countries, we estimate that this sample size can detect a minimal effect size of approximately 0.20 standard deviations for each moderator. We use two-sided tests with adjusted p-values < 0.05 as our measure of statistical significance for all hypotheses. Specifically, for M1 we estimate minimum detectable effect sizes of 0.20 SD, for M2 of 0.20 SD, for M3 of 0.10 SD, and for M4 of 0.19 SD. The smaller MDE for M3 comes from the fact that our pilot experiments have more variation on the levels of democracy across the three countries.	For secondary hypotheses M1-M4, we will run linear multilevel models that test for the conditional effects of the social media reduction treatment interaction with the moderator described in the hypotheses. Democracy levels are measured using the Liberal Democracy Index from the Varieties of Democracy (V-Dem) project. Economic inequality is captured using the World Bank's Gini estimates. We will use continuous indicators for each moderator as the primary specification, but will also report results using binary indicators to test the robustness of our findings. For each outcome, we will use the same set of covariates selected via lasso for the primary hypotheses. All models will include a country-level random	For secondary hypothesis M1, if we find a significant *positive* interaction effect of the treatment and the country's income level (ordered from low to high-income), we will reject the null and find evidence in favor of the moderation effect. We will also report marginal effects at different levels of the moderator, regardless of the significance of the interaction term, and provide interpretation where relevant. For secondary hypothesis M2, if we find a significant *negative* interaction effect of the treatment and the country's income level (ordered from low to high-income), we will reject the null and find evidence in favor of the moderation effect. For secondary hypothesis M3, if we find a significant *negative* interaction effect of the treatment and the country's strength of democracy (ordered from emerging

	democracies than in emerging democracies M4: The effects of the treatment on *online hostility* will be moderated by a country's level of inequality such that social media reduction produces more substantial declines in online hostility in countries with more economic inequality than in countries with less economic inequality.		intercept to account for country-level variation in outcomes. See the ***Analysis*** section for full models. For all four moderators, we will apply the Benjamini-Hochberg False Discovery Rate (FDR) adjustment for the potential for false discovery with our four secondary moderation hypothesis comparisons. For full transparency, we will present all statistical tests using unadjusted (*p*-values) and adjusted (*q*-values*)*. Our measure of statistical significance for all tests will take as a reference adjusted two-sided tests p-values < 0.05. See the ***inference*** subsection.	to established democracies), we will reject the null and find evidence in favor of the moderation effect. For secondary hypothesis M4, if we find a significant *positive* interaction effect of the treatment and the country's level of inequality (ordered from low to high inequality), we will reject the null and find evidence in favor of the moderation effect. For any effects with adjusted *p*-values greater than 0.05, we will fail to reject the null hypothesis. To evaluate the robustness of these null effects, we will compute one-sided default Bayes Factors (BF10) to quantify the relative evidence the data provides in favor of the null hypothesis compared to the alternative. We will interpret a Bayes Factor of at least 10 times in favor of the null hypothesis over the alternative hypotheses as strong evidence in favor of the null. See

				the ***inference*** subsection for more details on the Bayes factor models.

We will make our anonymized dataset, code, and materials publicly available to other researchers on the Open Science Framework (https://osf.io/xne6k/?view_only=412204c27a56411d8a5bae01c8c9d8e2). Since we anticipate this dataset being a valuable resource for secondary papers among scholars in the field, we have included a number of exploratory outcome variables and moderators in our surveys that can be analyzed for secondary papers.

Methods

Ethics information

Ethics approval for this project was received by the NYU Institutional Review Board (Protocol #: IRB-FY2023-7927). All translations for individual countries will be submitted to the IRB before data collection. Participants will provide informed consent at the beginning of each survey, and will be debriefed after the entire study is over. All participants will be compensated for their participation in all surveys, including the pre-recruitment, treatment assignment, and post-treatment surveys. Survey and bonus payments will be based on the typical payment rate provided by the survey company in each country, adjusted for the local standard of living (See ***Supplemental Appendix F.2*** for a full breakdown of payments by country). Participants will also be paid a bonus payment of $30 (or the equivalent amount in a given country) for complying with the experimental treatment (e.g., reducing their social media screentime for a specified amount of time and confirming this via screen time screenshots). In the control condition, participants will receive a bonus payment of $10 (or the local equivalent) for uploading relevant screenshots.

Design

Our design is based on insights from three pilot experiments (detailed in ***Pilot Experiments*** and ***Supplementary Appendix D***), which demonstrated the feasibility of our approach across three countries (n = 894) and provided promising initial results.

Country-Selection. We selected at least 23 countries in which to run these experiments, with our decision guided by feasibility (number of collaborators and the ability of survey companies to aid recruitment), diversity of selected contexts, and the extent to which the context was under-researched (for detailed information, see ***Supplemental Appendix, Section A***). The final sample includes at least 23 countries (with up to 26 possible

countries), spanning all continents except Antarctica. These countries represent a combined population of approximately 3.2 billion people. These countries are shown visually in ***Figure 1***, and a summary of the characteristics of these countries is in ***Table 1***. We have assembled research teams (which we refer to as "country teams") in each of these countries. Members of country teams will help adapt survey materials to their local context. The final countries included are subject to change (based on potential subject panel unavailability, collaborator dropout, or other factors beyond our control).

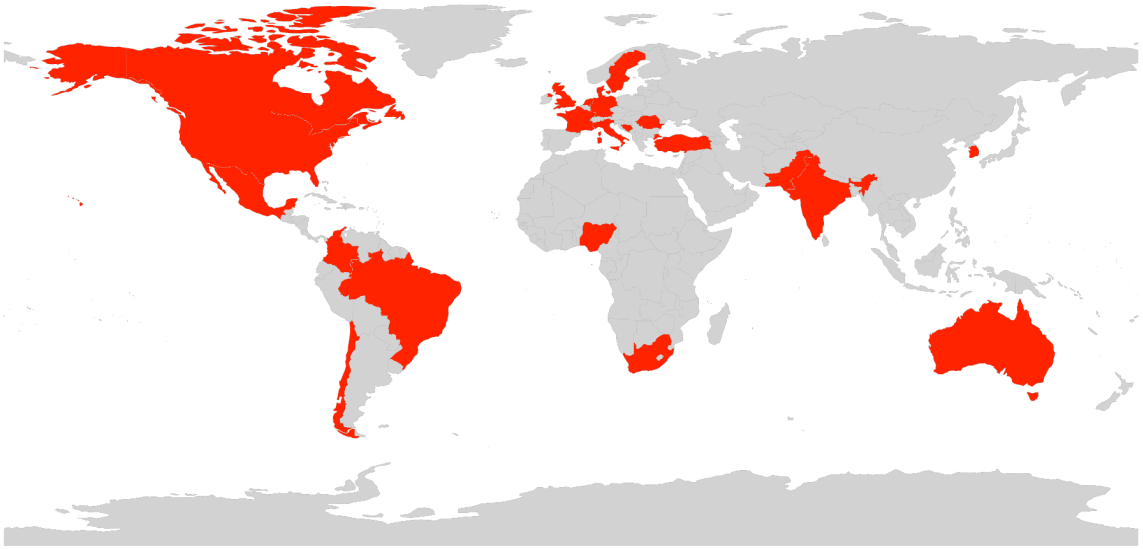

Figure 1. Participants from at least 23 countries (shown in red) will be included in this global experiment: Italy, Bosnia & Herzegovina, Sweden, Germany, the United Kingdom, Romania, Denmark, France, the Netherlands, Mexico, Canada, the United States, Chile, Brazil, Colombia, Nigeria, South Africa, India, Turkey, Pakistan, Singapore, South Korea, and Australia. We have assembled research teams in each of these countries who will help adapt survey materials to their local context.

Criteria	No. of Countries	% of Sample
Income		
High income	13	56.52%
Upper middle income	7	30.43%
Lower middle income	3	13.04%
Democracy level		
More democratic (>0.5 V-Dem liberal democracy score)	17	74%
Less democratic (<0.5 V-Dem liberal democracy score)	6	26%
Geographical coverage		
Europe	9	39%
Africa	2	9%
North America	3	13%
South America	3	13%
Asia	5	22%
Australia	1	4%

Table 2. Characteristics of the 23 countries selected for the study. The selected countries represent all six continents and vary among a variety of key dimensions of interest. The countries represent a combined total population of 3.2 billion people. Income classification scores come from the World Bank and democracy level scores come from the Varieties of Democracy (V-Dem) Liberal Democracy Index scores.

Translations. Surveys will be adapted into each country's most widely spoken language (see translation guidelines in ***Supplementary Appendix B.1***). Each country team will have forward and back-translators, along with translators assigned to adapt screenshot instructions. In addition to translating the language, participating country teams will provide a list of country-relevant in-groups and out-groups (both political and social) for affective polarization questions. Country teams will also select relevant news stories for the news knowledge index (Procedures for selecting included groups and news headlines are in ***Supplementary Appendices B.2 and B.3***.)

Participant Recruitment. To recruit high-quality, representative samples across international contexts, we will recruit participants via the survey companies Bilendi and Netquest, which are experienced firms used successfully in prior cross-cultural studies[39]. Based on our available funding for participant expenses and quotes from these survey companies, our minimum target sample size is at least 8,000 participants across at least 23 countries (All data collection sources and anticipated sample sizes per country are shown in ***Supplementary Appendix F.2***.)

Experimental Design. This longitudinal between-subjects field experiment will consist of three surveys: (1) a *pre-treatment survey*, (2) a *treatment assignment survey*, and

(3) a *post-treatment survey.* We plan to administer the treatment survey one week after the pre-treatment survey. The post-treatment survey will be administered two weeks after the treatment assignment survey. A recent meta-analysis[32] focusing on mental health suggests that social media reduction interventions should last at least one week to yield the strongest effects. Given logistical constraints – and because our intervention, unlike others, spans multiple social media platforms – we have selected a two-week reduction period as a meaningful yet feasible duration. The experimental design is represented visually in ***Figure 2***.

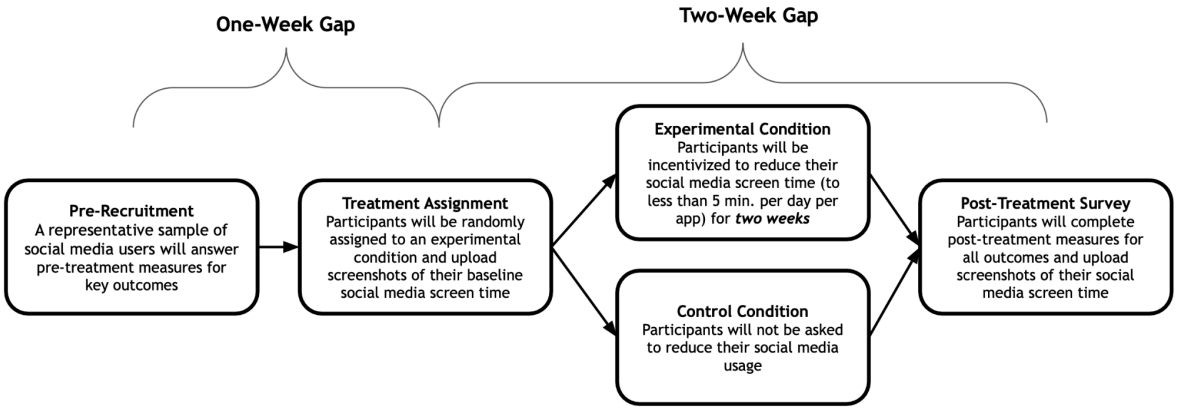

Figure 2. *Experimental design.* A representative sample of social media users (by age and gender) from at least 23 countries will first complete a pre-treatment survey in which they confirm their eligibility and answer pre-treatment measures for the key outcome variables. Then, one week later, participants will complete a treatment assignment survey in which they are randomized to an experimental or control condition and upload screenshots of their baseline social media screen time. Participants will be incentivized to substantially reduce their social media screen time for four key apps (Facebook, Instagram, TikTok and X) to less than five minutes per day per app for either two weeks (in the experimental condition) or zero days (in the control condition). Then, two weeks later, participants will complete a post-treatment survey in which they again upload screenshots of their social media screen time (to ensure compliance with the experiment) and fill out all post-treatment outcome variables.

(1) **Pre-Treatment Survey.** The pre-treatment survey will assess whether participants are eligible and willing to participate in the experiment. To proceed, participants must confirm they are over 18 and provide consent. We are focusing on mobile social media usage because 84% of social media visits are on a mobile device[58]. From a practical perspective, smartphones have built-in software that tracks the amount of time users spend on different apps, allowing for behavioral compliance to be monitored via screen-time screenshots. Screenshots have been successfully used to measure compliance in a number of prior social

media deactivation and reduction experiments.[26,59] To make the treatment maximally effective, we are focusing on users who primarily use social media on their smartphones. Specifically, participants must confirm that they: 1) own a smartphone with an iOS or Android operating system, 2) use at least one of the social media platforms included in the experiment for at least thirty minutes per day, 3) use social media at least half the time on their smartphone, 4) are willing to participate in two follow-up surveys, and 5) are willing to reduce their smartphone social media such that it does not exceed 5 minutes per day per platform. This final condition ensures that treatment and control participants are comparable in their baseline willingness to participate, irrespective of their assigned group. Participants who do not meet these five criteria are screened out and will not proceed with the pre-treatment survey. If we observe challenges in recruiting our target sample size, we will amend the screening criteria to allow participants with fifteen minutes of baseline social media usage of a key platform to participate. Eligible participants will complete pre-treatment measures of our key outcome variables (e.g., news knowledge, well-being, affective polarization, and exposure to hostility). See ***Supplementary Appendix C*** for the phrasing of all of these pre-treatment variables. Participants will have five days to complete the pre-treatment survey. Instructions for uploading screenshots are provided in ***Supplementary Appendix D.3.***

(2) **Treatment Assignment Survey.** The treatment assignment survey will be distributed approximately one week after the pre-recruitment survey. Afterwards, participants will be randomly assigned (via a random number generator in Qualtrics) to either the experimental or control condition. In the *experimental* condition, participants will be asked to reduce the usage of four widely used social media apps to less than five minutes per day for each app for two weeks in exchange for a bonus payment of up to $30 (or the equivalent amount in a given country). This bonus payment will differ by country and will be decided by the survey company based on the typical survey payment rate in each country (See ***Supplementary Appendix F.2***). We selected a five-minute screen time limit to effectively eliminate all usage of these four apps, while also allowing for the possibility of accidental or limited essential use. We primarily focus on the largest social media platforms with algorithmic "news feeds," rather than messaging apps such as WhatsApp, due to widespread concerns about the effects of algorithmic social media[60] and practical concerns associated with removing key communications channels.

To receive a bonus payment, participants will be informed that they must upload screenshots in the *post-treatment survey* – two weeks later – demonstrating that they kept their screen time below five minutes per day (or 35 minutes per week) for each designated app throughout the two-week treatment period. Participants will also be asked to set reminders on their phones to alert them when they reach this five-minute daily limit for the key apps. Participants will be told that screenshots will be closely reviewed to ensure they have not been digitally manipulated, are not sourced from the internet, and are not duplicates of screenshots submitted by other participants to discourage cheating. Screenshots will be reviewed by members of each country's research team (as well as with a commercial large language vision model) to measure the degree to which participants complied with the treatment. Additionally, we will ask respondents at both screening and post-treatment stages how they access social media. At screening, they will report whether they primarily use a smartphone or other devices. We will only invite participants who, during the initial survey – before learning about the study – report using social media mostly or exclusively on smartphones. At post-treatment, they will be asked whether they accessed social media via (a) their smartphone's internet browser or (b) a browser on another device. We will also present results separately for those who did and did not use browsers to further evaluate any influence.

In the *control* condition, participants will be instructed to continue using social media as usual in their daily habits. *Control* condition participants will receive $10 (or the equivalent amount in a given country) for correctly uploading screenshots of their screen time at the end of the experiment. Pilot testing illustrated that the amount paid in the control condition ($30 vs. $10) did not significantly impact attrition or key outcome variables (see ***Supplementary Appendix D.2.2***).

Some prior deactivation experiments asked control participants to deactivate for a shorter period to ensure both groups were willing to deactivate[20], while others – especially those with shorter interventions – did not[21,22]. We adopt the latter approach in our reduction experiment for several reasons. In our three pilots, asking control participants to reduce usage for one day led to unintended reductions at posttreatment (likely due to priming about ways to reduce social media usage) and confusion among control group participants. Given these threats to internal validity, we will not ask control participants to reduce their usage in the full study. Instead, we will pre-screen for willingness to deactivate. Only participants who confirm their willingness to deactivate for up to 14 days will be included. To ensure

comparability in motivation and effort across groups, all participants will be required to upload screenshots.

Immediately after being randomly assigned to condition, participants will upload screenshots of their social media screen time (using either the Screen Time application automatically installed on iPhone devices or the Digital Wellbeing application automatically installed on Android devices) to get a baseline measure of social media usage. Participants will be asked to take screenshots that show their total screen time for the past week, as well as screenshots showing all apps they use for more than five minutes per day (or 35 minutes per week) for the past week. In order to ensure understanding, we will provide: a) video instructions of how to access the relevant app usage statistics; b) insert an on-the-spot mini-quiz to check they know how to take a phone screenshot; c) ask how much confidence they have in uploading the correct screenshots; d) add a bonus reminder that they might not receive the $10/$30 (or local equivalent) payment if they upload the incorrect screenshot (described in detail in ***Supplementary Appendix Section D.3.6***).

(3) **Post-Treatment Survey.** In the post-treatment survey, participants in both conditions will again upload screenshots of their social media screen time from the past week. Afterwards, participants will complete a survey containing all main outcome measures and secondary outcome measures. Secondary outcome measures will not be reported in this paper and will be analyzed in secondary papers.

Measuring Compliance. On both iPhone and Android, participants will self-report their total screen time and screen time for specific social media apps from the past week in addition to uploading screenshots of their screen time for these apps. Due to survey length constraints, participants will upload screen time screenshots only from the second week of the intervention, likely yielding a more conservative estimate of reduction since a full week will have passed since the treatment began. iPhone and Android users will follow different procedures, as Android's Digital Wellbeing feature only tracks daily usage. To avoid increased burden and potential dropout, Android users will report screen time for a randomly selected day with supporting screenshots. Given these platform differences, iPhone and Android data will be analyzed separately in supplemental analyses.

All screenshots will be classified using image-to-text classification from multimodal large language models from OpenAI. Based on data from the three pilot studies, language-model-based classification to measure compliance (asking the models to extract

usage based on screenshots) achieved 90% accuracy or over when compared to human coding. In the case of one pilot (Brazil), the accuracy was slightly lower, likely due to some confusion relating to the screenshot upload instructions. In addition, manual annotators from each country team (see ***Supplementary Appendix Section F.1***) will be instructed to manually annotate at least 20% of randomly selected screenshots from each country. In countries where the LLM-based classification achieves less than 90% accuracy compared to our team's human annotations, we will ask annotators from each country to manually label all the screenshots. Additionally, we will be using a four-step procedure to mitigate the potential of gaming compliance. This is designed to screen out those who use different devices, edit screenshots or use available images from the internet. We set out this procedure in full in ***Supplementary Appendix Section D***.

Outcomes. The key outcome variables, which will be included in the pre-treatment and post-treatment surveys, are described below. Full question wording of the main outcome variables is available in the ***Supplementary Appendix Section C***.

News knowledge. Participants will be asked about their knowledge of eight news stories that were posted in their country during the experimental window on the post-treatment survey. Six of these stories will focus on domestic issues, while the other two will focus on international issues. We will measure news knowledge using two outcomes. First, we will use a headline recall task[27,61], in which we show participants a set of headlines, and ask them a recall question ("Do you remember seeing this news headline in the past two weeks?"), which will have the answer options "Yes", "No", and "Unsure". Second, with the same set of headlines, we will ask them an accuracy question[20,21] ("To the best of your knowledge, is the claim in the above headline accurate?"), which will have the answer options "True", "False", and "Unsure". To select these headlines, country teams will use news collected from Google News over the past two weeks, and will consider these Google News headlines together with headlines from major local newspapers that they consider to be more relevant in each country. Among the eight headlines, six will be factual news articles and two will be placebo fake articles (i.e., changing the direction of the news articles so that the headlines refer to facts that did not happen). The latter are selected to avoid acquiescence bias, following other deactivation designs[21,29]. Our outcome of News Knowledge is captured as an index summing the number of correct accuracy questions and the number of correctly recalled headlines.

Exposure to online hostility: We will ask participants to reflect on the past two weeks and indicate how often they have observed or experienced the following while using the Internet: hostile content (e.g., humiliation, bullying, threats) and rude content (e.g., ridicule, offensive language) on a 7-point scale from "never" to "very frequently." The categories are adapted from an existing survey measure of online hostility[57]. Our outcome of exposure to online hostility is captured as an index of standardized responses to these two items, but we will also report results for each individual item separately in the supplementary appendix as a robustness check.

Affective Polarization: Affective polarization refers to animosity toward individuals, whether political or social. We measure it using two widely used and validated approaches: the feeling thermometer and social distance[62]. The feeling thermometer assesses general attitudes by measuring how positively or negatively individuals feel toward their ingroup versus outgroup (on a scale from 0 to 100), while social distance captures comfort with varying levels of closeness to the outgroup, reflecting attitudes toward specific interactions (such as being neighbors or co-workers with an outgroup member). Participants will indicate all behaviors they are comfortable with by selecting from a predefined list. Our affective polarization index is constructed by averaging the standardized values of the feeling thermometer and social distance scores.

While we measure both political and social group polarization, our main pre-registered analysis relies on the expertise of our collaborators to determine which dimension – political or social – is most salient in their country. See ***Supplementary Appendix Table 2*** for collaborators' assessments of whether political or social polarization is most relevant in their country. As a secondary analysis, we report political and social polarization separately.

Our methods for identifying political in-groups and out-groups are as follows: To identify political out-groups, participants indicate the party they feel closest to. To identify social in-groups, participants indicate their ethnic, racial, or religious identity. Outgroups (whether political or social) are defined as the group participants report feeling most negatively toward from a set of predefined options (see ***Supplementary Appendix Section B.2*** for the guidelines for selecting groups; see ***Supplementary Appendix Table 2*** for the groups selected by collaborators). In line with recommended best practices for measuring polarization cross-culturally[63], we rely on local knowledge from our collaborators to help

define political and social cleavages, and also use a variety of different outcomes (such as social distance measures and feeling thermometer measures) to build a composite index for polarization. As a robustness check, we will report results separately for the feeling thermometer and social distancing items in the supplementary appendix, in addition to reporting in-party affect, out-party affect, and the gap between in-party affect and out-party affect separately. We will also report supplementary analysis with an alternative operationalization of the outgroup as the average feeling toward all groups other than the one with which the participant identifies.

Subjective well-being. Participants will be asked about their subjective well-being using an index adapted from prior deactivation studies[20–22,40,63]. This index will include two questions about happiness, as well as one question each about life satisfaction, loneliness, anxiety, depression, and ability to focus over the past two weeks[26]. Participants will say how frequently they felt each of these over the past two weeks on a five-point scale from "not at all" to "very much." We will also report results for the individual well-being items separately in the supplementary appendix.

Demographics and other baseline measures. Participants will be asked to report their age, gender, political ideology, subjective social status, and the zip code of their current residence in the pretreatment survey.

Auxiliary measures: These are survey measures that we capture to help interpret the results, but they will not be answering an explicit research question. These include: self-reported measure of compliance, confidence in the news recall, offline and online network heterogeneity, substitution behaviors, expectations of future social media use, and information about the content participants saw online. These variables will help contextualize and guide the interpretation of our findings. All auxiliary and secondary survey measures are available on our OSF.

Sampling plan

We are aiming for roughly 400 participants at the post-treatment survey per country (expected sample size in each country shown in ***Supplementary Appendix Section F.2***). To account for limited participant availability in certain countries, we will over-recruit in other countries where higher participation is feasible. We plan to collect more than 8,000 participants across 23 countries and have grant funding set aside for this data collection. We

plan to collect data within a six-month window, launching four to five country-experiments per month. The survey companies will recruit participants using non-interlocking quotas, aiming for a balanced distribution across three age groups (18–34, 35–49, 50+) and two sex categories (male and female).

Data inclusion and exclusion. Once data collection is complete, we will merge data files from all countries and include only participants who complete all surveys. This will automatically exclude participants screened out during the pre-treatment survey (e.g., those who did not consent, failed an attention check, or did not meet screening criteria).

Target Population. Our target population includes users of Facebook, Instagram, TikTok, or X who: 1) own a smartphone with an iOS or Android operating system; 2) use at least one of these platforms for a minimum of thirty minutes per day on their smartphone; 3) primarily access social media via their smartphone; 4) are willing to participate in two follow-up surveys; and 5) are willing to reduce their social media usage such that it does not exceed 5 minutes per day per platform for a two-week period. Our main analysis estimates the sample average treatment effect of providing incentives to reduce smartphone-based social media use across these four platforms, relative to a control group who is similarly willing to reduce their usage but receives no incentive.

Analysis Plan

Main Analysis. Our main analysis will use intent-to-treat effects (ITT) on the combined sample across all countries. Our primary specification will use a multilevel linear mixed effects model that tests for an effect of the treatment on our main outcome variables (i.e., news knowledge, exposure to hostility, affective polarization, subjective well-being), as specified below:

$$Y_{it} = \alpha_j + \tau T + \beta_k X_{it-1} + \epsilon_{it}$$

Where α_j is a random intercept for each country j, T indicates the treatment assignment to reducing social media usage for two weeks, and X_{it-1} is a vector of baseline covariates. We will select covariates X_{it-1} using a lasso regression on all pretreatment variables (e.g. pre-treatment outcomes, social media usage, political attitudes, political interest, news consumption, etc.) and demographic variables (e.g., age, gender, race,

ethnicity). We will use all covariates measured in the pre-treatment and treatment assignment survey in the Lasso selection procedure, and estimate one model for each primary outcome. We will also report non-covariate-adjusted results as a robustness check.

For our secondary moderation analyses, we will estimate multilevel linear mixed effects models that test for an interaction between the experimental treatment and our key moderator effects of interest at the outcome variable of interest, adjusting for the same covariate variables selected via lasso in the primary hypothesis models. The model is specified below:

$$Y_{it} = \alpha_j + \tau_1 T + \tau_2 M + \tau_3 M * T + \beta_k X_{it-1} + \epsilon_{it}$$

Inference. For all statistical tests, we will use two-sided tests with $p < 0.05$ as our measure of statistical significance. We will use a Benjamini-Hochberg False Discovery Rate (FDR) adjustment as our main inferential threshold to deal with multiple hypothesis testing. For the primary hypotheses, we will use *p*-values adjusted for the number of primary hypotheses pre-registered. For our country-level moderator hypotheses, which we consider a distinct set of hypotheses, we will use *p*-values adjusted for the four secondary moderation hypotheses. For pre-registered exploratory analysis, we report unadjusted *p*-values. All statistical tests will be reported using unadjusted and adjusted *p*-values.

Even in the case of null effects, we cannot rule out the possibility that social media may have impacts in other ways. For example, it may impact people on different outcomes (e.g., attention), over longer periods of time, or by changing collective behavior[64,65]. It may also have small effects that cannot be measured even through this high-powered sample. In the case of null effects, we will also measure Bayes factors to test the strength of evidence for the null hypothesis, as discussed in ***Table 1***. For any statistical tests where we fail to reject the null hypothesis, we will compute one-sided default Bayes Factors (BF10) to quantify the relative evidence the data provides in favor of the null hypothesis compared to the alternative. We will use the default JZS (Jeffreys–Zellner–Siow) prior, which corresponds to a Cauchy distribution centered at 0 with a scale parameter of 0.707, as is standard in Bayes Factor applications.

Pre-Registered Exploratory Analysis. As an additional analysis, we will examine the within-country effects for the four primary outcomes using multilevel models with a random intercept per country and a random slope for the treatment assignment varying by country. For these models, we will use the same lasso-selected covariates in the primary

hypotheses specification. We will also examine whether country-level characteristics – such as income, inequality, and democracy – moderate the effects of social media reduction on each of the four outcome variables, beyond the theoretically specified relationships outlined in ***Table 1***. We will also report robustness checks for moderation, examining results using both binary and continuous moderators. Additionally, we will test moderation (M1, M2, and M4) using an individual-level proxy for socio-economic status. These are additional tests for which we have no strong theoretical expectations of moderation, so we do not pre-register any directional hypotheses and consider the results exploratory.

Compliance Average Causal Effects. As an additional analysis reported in supplementary materials, we will also run Complier-Average Causal Effects (CACE) using the two-stage least squares regression approach for our primary hypotheses[62]. As before, we will use the same set of covariates selected via lasso in the primary hypotheses specification. In section ***Supplementary Appendix D.3.6***, we propose an extensive procedure to measure compliance and detect fraud in the screenshot submission. We will report CACEs estimates for various levels of compliance (e.g., those who reduced their screentime, those who reduced their screentime and passed all screenshot tampering checks, etc.).

Differential Attrition. We will test for differential attrition using two procedures: i) using a *t-test* of the null hypothesis that the attrition rate is equal in treatment and control, and ii) by using an attrition bias test[66] to evaluate bias using the difference in the distribution of baseline outcomes between attriters and non attriters. If either of these tests indicates the presence of significant differential attrition, we will use treatment-effect bounds[67] and report results for all our primary treatment effects with and without treatment-effect bounds.

Baseline Balance. We will also report differences in baseline demographics and social media usage between the treatment and control groups, as well as between participants who dropped out and those who completed the post-treatment survey. We report *p*-values for equality in each of the individual comparisons, as well as the *p*-value for the *F*-test of joint significance of all differences.

Baseline Usage and Platform Differences. To examine whether the treatment is different for heavy social media users, we will examine treatment effect heterogeneity based on baseline usage of the four key social media apps. To test whether effects are driven by any platforms in particular, we will compute effect sizes separately for frequent users of each platform. Specifically, we will create two variables: 1) the platform (of the four key platforms included in this study) participants reported using most at baseline (pre-treatment), and 2) the

platform on which we observed the greatest change in usage among participants during the treatment period. We will plot the four main outcomes across each of these variables in the supplementary materials, and we will also test if each of these variables interacts with the treatment's effects on the four key variables. For these analyses, we do not pre-register any directional hypotheses and consider the results exploratory.

Demand Effects: To examine for the presence of experimenter demand effects, we propose to follow a strategy similar to prior Facebook deactivation studies[20]. In the post-treatment survey, we will include a multiple-choice question asking whether participants believe the researchers had an agenda. We will also ask whether participants adjusted their response in any way given what they thought the researchers were looking for. In the supplemental materials, we will present three analyses based on their responses: 1) the share of participants who believed the researchers had an agenda; 2) the difference in this belief between treatment and control (the key identification condition for experimenter demand bias); and 3) treatment effects excluding participants who both believed researchers had an agenda and self-reported altering their responses.

Pilot Experiments

To ensure the feasibility of our approach, we conducted three pilot experiments using the proposed method for our research (total $n = 894$) in the United Kingdom (June-July 2024, $n = 188$), the United States (October-November 2024, $n = 401$), and Brazil (January-February 2025, $n = 305$). Participants in the UK and US were recruited via the survey company Bilendi, and participants in Brazil were recruited via the survey company NetQuest. Details of the pilot experiments are in ***Supplementary Appendix Section D***.

These pilot experiments informed both the research design and analysis and provide evidence of the feasibility of this multi-country experiment. They allowed us to test our main outcome variables in a multi-country context and confirmed that the survey companies could successfully recruit relevant samples. The pilots also demonstrated that key design features (e.g., differential payments for the experimental vs. control conditions) did not substantially impact attrition (***Supplementary Appendix Section D.2).*** They confirmed that the social media reduction method successfully reduced screen time (***Supplementary Appendix Section D.3)*** and that the screenshot method was successful in checking compliance. We also validated our approach to using AI classification to analyze screenshot data (***Supplementary Appendix Section E).*** In addition, we conducted preliminary intent-to-treat analyses using the

pilot data (***Supplementary Appendix Section D.4)*** and used the pilot data to conduct a well-grounded power analysis (***Supplementary Appendix Section D.5)***.

The pilot experiments provide a strong foundation for rolling out the global field experiment across three critical areas. First, we confirmed that Bilendi can meet our recruitment goals and refined our procedures accordingly. Second, we confirmed that participants complied with the treatment, and that our approach effectively detected reductions in screen time. Third, we confirmed that the process does not induce differential attrition between treatment and control groups, specifically testing the role of differential payments.

Power Analysis. We used these pilot studies to conduct a power analysis (see ***Supplementary Appendix Section D.5***). For our primary hypothesis, our expected sample size can detect a minimal effect size of approximately 0.10 standard deviations with 95% power and two-sided tests with adjusted p-values < 0.05 (***Supplementary Appendix Figure 11***). Our minimum detectable effect size for the secondary moderation hypotheses is approximately 0.20 SD, with 95% of power, and two-sided tests with adjusted p-values < 0.05 (***Supplementary Appendix Figure 12***).

Preliminary Results. We also used the three pilot studies to present initial results as we intend to report in the main manuscript (see ***Supplementary Appendix Section D.4***). Our initial pilot results indicated that the social media reduction (H1) did not significantly impact news knowledge, $\beta = 0.07$, 95% CI = [-0.05, 0.20], $t = 1.09$, $p = 0.275$. It did, however, (H2) significantly reduce exposure to online hostility, $\beta = -0.21$, 95% CI = [-0.36, -0.069], $t = -2.88$, $p = 0.004$. It did not significantly impact (H3) party affective polarization, $\beta = -0.05$, 95% CI = [-0.18, 0.07], $t = -0.78$, $p = 0.435$, or social group polarization, $\beta = 0.09$, 95% CI = [-0.03, 0.22], $t = 1.44$, $p = 0.149$. It did (H4) significantly improve subjective well-being, $\beta = 0.15$, 95% CI = [0.019, 0.28], $t = 2.25$, $p = 0.024$. These preliminary findings go in the expected direction for online hostility (H2), partisan affective polarization (H3), and subjective well-being (H4). Meanwhile, the preliminary findings for news knowledge (H1) and social group polarization (H3) go in the opposite direction of the pre-registered hypotheses. However, the pilot results have substantially less power than the anticipated final sample size.

Data availability

All raw data and code will be shared upon Stage 2 acceptance. It will be made available at the following link: https://osf.io/xne6k/?view_only=412204c27a56411d8a5bae01c8c9d8e2. Pilot data and analysis code is currently available on our OSF.

Code availability

Our pilot analysis scripts are available at the following link: https://osf.io/xne6k/?view_only=412204c27a56411d8a5bae01c8c9d8e2

References

1. Number of worldwide social network users 2028. *Statista* https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ (2025).
2. Daily time spent on social networking by internet users worldwide from 2012 to 2025. *Statista* https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/ (2025).
3. Newman, N., Fletcher, R., Schulz, A., Andi, S. & Nielsen, R. *Reuters Institute Digital News Report 2020*. 111 https://www.digitalnewsreport.org/2020/ (2020).
4. Van Bavel, J. J. *et al.* Political psychology in the digital (mis) information age: A model of news belief and sharing. *Soc. Issues Policy Rev.* **15**, 84–113 (2021).
5. Van Der Linden, S. Misinformation: susceptibility, spread, and interventions to immunize the public. *Nat. Med.* **28**, 460–467 (2022).
6. Lazer, D. M. J. *et al.* The science of fake news. *Science* **359**, 1094–1096 (2018).
7. Van Bavel, J. J., Rathje, S., Harris, E., Robertson, C. & Sternisko, A. How social media shapes polarization. *Trends Cogn. Sci.* (2021).
8. Bail, C. *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. (Princeton University Press, Princeton, 2022).
9. Persily, N. & Tucker, J. A. *Social Media and Democracy: The State of the Field, Prospects for Reform*. (Cambridge University Press, 2020).
10. Kubin, E. & von Sikorski, C. The role of (social) media in political polarization: a systematic review. *Ann. Int. Commun. Assoc.* **45**, 188–206 (2021).
11. Kross, E. *et al.* Social media and well-being: Pitfalls, progress, and next steps. *Trends Cogn. Sci.* **25**, 55–66 (2021).
12. Orben, A. Teenagers, screens and social media: a narrative review of reviews and key studies. *Soc. Psychiatry Psychiatr. Epidemiol.* **55**, 407–414 (2020).
13. Twenge, J. M. Increases in Depression, Self‐Harm, and Suicide Among U.S. Adolescents After 2012 and Links to Technology Use: Possible Mechanisms. *Psychiatr. Res. Clin. Pract.* **2**, 19–25 (2020).
14. Braghieri, L., Levy, R. & Makarin, A. Social media and mental health. *Am. Econ. Rev.* **112**, 3660–3693 (2022).
15. Ghai, S., Fassi, L., Awadh, F. & Orben, A. Lack of sample diversity in research on adolescent depression and social media use: A scoping review and meta-analysis. *Clin.*

Psychol. Sci. 21677026221114859 (2021).

16. Ghai, S., Magis-Weinberg, L., Stoilova, M., Livingstone, S. & Orben, A. Social media and adolescent well-being in the Global South. *Curr. Opin. Psychol.* **46**, 101318 (2022).
17. Statista. Facebook users by country 2023 | Statista. https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/ (2023).
18. Lorenz-Spreen, P., Oswald, L., Lewandowsky, S. & Hertwig, R. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nat. Hum. Behav.* **7**, 74–101 (2023).
19. Wike, R. *et al. Social Media Seen as Mostly Good for Democracy Across Many Nations, But US Is a Major Outlier*. https://pewrsr.ch/3P7eQ5X (2022).
20. Allcott, H., Braghieri, L., Eichmeyer, S. & Gentzkow, M. The Welfare Effects of Social Media. *Am. Econ. Rev.* **110**, 629–676 (2020).
21. Asimovic, N., Nagler, J., Bonneau, R. & Tucker, J. A. Testing the effects of Facebook usage in an ethnically polarized setting. *Proc. Natl. Acad. Sci.* **118**, (2021).
22. Asimovic, N., Nagler, J. & Tucker, J. A. Replicating the effects of Facebook deactivation in an ethnically polarized setting. *Res. Polit.* **10**, 20531680231205157 (2023).
23. Hanley, S. M., Watt, S. E. & Coventry, W. Taking a break: The effect of taking a vacation from Facebook and Instagram on subjective well-being. *Plos One* **14**, e0217743 (2019).
24. Lambert, J., Barnstable, G., Minter, E., Cooper, J. & McEwan, D. Taking a one-week break from social media improves well-being, depression, and anxiety: a randomized controlled trial. *Cyberpsychology Behav. Soc. Netw.* **25**, 287–293 (2022).
25. Allcott, H., Gentzkow, M. & Song, L. Digital addiction. *Am. Econ. Rev.* **112**, 2424–63 (2022).
26. Thai, H. *et al.* Reducing social media use improves appearance and weight esteem in youth with emotional distress. *Psychol. Pop. Media* **13**, 162–169 (2024).
27. Ventura, T., Majumdar, R., Nagler, J. & Tucker, J. A. WhatsApp Increases Exposure to False Rumors but has Limited Effects on Beliefs and Polarization: Evidence from a Multimedia-Constrained Deactivation. *Available SSRN 4457400* (2023).
28. Allcott, H. *et al.* The effects of Facebook and Instagram on the 2020 election: A deactivation experiment. *Proc. Natl. Acad. Sci.* **121**, e2321584121 (2024).
29. Arceneaux, K., Foucault, M., Giannelos, K., Ladd, J. & Zengin, C. Facebook increases political knowledge, reduces well-being and informational treatments do little to help. *R. Soc. Open Sci.* **11**, 240280 (2024).

30. Ferguson, C. J. Do social media experiments prove a link with mental health: A methodological and meta-analytic review. *Psychol. Pop. Media* **14**, 201–206 (2025).
31. Ramadhan, R. N. *et al.* Impacts of digital social media detox for mental health: A systematic review and meta-analysis. *Narra J* **4**, e786–e786 (2024).
32. Thrul, J. *et al.* Social media reduction or abstinence interventions are providing mental health benefits—Reanalysis of a published meta-analysis. *Psychol. Pop. Media* **14**, 207–209 (2025).
33. Lemahieu, L. *et al.* The effects of social media abstinence on affective well-being and life satisfaction: a systematic review and meta-analysis. *Sci. Rep.* **15**, 7581 (2025).
34. Biggest social media platforms by users 2025. *Statista* https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/.
35. Vuorre, M. & Przybylski, A. K. Estimating the association between Facebook adoption and well-being in 72 countries. *R. Soc. Open Sci.* **10**, 221451 (2023).
36. Milkman, K. L. *et al.* Megastudies improve the impact of applied behavioural science. *Nature* **600**, 478–483 (2021).
37. Coles, N. A., Hamlin, J. K., Sullivan, L. L., Parker, T. H. & Altschul, D. Build up big-team science. *Nature* **601**, 505–507 (2022).
38. Van Bavel, J. J. *et al.* National identity predicts public health support during a global pandemic. *Nat. Commun.* **13**, 1–14 (2022).
39. Vlasceanu, M. *et al.* Addressing climate change with behavioral science: A global intervention tournament in 63 countries. *Sci. Adv.* **10**, eadj5778 (2024).
40. Rathje, S. *et al.* Unfollowing hyperpartisan social media influencers durably reduces out-party animosity. Preprint at https://doi.org/10.31234/osf.io/acbwg (2024).
41. Frimer, J. A. *et al.* Incivility Is Rising Among American Politicians on Twitter. *Soc. Psychol. Personal. Sci.* **14**, 259–269 (2023).
42. Rathje, S., Van Bavel, J. J. & Van Der Linden, S. Out-group animosity drives engagement on social media. *Proc. Natl. Acad. Sci.* **118**, e2024292118 (2021).
43. Sunstein, C. R. *# Republic: Divided Democracy in the Age of Social Media*. (Princeton University Press, 2018).
44. González-Bailón, S. *et al.* Asymmetric ideological segregation in exposure to political news on Facebook. *Science* **381**, 392–398 (2023).
45. Settle, J. E. *Frenemies: How Social Media Polarizes America*. (Cambridge University Press, Cambridge, 2018). doi:10.1017/9781108560573.

46. Kowal, M. *et al.* Predictors of enhancing human physical attractiveness: Data from 93 countries. *Evol. Hum. Behav.* **43**, 455–474 (2022).

47. Blanchflower, D. G. & Bryson, A. The Mental Health of the Young in Africa. Working Paper at https://doi.org/10.3386/w33280 (2024).

48. Ruscio, A. M. *et al.* Cross-sectional Comparison of the Epidemiology of DSM-5 Generalized Anxiety Disorder Across the Globe. *JAMA Psychiatry* **74**, 465–475 (2017).

49. Newman, N., Fletcher, R., Robertson, C. T., Ross Arguedas, A. & Nielsen, R. K. *Reuters Institute Digital News Report 2024*. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2024 (2024) doi:10.60625/RISJ-VY6N-4V57.

50. Poushter, J., Bishop, C. & Chwe, H. *Social Media Use Continues to Rise in Developing Countries but Plateaus Across Developed Ones*. https://www.pewresearch.org/global/2018/06/19/social-media-use-continues-to-rise-in-developing-countries-but-plateaus-across-developed-ones/ (2018).

51. David, C. C., Pascual, M. R. S. S. & Torres, M. E. S. Reliance on Facebook for news and its influence on political engagement. *PLOS ONE* **14**, e0212263 (2019).

52. Amsalem, E. & Zoizner, A. Do people learn about politics on social media? A meta-analysis of 76 studies. *J. Commun.* **73**, 3–13 (2023).

53. Stroud, N. J., Scacco, J. M. & Kim, Y. Passive learning and incidental exposure to news. *J. Commun.* **72**, 451–460 (2022).

54. Schäfer, S. Incidental news exposure in a digital media environment: a scoping review of recent research. *Ann. Int. Commun. Assoc.* **47**, 242–260 (2023).

55. Lorenz-Spreen, P., Oswald, L., Lewandowsky, S. & Hertwig, R. A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nat. Hum. Behav.* **7**, 74–101 (2023).

56. Yuen, S. & Lee, F. L. F. Echoes of silence: how social influence fosters self-censorship under democratic backsliding. *Democratization* **32**, 1213–1238 (2025).

57. Bor, A., Marie, A., Pradella, L. & Petersen, M. B. Undemocratic and unequal countries experience more political hostility on social media – Evidence from 30 Countries. Preprint at https://doi.org/10.31234/osf.io/spkyz (2024).

58. Statista. U.S. mobile social media traffic share 2019. *Statista* https://www.statista.com/statistics/477368/us-social-media-visits-share/ (2023).

59. Ventura, T., Majumdar, R., Nagler, J. & Tucker, J. Misinformation Beyond Traditional Feeds: Evidence from a WhatsApp Deactivation Experiment in Brazil. *J. Polit.* (2025)

doi:10.1086/737172.

60. Rathje, S., Robertson, C., Brady, W. J. & Van Bavel, J. J. People think that social media platforms do (but should not) amplify divisive content. *Perspect. Psychol. Sci.* (2023).
61. Bode, L. Political News in the News Feed: Learning Politics from Social Media. *Mass Commun. Soc.* **19**, 24–48 (2016).
62. Druckman, J. N. & Levendusky, M. S. What Do We Measure When We Measure Affective Polarization? *Public Opin. Q.* **83**, 114–122 (2019).
63. Allcott, H. *et al.* The Effect of Deactivating Facebook and Instagram on Users' Emotional State. Working Paper at https://doi.org/10.3386/w33697 (2025).
64. Bursztyn, L., Handel, B. R., Jimenez, R. & Roth, C. When Product Markets Become Collective Traps: The Case of Social Media. Working Paper at https://doi.org/10.3386/w31771 (2023).
65. Bak-Coleman, J. B. *et al.* Moving towards informative and actionable social media research. Preprint at https://doi.org/10.48550/arXiv.2505.09254 (2025).
66. Ghanem, D., Hirshleifer, S. & Ortiz-Becerra, K. Testing Attrition Bias in Field Experiments. *J. Hum. Resour.* 0920-11190R2 (2023) doi:10.3368/jhr.0920-11190R2.
67. Lee, D. S. Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. Working Paper at https://doi.org/10.3386/w11721 (2005).

Acknowledgements

Funding for this project was obtained from the National Science Foundation grant awarded to JVB and JT (#2334148), a National Science Foundation SBE Postdoctoral Fellowship awarded to SR (Grant #2404649), an AXA postdoctoral fellowship awarded to SR, a Templeton World Charity Foundation grant awarded to JVB (TWCF-2023-31570, https://doi.org/10.54224/31570), a Russell Sage Foundation grant awarded to SR and JVB (G-2110-33990), an NYU Seed Grant awarded to SR and JVB.

Author contributions

S.R., N.A., T.V., S.M., C.E.R., C.B., J.A.T, and J.V.B contributed to conceptualization, investigation, methodology, project administration, data curation & validation, and supervision. S.R., N.A., C.E.R., J.A.T., and J.V.B. contributed to funding acquisition. S.R., N.A., T.V., C.E.R., and C.B. contributed to formal analysis and visualization. S.R., N.A., T.V., S.M., C.E.R., and C.B. contributed to writing (original draft), and J.A.T and J.V.B contributed to writing (review and editing). All author contributions (outside of the core authorship team) are acknowledged in ***Supplementary Appendix F.4***.

Competing interests

J.V.B. has received funding from Google Jigsaw, has consulted for Microsoft News (MSN), and has participated in expert testimony for the state of New Mexico in a case related to Meta. S.R. and C.E.R. provided consulting for the state of New Mexico regarding a case related to Meta. J.A.T. received a small fee from Facebook to compensate him for administrative time spent in organizing a 1-day conference for approximately 30 academic researchers and a dozen Facebook product managers and data scientists that was held at NYU in the summer of 2017 to discuss research related to civic engagement. J.A.T. is also one of the co-leads of the external academic team for the 2020 U.S. Facebook & Instagram Election Study, a project that began in early 2020 and is still ongoing at the time of the submission of this registered report. He was not compensated financially for his participation in this project by Meta, but the project involves working collaboratively with Meta researchers. J.A.T. received a 2024 Google Research Grant to support a research project on “From Search Engines to Answer Engines: Testing the Effects of Traditional and LLM-Based Search on Belief in the Veracity of News.”