Social media advertising loads as prices*

George Beknazar-Yuzbashev[†] **Rafael Jiménez-Durán**[‡]
Andrey Simonov[§] **Mateusz Stalinski**[¶]

July 12, 2026

Abstract

Most social media platforms are funded through advertising rather than direct payments. Why? We argue that three main factors could explain this prevalence: users are more sensitive to monetary prices than to ad loads, microtargeting improves the match quality between users and ads, and platforms can personalize ad loads and thus price discriminate. We conduct a field experiment on Facebook with 1,810 users who install a browser extension that (i) hides nearly all ads or (ii) replaces microtargeted ads with untargeted ones. We find that hiding 82% of ads increases time on the platform by only 6%, showing that users are highly insensitive to ad loads. Removing targeting sharply reduces ad clicks and long-run engagement, indicating that targeting increases the match quality between users and ads. Finally, roughly two-thirds of ad-load variation occurs across users, consistent with ad-load discrimination. Counterfactual simulations indicate that an ad-funded model performs better than a subscription model in terms of profits and consumer surplus. The key mechanisms are that microtargeted ads are engaging and that users are much less sensitive to ad loads than to monetary prices, making advertising a relatively efficient revenue source.

Keywords: social media platfo rms, online advertisement, user engagement, field experiment

*This research is funded by the University of Chicago's Becker Friedman Institute, Columbia Business School's Digital Future Initiative, Columbia Business School's Jerome A. Chazen Institute for Global Business, George Mason University Antonin Scalia Law School, Law & Economics Center's Program on Economics & Privacy, and Marketing Science Institute. We thank Guy Aridor, Luca Braghieri, Jacob Conway, Jean-Pierre Dubé, Soheil Ghili, Ali Goli, Avi Goldfarb, Yufeng Huang, Garrett Johnson, Ro'ee Levy, Tesary Lin, Andrea Prat, Matthew Ridley, Lena Song, Nils Wernerfelt, seminar participants at VQMS, Economics of Advertising Workshop, Marketing Science, UChicago IO+ Conference, Stigler Center Affiliate Fellows conference, HKUST, Mannheim, MIT Sloan, UCLA Anderson, BSE Summer Forum: Applied IO, and SICS. The study was approved by the Columbia University Institutional Review Board (IRB-AAAU8500). The project is logged as externally approved by the Humanities and Social Sciences Research Ethics Committee at the University of Warwick under the reference number HSSREC 264/23-24. It was pre-registered in the American Economic Association Registry for randomized controlled trials under trial number AEARCTR-0014227.

[†]University of Chicago, beknazar@uchicago.edu.

[‡]Bocconi University, IGIER, Stigler Center, CESifo, and CEPR, rafael.jimenez@unibocconi.it.

[§]Columbia University and CEPR, as5443@gsb.columbia.edu.

[¶]University of Warwick and CAGE, mateusz.stalinski@warwick.ac.uk.

1 Introduction

Advertising-funded social media platforms have been criticized for maximizing engagement at the expense of content quality, contributing to the spread of misinformation, clickbait, and divisive content (Liu et al. 2022; Acemoglu et al. 2024; Beknazar-Yuzbashev et al. 2024). Subscriptions are often proposed as a remedy, yet advertising has dominated since the earliest social platforms (e.g., SixDegrees.com in 1997) and still accounts for over 97% of Meta's revenue—even as technology has made direct payments viable.[1] Why does the ad-funded model remain so prevalent? This question has become increasingly urgent as regulations—such as the EU's Digital Markets Act and Digital Services Act—and privacy-related product changes constrain platforms' ability to advertise and microtarget.

This paper investigates why advertising dominates social media monetization and what happens when this model is restricted. Treating ads as implicit prices that users pay in attention rather than money (Anderson and Coate 2005; Calvano and Polo 2021), we suggest three channels that may explain advertising's prevalence. First, users may be less sensitive to attention prices than to monetary prices, making advertising a relatively efficient way to extract surplus. Second, microtargeting may improve the match between users and ads—a channel that depends on data practices increasingly subject to privacy regulation. Third, platforms can personalize ad loads across users, a form of price discrimination that may be harder to implement with monetary prices.

We first develop a modeling framework to formalize these channels. Users allocate time to a platform that monetizes through advertising, and the platform can personalize both ad loads and ad targeting based on observable user characteristics. The model shows that optimal ad loads follow an inverse-elasticity logic: users less sensitive to ads face higher loads. Network effects also matter: users whose participation generates larger spillovers receive lower ad loads, as the platform protects their engagement. Advertiser value provides a third force: users who are more valuable to advertisers receive higher ad loads. Finally, we use a decomposition to compare ad-based and subscription-based profits, isolating the contributions of microtargeting, ad-load personalization, and differential demand sensitivity to attention versus monetary prices.

To estimate the key model parameters, we conduct a field experiment with 1,810 U.S.

[1] See https://www.sec.gov/Archives/edgar/data/1326801/000132680125000017/meta-20241231.htm, accessed February 04, 2026.

Facebook users who install a custom desktop browser extension. The extension passively collects platform activity and browsing data for six weeks. Users are then randomly assigned to one of two interventions for six more weeks. In the first treatment arm, the extension hides most ads, reducing ad load to near zero. In the second arm, ads appear at the same frequency but are replaced with ads originally shown to other users, degrading match quality while holding quantity fixed. Each intervention has a dedicated control group, with the *Replace Control* replacing the ad with itself.[2] Comparing outcomes across treatment and control groups before and after randomization identifies the causal effects of changing ad load and ad targeting.

We measure four categories of outcomes: ad and platform engagement on Facebook, time spent on competing platforms, and platform valuation. Ad engagement is measured through ad impressions, ad clicks, and a summary index of the two; platform engagement is measured through active time, the volume of posts and comments consumed, and an analogous index. To capture potential substitution, we track time spent on competing platforms, including X and Reddit. For platform valuation, we elicit users' willingness to accept (WTA) to deactivate Facebook for four weeks (Allcott et al. 2020), surveyed at both baseline and endline. The endline survey was administered at the end of the six-week intervention period. Subsequently, the extension continued the intervention for approximately thirteen additional weeks, allowing us to assess whether treatment effects persist.

Our first result is that social media feed ad loads—the ratio of ad content to total platform content in the main feed—are highly heterogeneous across users even in the absence of intervention, consistent with systematic ad-load discrimination. Averaged at the user level, untreated feed ad loads on Facebook were approximately 13% (roughly one ad per eight pieces of content). We document substantial underlying heterogeneity: the cross-user standard deviation is 4.6 percentage points, roughly one-third of the mean. Indeed, 66% of user-day variation in feed ad load on Facebook is attributable to persistent between-user differences, suggesting that platforms systematically charge different users different "attention prices." This cross-user variation correlates with observable demographics—most notably age, desktop usage share, and political affiliation. To assess whether ad-load discrimination is specific to Facebook or a broader feature of social media monetization, we also examine

[2]This design isolates the effect of targeting from the potential impact of the replacement arm's technical implementation. For example, because the possibility to comment was disabled on replaced ads, comparisons between *Replace* and *Hide Control*—where commenting remained enabled—could confound targeting changes with functionality changes. *Replace Control* is designed to avoid such contamination. See Section 3 for details.

ad loads on X and Reddit, finding similar averages (11% and 10%) and similar between-user variance shares (65% and 71%). Moreover, individual ad loads are positively correlated across Facebook, X, and Reddit, consistent with different platforms tailoring ad loads to similar user characteristics.

Turning to the experiment, we find that users are highly inelastic with respect to ad loads. The hiding intervention reduced experienced ad loads by 82%, yet content consumption increased by only 9% and active time by only 6% over the six-week intervention period. The preregistered platform engagement index rose by just 0.03 standard deviations (SD). These modest responses imply an elasticity below 0.1. Moreover, the effects remain stable over nineteen weeks of continued intervention (0.02 SD), suggesting muted long-run responses.

Our replacement intervention degraded ad targeting by substituting users' algorithmically-selected ads with ads randomly drawn from those shown to other users: a measure of distance between user characteristics and the ads displayed to them increased by 0.5 SD. This intervention reduced ad engagement: the index fell by 0.15 SD, driven primarily by a 0.25 SD decline in ad clicks. Even more strikingly, the intervention reduced platform engagement as well—in the long run, active time and content consumption both fell by 13%, and the platform engagement index declined by 0.03 SD. Thus, degrading ad targeting decreases both ad engagement and platform engagement.

We report several supplementary results and robustness checks. First, Facebook's algorithm showed no response to the hiding intervention—ad loads supplied by the platform remained unchanged—but responded mildly to the replacement intervention by supplying less-targeted ads. This asymmetry suggests the platform treats users' ad-load sensitivity as stable while assuming that preferences for ad content can evolve. Second, we find little evidence of substitution toward competing platforms, consistent with Aridor (2025)—even large changes in ad load or targeting did not drive users to competitors. Third, neither intervention affected willingness to accept payment to deactivate Facebook, our proxy for platform valuation. Finally, our results are robust to differential attrition: survival at the end of the six-week intervention period exceeds 75% across all treatment arms and does not vary by assignment. Attrition remains non-differential after nineteen weeks of continued intervention, with survival above 50% in all groups. Although our sample consists of desktop users who responded to a Facebook recruitment ad, our central inelasticity estimate matches the full-population experiment of Brynjolfsson et al. (2025)—which covers both mobile and desktop and does not recruit through ads—suggesting that these sampling restrictions do

not drive our conclusions.

What mechanisms explain such muted responses to ad load? In a separate, preregistered data collection, we recruited 215 U.S. Facebook users to browse their own feeds while the extension recorded how they attended to each ad. The evidence points to low-cost ad avoidance rather than equal enjoyment of ads and organic content. Users devote about 10–11% less viewport time to ads than to organic posts and rate ads as less interesting and less valuable, yet they still correctly recognize roughly 51% of the ads from their own session (against 7% for placebo ads they never saw). Users thus notice ads but scroll past them at little cost, which helps explain why removing ads changes engagement so little.

Next, we use our experimental results to estimate a structural model of engagement that allows us to simulate counterfactual monetization regimes. We compare outcomes under three scenarios: the observed advertising-based model with personalized ad loads and microtargeting, a subscription-based model with uniform monetary pricing and no ads, and a hybrid "opt-in" alternative where users choose between an ad-supported experience and a paid ad-free tier. For each regime, we evaluate profits, user engagement, platform demand, and consumer surplus. To identify which mechanisms drive the differences across regimes, we decompose performance gaps into four components: (i) the loss of ad-user match quality from eliminating ad targeting, (ii) the loss of ad price differentiation that ad targeting enables on the advertiser side, (iii) the loss of ad-load personalization across users, and (iv) the direct effect of switching from attention-based to monetary pricing.

The counterfactual simulations indicate that the ad-funded model generates higher profits and consumer surplus compared to a subscription-based alternative. Two key drivers of these patterns are that microtargeted ads are engaging and that users respond far more strongly to monetary prices than to advertising intensity. Even the hybrid opt-in model—similar to the paid ad-free tier Meta introduced in Europe under the Digital Markets Act—fails to outperform pure advertising: only around 1% of users in our simulations choose the ad-free subscription option, leaving the aggregate outcomes largely unchanged.

Taken together, these results help explain why advertising has remained the dominant revenue model for social media platforms despite the technological feasibility of alternatives. The persistence of this business model does not appear to reflect consumer exploitation; rather, advertising functions as an efficient pricing mechanism when microtargeted ads are engaging and users are highly sensitive to monetary charges but tolerant of attention costs. These findings paint a nuanced picture of advertising's welfare effects: while previous research

has highlighted that engagement-maximization incentives can inadvertently lead platforms to prioritize "toxic" content (Beknazar-Yuzbashev et al. 2024; Beknazar-Yuzbashev et al. 2025), this paper shows that ad-based monetization is nonetheless a relatively efficient revenue source both for platforms and users—holding constant the quality of content on the platform.

Our paper builds on the theoretical insight that advertising intensity functions as an implicit "attention price" (e.g., Anderson and Coate 2005; Anderson and De Palma 2012; Anderson and Peitz 2023; Calvano and Polo 2020; see Calvano and Polo 2021 for a review). Empirically, a negative relationship between ad loads and consumption has been documented in traditional media (Wilbur 2008) and in digital settings such as sponsored search, streaming audio, and e-commerce (Moshary 2021; Goli et al. 2025a; Goli et al. 2025b). In the context of social media, Brynjolfsson et al. (2025) report results from a nine-year-long field experiment on Facebook and find an elasticity of time spent with respect to ad loads of 0.094, consistent with our finding of modest engagement responses.[3] We extend this literature in two ways: by experimentally varying not only ad quantity but also ad-user match quality, and by documenting systematic ad-load personalization across users consistent with price discrimination.

We also contribute to the growing literature on the economics of social media and platform monetization (Allcott et al. 2020; Aridor et al. 2024; Aridor et al. 2025b). This work highlights the impact of ad-driven incentives on the proliferation of harmful content (Liu et al. 2022; Acemoglu et al. 2024; Jimenez-Duran 2022; Beknazar-Yuzbashev et al. 2024). Ad load policies are central to competition-policy and antitrust questions related to social media (e.g. see Ambrus et al. 2016; Athey et al. 2018; Center 2019; Argentesi et al. 2021; Aridor 2025; Banchio et al. 2026). The most closely related paper is Katz and Allcott (2025), who develop a two-sided market model of digital media mergers that centers on advertiser demand, duplication inefficiencies, and equilibrium ad prices. Compared to Katz and Allcott (2025), we focus on the consumer side of the market, estimating user ad-load sensitivity and the engagement effects of microtargeting. Our counterfactual analysis of alternative ad load designs and monetization regimes speaks directly to ongoing regulatory scrutiny, including the European Commission's investigation of Meta under the Digital Markets Act (European Commission 2025). Our results help to rationalize why advertising-based business models

[3]Relatedly, Goodman et al. (2026) apply a Beckerian time-allocation model to the experiment in Brynjolfsson et al. (2025) and an additional experiment, and show that Facebook and Instagram demand is highly inelastic to ad load and that time costs and time shares shape diversion patterns across online and offline activities.

are so prevalent in the context of digital platforms.

Lastly, we contribute to the literature on the economics of privacy (Lin 2022; Goldfarb and Que 2023; Dubé et al. 2025). Prior work has shown that limiting tracking reduces advertising effectiveness and platform revenues (Goldfarb and Tucker 2011; Alcobendas et al. 2023; Aridor et al. 2023; Johnson 2024; Aridor et al. 2025a; Wernerfelt et al. 2025), but these estimates combine effects on both sides of the market. We isolate the consumer response by experimentally degrading ad targeting while holding the advertiser side fixed, finding that worse targeting reduces user engagement but does not affect platform valuation. This is consistent with the privacy paradox (Acquisti and Grossklags 2005): while users express growing privacy concerns across ever-broader contexts (Goldfarb and Tucker 2012), their revealed behavior suggests limited welfare costs from targeting.

The paper proceeds as follows. Section 2 presents a model. Section 3 presents our experimental design. Section 4 presents our results. Section 5 evaluates counterfactual business models and mechanisms. Section 6 concludes.

2 Model

We now present a model where users spend time on a platform. The platform can personalize both the quantity of ads it serves to the users and their average match quality, and can in principle charge users a monetary subscription fee. We use this model 1) to generate predictions of how a profit-maximizing platform would optimally set individual ad loads and 2) to understand the conditions that determine the profitability of an ad-based platform relative to the subscription-based platform. We consider these two extremes as a starting point.

Users. There is a set of individuals who choose how much time to spend on the platform to maximize their utility. Individuals derive utility from joining the platform, from the time they spend consuming content, and from the ads they are exposed to. They have observable types $\theta \in \Theta$, distributed according to the discrete probability mass function g^{θ}. We interpret θ as the set of characteristics that the platform can observe and condition on when personalizing ad loads or targeting. Within each observable type θ, there is a continuum of users indexed by an idiosyncratic type ε, which determines the "membership benefits" of joining the platform (and becoming a user). This idiosyncratic type is distributed according

to the conditional density $f_{\varepsilon|\theta}$.

Consider an individual of type θ who allocates time $e^\theta \geq 0$ to consume content, conditional on joining the platform. The individual takes as given the (potentially personalized) ad load denoted by $a^\theta \in [0,1]$, which denotes the amount of time spent consuming ads for each unit of time spent consuming content. Thus, the total amount of ads consumed is $A^\theta = a^\theta e^\theta$. The individual also takes as given the average ad-user match quality denoted by $q^\theta \in [0,1]$. The total amount of time spent on the platform, conditional on joining, is $e^\theta(1+a^\theta)$. Each unit of time spent carries an opportunity cost w^θ, reflecting either the value of time or the implicit cost of sharing personal data through engagement.

Conditional on joining the platform, the user's optimization problem, considering the utility from content consumption, ad consumption, and the opportunity cost of time, is:

$$v^\theta(a^\theta, q^\theta, x) = \max_{e \geq 0} \; u^{\theta,c}(e) + u^{\theta,a}(ea^\theta, q^\theta) + u^{\theta,j}(x) - w^\theta e(1+a^\theta),$$

where $u^{\theta,c}$ is the utility of content consumption. We assume that the marginal utility of time spent is positive, $u_e^{\theta,c} > 0$. The utility from consuming ads is $u^{\theta,a}$, which depends both on the quantity and match quality of ads consumed. The marginal utility from ads, represented by $u_A^{\theta,a}$, can be positive or negative, but we restrict it to be lower than the marginal utility of engagement: $u_A^{\theta,a} < u_e^{\theta,c}$, in other words, users prefer to spend time consuming content rather than consuming ads. The match quality of ads has a "vertical" nature, in the sense that it increases the marginal utility of ads, $u_{Aq}^{\theta,a} \geq 0$. However, users can still derive positive or negative utility from consuming ads with a better match quality; for example, reflecting that they might have some instrumental utility from privacy (Lin 2022), which would result in users experiencing a negative marginal utility from a better match quality.[4] The function $u^{\theta,j}(x)$ depends on the platform's market share, x, to capture direct network effects from other users' participation. Lastly, to ensure that second-order conditions hold, we assume that utilities are concave in e. We denote by $e^\theta(a^\theta, q^\theta)$ the solution to this optimization problem.

Let $a = (a^\theta)_{\theta\in\Theta}$ and $q = (q^\theta)_{\theta\in\Theta}$ denote the vectors of ad loads and qualities for all user types, respectively. Additionally, let s denote the "subscription" (monetary price) charged by the platform.[5] The demand for the platform of θ users is given by $X^\theta(a^\theta, q^\theta, s, x) =$

[4] We normalize $u^{\theta,a}(0,q) = 0$ for all q: when there are no ads, their match quality is irrelevant.

[5] We assume that the platform cannot price-discriminate through subscription fees, but can personalize ad loads. This reflects common practice in real-world platforms.

$g^\theta \int_{s-v^\theta(a^\theta,q^\theta,x)}^{\infty} f_{\varepsilon|\theta}(\varepsilon \mid \theta)d\varepsilon$, and total platform demand is $X(a,q,s,x) = \sum_\theta X^\theta(a^\theta, q^\theta, s, x)$. In equilibrium, this demand has to be consistent with expectations, $x = X(a,q,s,x)$. As standard in the network-effects literature, we assume that network effects are not too strong, such that, for any (a,q,s), there exists a unique solution $x(a,q,s)$ to the fixed-point problem.

Platform. Let p^θ denote the price (in dollars per unit of time) that advertisers pay per unit of engagement of θ users with their ads, and p denote the vector of ad prices for all users. The platform's profit function is then:

$$\pi(p,a,q,s) = \sum_\theta \left[\underbrace{p^\theta \times a^\theta \times e^\theta(a^\theta,q^\theta) \times x^\theta(a,q,s)}_{\text{Ad revenue}} + \underbrace{s \times x^\theta(a,q,s)}_{\text{Subscription revenue}} \right],$$

where $x^\theta(a,q,s) = X^\theta(a^\theta, q^\theta, s, x(a,q,s))$ is the equilibrium demand of θ users. We can allow ad prices p^θ to depend on the ad load and match quality vectors, to capture potential information overload (Anderson and De Palma 2012), that advertiser willingness to pay increases with a better match quality, or that the platform has market power in digital ads markets (Gentzkow et al. 2024).

The optimization problem of an ad-based platform is then:

$$\max_{a,q} \ \pi(p(a,q), a, q, s = 0),$$

with solution a^*, q^* and a maximum of $\pi(p^*, a^*, q^*, s = 0)$. Instead, the optimization problem of a subscription-based platform is:

$$\max_{s} \ \pi(p, a = 0, q, s),$$

with solution s^* and maximum of $\pi(p, a = 0, q, s^*)$.

Relative profitability. We start by decomposing the difference in profits between a subscription-based business model and an ad-based business model into four components. The decomposition is an accounting identity and does not rely on the assumptions on user

behavior above. For any scalars $\overline{a}$, $\overline{q}$, and $\overline{p}$:[6,7]

$$\begin{aligned}\pi(p, a=0, q, s^*) - \pi(p^*, a^*, q^*, s=0) = & \underbrace{\pi(p^*, a^*, \overline{q}, s=0) - \pi(p^*, a^*, q^*, s=0)}_{\text{1. Reduced ad match quality}} \\ & + \underbrace{\pi(\overline{p}, a^*, \overline{q}, s=0) - \pi(p^*, a^*, \overline{q}, s=0)}_{\text{2. No ad price differentiation}} \\ & + \underbrace{\pi(\overline{p}, \overline{a}, \overline{q}, s=0) - \pi(\overline{p}, a^*, \overline{q}, s=0)}_{\text{3. No ad load discrimination}} \\ & + \underbrace{\pi(\overline{p}, a=0, \overline{q}, s^*) - \pi(\overline{p}, \overline{a}, \overline{q}, s=0)}_{\text{4. Business model change}} \end{aligned} \tag{1}$$

The first two components correspond to the change in profits due to the elimination of microtargeting, all else equal. The first component speaks to the user side of the market: it is the difference in profits driven by the change in engagement when users face ads with a worse ad match quality. The second component speaks to the advertiser side of the market: it gives the change in profits when the platform cannot charge different prices for advertising to different segments of the market. The third component corresponds to the change in profits when the platform can no longer set different ad loads to different users, all else equal. The fourth component informs about the change in profits due to the change in business model from an ad-based to a subscription-based model.

Two objects are needed to calculate this decomposition: a counterfactual estimate of engagement, $e^\theta(a^\theta, q^\theta)$, and a counterfactual estimate of consumer surplus, $v^\theta(a^\theta, q^\theta, x)$. The goal of our experiment is to vary the ad load a^θ and ad microtargeting q^θ that individuals face to construct these counterfactuals.[8]

Optimal ad policy. We now consider the optimal ad load and optimal ad match quality that an ad-based platform should set. Rearranging the first-order condition with respect to

[6]Note that: $\pi(p, a=0, q, s^*) = \pi(\overline{p}, a=0, \overline{q}, s^*)$ since changes in ad prices and ad match quality do not change profits when there are no ads.

[7]Note that this decomposition can be used for other outcomes beyond profits, such as engagement or consumer surplus.

[8]Armed with these counterfactuals, we compute the platform's optimal subscription fee s^* in our sample. Additionally, our experiments do not produce estimates of network effects or how they impact consumer surplus, but we use estimates from the literature, particularly from Bursztyn et al. (2025).

a^θ, we obtain the optimal advertising load for users of type θ:

$$a^\theta = \frac{p^\theta e^\theta + \sum_{\tilde{\theta}} \frac{p^{\tilde{\theta}} a^{\tilde{\theta}} e^{\tilde{\theta}}}{x^\theta} \frac{\partial x^{\tilde{\theta}}}{\partial a^\theta}}{\left|\frac{\partial e^\theta}{\partial a^\theta}\right| p^\theta + \left|\frac{\partial p^\theta}{\partial a^\theta}\right| e^\theta}. \tag{2}$$

Equation (2) highlights three key forces behind optimal ad allocation:

1. **Ad load sensitivity.** Users whose engagement is less sensitive to ads—i.e., those with small $\left|\partial e^\theta / \partial a^\theta\right|$—receive higher ad loads. This condition mirrors the classic inverse-elasticity pricing logic (Pigou 1920).
2. **Network spillovers.** The term $\sum_{\tilde{\theta}} \frac{p^{\tilde{\theta}} a^{\tilde{\theta}} e^{\tilde{\theta}}}{x^\theta} \frac{\partial x^{\tilde{\theta}}}{\partial a^\theta}$ captures how ads on user θ affect the platform's revenue from other users. When θ's participation generates strong positive spillovers (the term is large and negative), the platform sets a lower ad load to protect that participation. Conversely, when these spillovers are weak or negative, the platform is more willing to tax θ's attention with ads. This parallels the logic of optimal pricing in networks (Candogan et al. 2012).
3. **Advertiser value.** Users who are more valuable to advertisers—those with a higher ad price p^θ—tend to receive higher ad loads. Formally, $\partial a^\theta / \partial p^\theta \geq 0$ if exposing user θ to more ads does not raise any user's participation.[9]

In terms of the optimality condition with respect to q^θ, our assumptions imply that the platform will reach a corner solution and maximize the ad match quality.

3 Experiment Design

3.1 Design Overview

We conduct a field experiment to estimate user engagement responses to changes in ad load and ad targeting on Facebook. Participants install a custom browser extension that passively

[9] To see this, differentiate (2) with respect to p^θ, holding fixed engagement e^θ, the sensitivities $\left|\partial e^\theta / \partial a^\theta\right|$ and $\left|\partial p^\theta / \partial a^\theta\right|$, and the spillover term $\Sigma^\theta \equiv \sum_{\tilde{\theta}} \frac{p^{\tilde{\theta}} a^{\tilde{\theta}} e^{\tilde{\theta}}}{x^\theta} \frac{\partial x^{\tilde{\theta}}}{\partial a^\theta}$. Writing the numerator of (2) as $N = p^\theta e^\theta + \Sigma^\theta$ and the denominator as $D = \left|\partial e^\theta / \partial a^\theta\right| p^\theta + \left|\partial p^\theta / \partial a^\theta\right| e^\theta$, we get

$$\frac{\partial a^\theta}{\partial p^\theta} = \frac{e^\theta D - \left|\partial e^\theta / \partial a^\theta\right| N}{D^2} = \frac{\left|\partial p^\theta / \partial a^\theta\right| \left(e^\theta\right)^2 - \left|\partial e^\theta / \partial a^\theta\right| \Sigma^\theta}{D^2},$$

which is nonnegative whenever $\Sigma^\theta \leq 0$—in particular, when $\partial x^{\tilde{\theta}} / \partial a^\theta \leq 0$ for all $\tilde{\theta}$, so that loading more ads on θ does not raise any user's participation. The inequality is strict unless $\left|\partial p^\theta / \partial a^\theta\right| = 0$ and $\Sigma^\theta = 0$: with constant ad prices and no participation spillovers, p^θ cancels from (2) and the ad load follows the inverse-elasticity rule $a^\theta = e^\theta / \left|\partial e^\theta / \partial a^\theta\right|$.

collects engagement data and allows us to modify the ads they see without altering organic content or the platform's ad-serving decisions (see Farronato et al. 2024 for a general-purpose implementation of this approach).

The experiment uses two treatment-control pairs rather than a single pooled control. The first pair isolates the effect of ad quantity: *Hide* removes most ads while *Hide Control* leaves the feed unchanged. The second pair isolates the effect of ad targeting: *Replace* substitutes users' ads with ads shown to other users while *Replace Control* applies the same technical procedure but preserves targeting. Having separate control groups ensures that comparisons are not confounded by implementation artifacts of the replacement technology.

Participants—recruited through Facebook ads—complete a six-week baseline period followed by an intervention period. We pre-registered a six-week intervention window for our main analysis, but the intervention remains active for approximately nineteen weeks total to assess persistence. Figure A1 summarizes the study flow.

3.2 Recruitment and Sample Construction

We recruit participants through Facebook ads (Figure A4 offers an example) between June 19, 2024 and December 24, 2024. Eligibility is restricted to adult U.S.-based users who use English as their primary language on Facebook and who primarily access the platform using Google Chrome. The recruitment advertisements direct interested users to a landing page that describes the study and links to the browser extension. Recruiting via Facebook ads ensures that enrollees do not use ad blockers on Facebook, since they would not have seen the recruitment ad otherwise. We describe the study as academic research on social media without disclosing any information about advertising interventions, minimizing the risk of self-selection based on ad preferences.

Enrollment requires installation of the browser extension and completion of both a baseline and an endline survey. Participants receive a $21 gift card upon completion. Our final sample includes 1,810 users who meet a minimum Facebook usage threshold during the baseline period, run a current version of the extension, and satisfy technical requirements detailed in Appendix D.

3.3 Treatments and Intervention Implementation

3.3.1 Hide Treatment

The *Hide* treatment reduces the experienced ad load to near zero by suppressing the visual display of ads after they are delivered to the browser (but before they are delivered to the user), without altering which ads are selected by the platform's ad auction or modifying any organic content. Figure A2 in the appendix illustrates the intervention.

The corresponding control group, *Hide Control*, experiences no modifications to the user experience. Thus, between these two groups, we exogenously vary the ad load—effectively altering the implicit price of Facebook consumption.

3.3.2 Replace Treatment

The *Replace* treatment eliminates microtargeting while holding ad quantity fixed. Ads selected for a user by the platform are replaced with ads originally shown to other users, severing the link between user characteristics and displayed advertising.

Replacement ads are drawn from a pool of previously collected English-language advertisements observed by the extension *across* the user base in the past few days. Ads are replaced on a one-for-one basis, so that the number and position of ads in the feed are identical to those in the control condition, ensuring that the visual structure and scrolling experience remain unchanged. Several constraints are imposed to ensure that treatment effects are not driven by artifacts of the replacement technology.[10] Certain ad formats—including video ads, carousel ads, and lead-generation forms—are excluded from replacement due to technical limitations (they are neither replaced nor used as replacements).

Figure A3 in the appendix shows an example of an ad that replaced one that Facebook intended to display. The left panel of the figure demonstrates an image of the replacement ad—it is formatted in the same way as any other Facebook ad. The right panel of the figure showcases possible interactions with the replacement ad. Users can interact naturally with replacement ads—liking, clicking through to the advertiser's website—though commenting and sharing are disabled.

The associated control group, *Replace Control*, undergoes the identical replacement pro-

[10] User interface elements are standardized across replaced ads, including a uniform call-to-action button ("Learn more"), and interactive features such as comments and shares are disabled for all replacement ads. In general, authors of Facebook posts can opt to disable commenting, so encountering ads with commenting disabled is not unusual. Moreover, it is rare for users to attempt commenting on or sharing ads.

cedure except that each ad is replaced with itself, so the only exogenous variation between groups is targeting quality.

3.3.3 Common Engineering Features

To ensure that comparisons across treatment arms are not confounded by differential processing loads, all advertisements in all groups are processed through identical detection and classification routines.[11]

Neither intervention modifies ads that are already visible on the user's screen (i.e., in the viewport) at the start of a Facebook session. This constraint arises because initial page loads occur concurrently with extension initialization, creating a risk that ads appear briefly before being modified. Keeping ads in the viewport at page load unchanged avoids abruptly altering content a user is actively viewing.[12] To prevent ads from appearing mid-scroll, the extension batches incoming feed content and holds each batch until processing is complete before releasing it to the user.[13]

3.3.4 Randomization

Randomization occurs immediately after completion of the baseline period and remains fixed for the duration of participation. Users are assigned to one of four experimental arms: *Hide, Hide Control, Replace*, or *Replace Control*. To ensure adequate sample size in the *Replace* arm despite an extension error affecting some participants (see Appendix C.2), we oversampled this condition during recruitment. Our empirical strategy accounts for the resulting unequal allocation probabilities using inverse probability weights (see Section 3.5).[14]

Participants enter the study on a rolling basis as they respond to recruitment advertisements. For the empirical specification, we group participants into cohorts defined by their intervention start date.

The resulting sample includes 413 users in *Hide*, 414 in *Hide Control*, 538 in *Replace*, and 445 in *Replace Control*. Tables B1 and B2 in the appendix present the covariate bal-

[11] For example, *Replace* goes through the routine of selecting a random ad (which includes verifying that the image of the ad is accessible), which increases latency. To ensure identical latency between treatment and control, we force the same action in *Replace Control*, even though the ad is ultimately discarded.

[12] In our dataset, 11% of Facebook content (posts, ads, or comments) are recorded as being in the viewport, meaning that at least a part of the element was visible on the screen at the point of rendering. The corresponding proportion for X is 15%.

[13] This happens off screen. The median time to screen for an element outside of the viewport was 4.87 seconds on Facebook, 1.72 seconds on X, and 29.7 seconds on Reddit.

[14] Results are robust to dropping users impacted by the error—see Appendix C.2.

ance between treatment and control conditions. In both cases, we assess balance on twelve covariates, including user demographics (age, sex, ethnicity, income, education, and political affiliation) and social media usage variables (number of Facebook friends and baseline engagement measures). We find no evidence of imbalance in either comparison—none of the covariates differs significantly by treatment at the 5% level and only one covariate is significant at the 10% level (in Table B2).

3.4 Data Collection and Measurement

3.4.1 Platform Activity and Ad Data

For Facebook, the extension captures the full content of the feed,[15] including organic posts, advertisements, and all observable user interactions. Ads are identified using platform-specific markers and structural features, allowing us to distinguish advertising content from organic content with high accuracy.

The extension also collects feed content and ad exposure data from Reddit[16] and X. The structural features used to identify ads differ by platform, but the procedures are analogous.

3.4.2 Ad Load Measures

We construct three measures of advertising intensity. The primary measure, *experienced ad load*, is defined as the number of advertisements *observed* by the user divided by the total volume of content (ads plus organic posts and comments) observed that day. This measure accounts for differences in consumers' typical usage patterns. For instance, it captures that consumers who spend more time in comment sections encounter fewer ads.

The second measure, *supplied ad load*, is defined analogously but uses content *served* by the platform before any intervention is applied, allowing us to test whether the platform's algorithm responds endogenously to our interventions.

The third measure, *feed ad load*, restricts the denominator to content appearing in the main news feed, excluding comment sections and other non-feed areas. Unlike experienced

[15] By "feed", here, we mean the central portion of the screen, which excludes any advertisement that appears on the side of the window in a static fashion.

[16] A change to Reddit's page structure in late October 2024 altered how advertisement elements were tagged, causing the extension to undercount Reddit ad impressions thereafter. We therefore restrict all Reddit ad-load measures to the window before this change. Facebook and X are unaffected. This shortens the Reddit observation window relative to the other platforms but leaves the cross-sectional ad-load patterns we report intact.

ad load, this measures what the platform shows in the feed, abstracting from user behavior. For X and Reddit, feed content is directly identifiable. For Facebook, we classify an element as feed content unless it appears in a sequence of four or more consecutive comments.[17]

3.4.3 Ad Match Quality

To measure how well ads match individual users, we construct an ad distance metric that captures the demographic gap between a user and the typical audience an advertiser reaches. The key idea is that a well-targeted ad should be shown to users whose characteristics align closely with the advertiser's revealed targeting profile; a poorly targeted ad will exhibit a large discrepancy.

Specifically, let $\mathcal{K}$ denote the set of five binary demographic variables: young (age at or below the median), male, white, bachelor's degree or higher, and household income above \$50,000. For each advertiser a and trait $k \in \mathcal{K}$, we define $\bar{X}_{ak}$ as the share of that advertiser's impressions delivered to users with trait k, computed using baseline-period data only (Period ≤ 0) to avoid contamination by the intervention. The distance for a given user-ad pair is then

$$d_{ia} = \sqrt{\sum_{k \in \mathcal{K}} (X_{ik} - \bar{X}_{ak})^2},$$

where $X_{ik} \in \{0, 1\}$ is user i's value for demographic trait k. A small value of d_{ia} indicates that user i's demographic profile closely matches the advertiser's typical audience, while a large value signals a poor match. We aggregate to the user-day level by averaging d_{ia} across all ads shown to user i on a given day, and normalize the resulting series by its baseline-period mean, so that a value of one corresponds to the average baseline distance and coefficients can be read as proportional changes.

3.4.4 Engagement and Time Measures

We measure engagement along two dimensions: time spent on platforms and content consumed. For Facebook, X, and Reddit, we construct an active time measure following Beknazar-Yuzbashev et al. (2025). Active time captures periods during which the user is actively interacting with the platform—scrolling, clicking, or otherwise engaging with the feed—while excluding idle time when the browser tab remains open but the user has stepped

[17] Comments in the feed do not occur in groups larger than three. A limitation is that posts with 1–3 comments are misclassified as feed elements, slightly overestimating feed content.

away. To avoid truncating sessions prematurely, we append a three-minute window following the last observed interaction, accounting for brief pauses (e.g., reading a long post) that plausibly reflect continued attention.

For other websites, we employ a coarser measure: the total daily duration the browser was open on a given domain, regardless of whether the user was actively interacting. Approximately every minute, the extension records the domain visible in the address bar. Though less precise than active time, this measure allows us to track potential substitution toward competing platforms—including other social media and related websites listed in Beknazar-Yuzbashev et al. (2025) and each participant's ten most-visited domains during baseline.

Beyond time, we record daily counts of content consumption at two levels: organic content (posts and comments viewed) and advertising content (impressions and, as a deeper measure of engagement, clicks). All measures are based on the feed data described in Section 3.4.1.

We combine time and content engagement measures into two summary indices for Facebook. The platform engagement index aggregates active time and organic content consumed, and the ad engagement index aggregates ad impressions and ad clicks. We follow the index construction method of Kling et al. (2007).

3.4.5 Long-Run Outcomes

Treatment effects may take time to fully emerge—or fade—so we extend observation beyond the pre-registered six-week intervention window. To ensure comparability across cohorts, we truncate all observation windows at a common endpoint: 176 days from enrollment. As a result, we track outcomes for approximately 19 weeks (134 days) post-randomization.

3.4.6 Survey Measures

The baseline survey collects demographic characteristics—including age, sex, race, education, household income, and political affiliation—as well as self-reported Facebook usage patterns. Both the baseline and endline survey elicit willingness to accept (WTA) payment to deactivate Facebook for four weeks.

3.5 Empirical Strategy

Our main specification exploits within-subject variation using a difference-in-differences design. For engagement outcomes recorded daily by the extension, we compare the six-week

baseline period (treatments disabled) and the six-week intervention period (treatments enabled), with the individual-day as the unit of observation. Specifically, we estimate the following two-way fixed effects (TWFE) regression:

$$Y_{i,t} = \alpha_i + \delta_{t,C_i} + \beta D_{i,t} + Errors_{i,t} + \varepsilon_{i,t}, \tag{3}$$

where α_i are individual fixed effects, δ_{t,C_i} are cohort $\times$ period fixed effects, and $D_{i,t}$ equals one if user i is assigned to treatment and period t falls in the intervention window. We index time in days relative to each individual's intervention start date, with $t = 0$ denoting the final day of the baseline period. The coefficient β identifies the treatment effect. We include cohort $\times$ period fixed effects, as opposed to just period fixed effects, to ensure that our estimation is robust to the staggered roll-out of the treatment—see Cengiz et al. (2019) for an application of this approach. We include controls for browser extension errors described in Appendix C and their interactions with treatment.[18]

We estimate two main comparisons: *Hide* versus *Hide Control* and *Replace* versus *Replace Control*. All regressions use inverse probability weights to account for the oversampling of the Replace condition described in Section 3.3.4. We use Driscoll and Kraay (1998) standard errors, which exploit the long panel in our study and are robust to both serial and cross-sectional dependence (see Cameron and Miller 2015; Hoechle 2007 provides evidence on finite-sample performance).

For outcomes measured only in surveys (baseline and endline), we estimate a two-period difference-in-differences: the outcome regressed on a treatment indicator, an endline indicator, and their interaction, with individual fixed effects.

To examine heterogeneity by baseline characteristics, we augment the main specification:

$$Y_{i,t} = \alpha_i + \delta_{t,C_i,M_i} + \beta D_{i,t} + \phi\left(D_{i,t}M_i\right) + Errors_{i,t} + \varepsilon_{i,t}. \tag{4}$$

[18]We account for three types of possible errors: (i) website update—Facebook made a sudden change to the HTML code of the website, which resulted in issues such as advertiser names displaying incorrectly, (ii) faulty replacement—eligible ads in the *Replace* group were incorrectly replaced, (iii) faulty hiding—the hiding intervention was incorrectly enabled during parts of the baseline period for a random subset of participants. Such errors are common in browser-extension-based research, as extensions must be adapted to ongoing changes in platform HTML structure. The share of user-day observations in which participants experienced at least one instance of each error was small: 0.049 for the website update, 0.005 for faulty replacement, and 0.004 for faulty hiding for the pre-registered study period. Depending on whether the error affected all treatment groups, occurred for different individuals on different days, or occurred during both the baseline and intervention periods, some variables in $Errors_{i,t}$ may be absorbed by the intervention start $\times$ period fixed effects. See Appendix C for a more comprehensive description of the errors and related robustness analysis.

where M_i indicates whether user i is above the median of the characteristic of interest. The coefficient β captures treatment effects for below-median users; $\beta + \phi$ captures effects for above-median users. Cohort $\times$ period fixed effects are interacted with M_i to allow differential trends.

4 Results

4.1 Descriptive Evidence on Ad Loads

The model in Section 2 highlights three channels that may sustain advertising model dominance: user insensitivity to ad loads, improved match quality through microtargeting, and ad load personalization across users. We begin by documenting descriptive patterns in ad loads that speak directly to the third channel.

4.1.1 Ad Load Heterogeneity

We first characterize ad loads on three major social media platforms. Figure 1 presents distributions of feed ad loads consumers see on average in the baseline period. The average consumer faces a 12.8% ad load on Facebook, 11.4% ad load on X, and 10.0% ad load on Reddit. Ad loads vary substantially across users on all three platforms: the standard deviations of ad loads across users are 4.6, 6.8, and 3.2 percentage points on Facebook, X, and Reddit, respectively.

Most user-day variation in feed ad load reflects persistent individual differences rather than day-to-day fluctuations. Using a random-intercept decomposition of daily feed ad load (Figure A6), we estimate between-user shares of 65.6% on Facebook, 64.5% on X, and 71.0% on Reddit. This dominance of user-specific differences suggests systematic ad-load personalization.

We find a consistent pattern using experienced ad loads, a measure that accounts for typical user activity on the platform. Figure A5 presents experienced ad loads both across user-day pairs (A5a) and across users (A5b); the two distributions are similar, confirming that the heterogeneity documented above is driven by across-user rather than within-user variation.

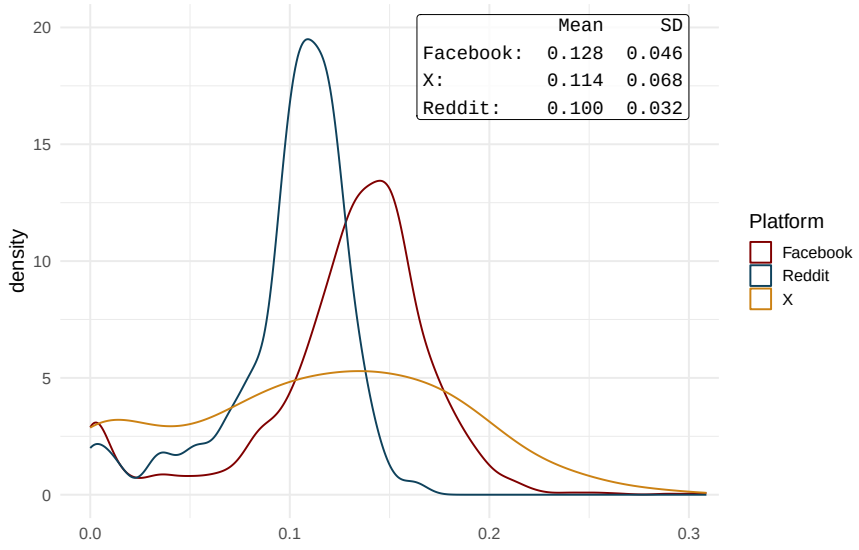


FIGURE 1: DISTRIBUTION OF INDIVIDUAL FEED AD LOAD BY PLATFORM

Note: The figure presents the distributions of the feed advertising load across users on Facebook, Reddit and X, averaged at the user level over the entire untreated period. The feed ad load is defined as the number of ads as a share of total content in the feed (both posts and comments) shown to the user on the platform. The browser extension can directly identify feed elements on Reddit and X. On Facebook, we consider an element to be in the feed unless it is a part of a sequence of at least four consecutive comments—comments in the feed do not occur in groups larger than three. The figure relies on the subsample of users who viewed at least 50 posts and comments throughout the relevant period.

4.1.2 Ad Load Correlates

Which user characteristics predict ad loads? Figure 2 reports general dominance—the average incremental R^2 each covariate contributes across all possible subset models—for a set of demographic and technology-use covariates predicting feed ad loads. Age is the strongest predictor by a wide margin: age and age squared rank first and second. The importance of age is consistent with older users being both less price-elastic and more valuable to advertisers (Gentzkow et al. 2024). Facebook desktop usage share, income, political affiliation, and platform valuation are weaker predictors.

The observed patterns are broadly consistent with our model's predictions (Section 2) on optimal ad-load setting. We pre-registered three tests, shown in Figure 3. First, the model predicts that users who generate larger network spillovers receive lower ad loads, as the platform protects their participation; consistent with this, users with above-median Facebook friends face significantly lower ad loads. Second, the model predicts that more ad-sensitive users receive lower ad loads; users with above-median predicted sensitivity—estimated via generalized random forests (Athey et al. 2019) using the differences between

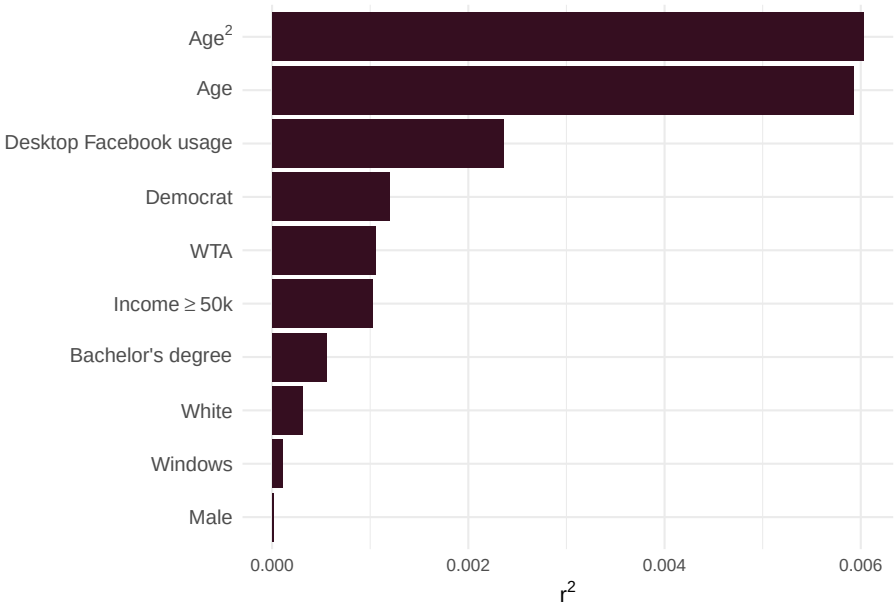


FIGURE 2: GENERAL DOMINANCE OF FEED AD LOAD PREDICTORS ON FACEBOOK

Note: The figure reports general dominance for a broad set of covariates predicting feed ad loads on Facebook, i.e., the average incremental R^2 a variable contributes when considering all possible subset models. The included variables are as follows: (1) age in years, (2) age squared, (3) an indicator for having at least a bachelor's degree, (4) an indicator for being a Democrat, (5) desktop share of Facebook usage (from the baseline survey), (6) an indicator for having household income above $50,000, (7) an indicator for being male, (8) an indicator for identifying as white/Caucasian, (9) an indicator for using Windows operating system, (10) willingness to accept to deactivate Facebook for four weeks during the baseline survey (WTA).

the *Hide* and *Hide Control* groups—receive feed ad loads about 3.5 percentage points lower, the largest gap among the three tests. Third, the model is ambiguous about ad value, since higher-value users generate more revenue per impression but may also be more worth retaining. Empirically, high-value users receive higher ad loads, suggesting that margin considerations dominate retention motives.

4.1.3 Ad Loads Across Platforms

Do different platforms charge the same users similar "attention prices"? Individual ad loads correlate positively across Facebook, X, and Reddit—most strongly between X and Reddit ($r = 0.43$, $p = 0.01$), more modestly between Facebook and X ($r = 0.15$, $p = 0.03$) and between Facebook and Reddit ($r = 0.29$, $p = 0.001$). This suggests that platforms tailor ad loads to similar user characteristics.

Platforms also show some overlap in which ad topics they show to the same users. We classify advertisers into 26 categories based on the IAB Content Taxonomy and compute, for

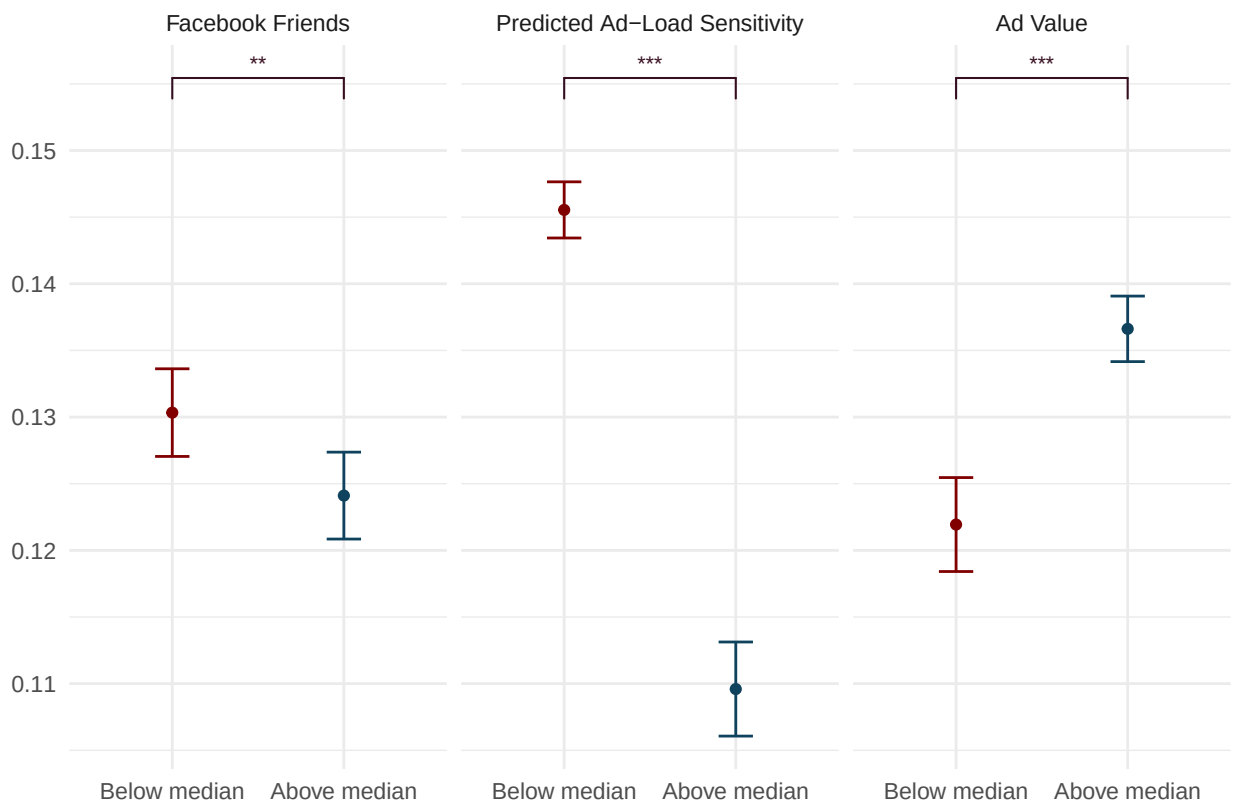


FIGURE 3: TESTING PREDICTIONS ON AD LOADS HETEROGENEITY

Note: The figure reports estimates from regressions of individual-level average feed ad load on an indicator equal to one if the heterogeneity variable exceeds its median. The heterogeneity variables are: (1) ad value, (2) number of Facebook friends, and (3) predicted ad load sensitivity. Ad value measures how valuable it is for the platform to sell ad slots targeting a given user, based on demographics. To construct individual-level ad value, we use $100 advertising campaigns targeted by sex, age, and high/low–income ZIP codes and obtain Facebook's estimates of cost per click; ad value is then computed as the product of cost per click and the individual's ad click rate. Predicted ad load sensitivities are computed using a generalized random forest approach, which estimates conditional average treatment effects of the hiding intervention as functions of individual covariates (baseline characteristics and demographics) on users with at least 50 observed feed elements in both the baseline and intervention periods after excluding user-days in documented measurement-error windows. Results are robust to excluding baseline ad load from the specification predicting sensitivity. Higher values mean more negative effect from an ad load increase. The figure reports point estimates with 95% confidence intervals.

each pair of platforms, within-topic correlations in how users are targeted.[19] The resulting correlations (Figure A10) are positive and significant for all platform pairs.[20] They are also modest: 0.19 for Facebook–Reddit, 0.05 for X–Reddit, and 0.10 for Facebook–X. Platforms thus share some commonality in how they target users by topic, but also appear to rely on distinct signals.

[19]Specifically, we classify each advertiser using Tier1 of the IAB Content Taxonomy v1.0. For each advertiser, we send its name and up to five recent ad texts to GPT-5-mini and repeat the classification three times, taking the majority vote as the final category. See Figure A7 in the appendix for distributions.

[20]To obtain platform-level measures, for each pair of platforms we take the mean of the diagonal correlation coefficients, which capture within-topic correlations in how users are targeted across platforms.

4.2 Experimental Results

4.2.1 First Stage

Hiding Intervention and Ad Load. Figure 4 shows the average Facebook experienced ad load by treatment condition over the study period. The difference-in-differences estimate indicates a reduction of 9.5 pp relative to the post-intervention average of 11.4% in *Hide Control* (Table B3).

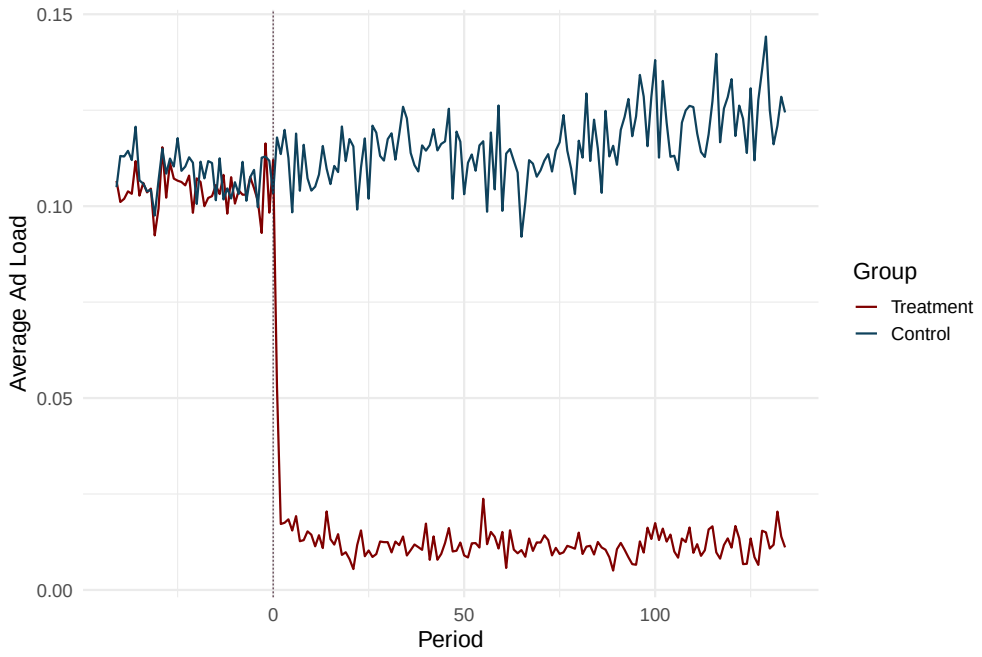


FIGURE 4: AVERAGE EXPERIENCED ADVERTISING LOADS ON FACEBOOK

Note: The figure shows the average daily advertising load on Facebook for the *Hide* condition and the *Hide Control* condition. Advertising load is defined as the number of ads shown as a share of all Facebook content viewed by the user (ads, posts, and comments). The x-axis represents study periods, measured in days relative to each individual's intervention start date. The vertical dashed line marks the final day of the pre-intervention period (period 0).

The platform's ad-serving algorithm did not respond to the intervention. Supplied ad load—measured before our filtering—was virtually unchanged between *Hide* and *Hide Control* (0.3 pp, Table B3, column 3, panel A), and long-term estimates are nearly identical (Table B4). Facebook continued to serve the same share of ads to treated users despite their dramatically different experienced ad loads.

Replacement Intervention and Microtargeting. Figure 5 shows average ad distance—our measure of ad-user match quality—over time. During the baseline period, ad distance

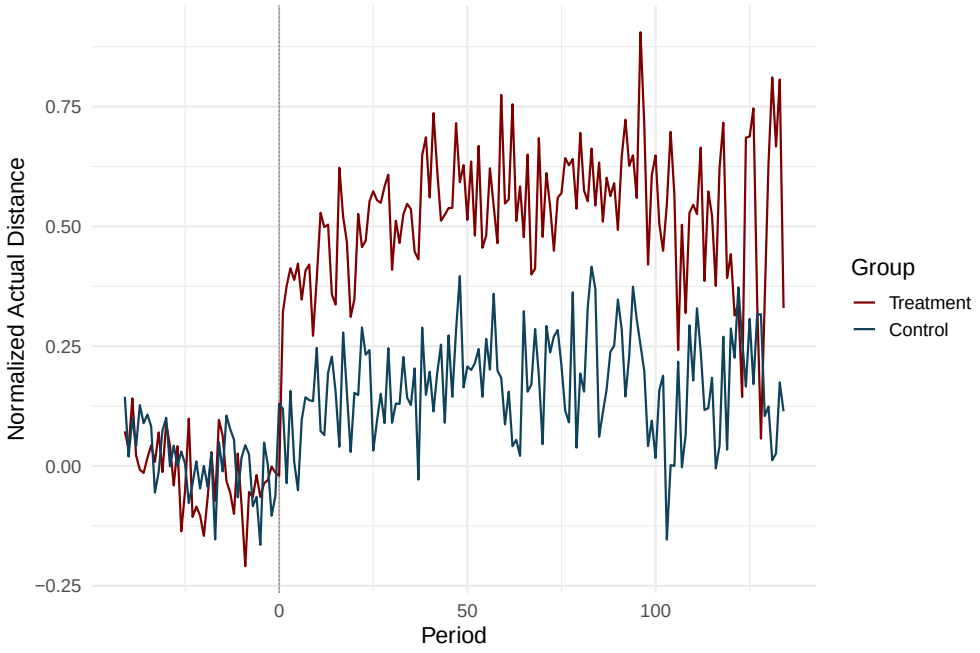


FIGURE 5: NORMALIZED AD DISTANCE ON FACEBOOK

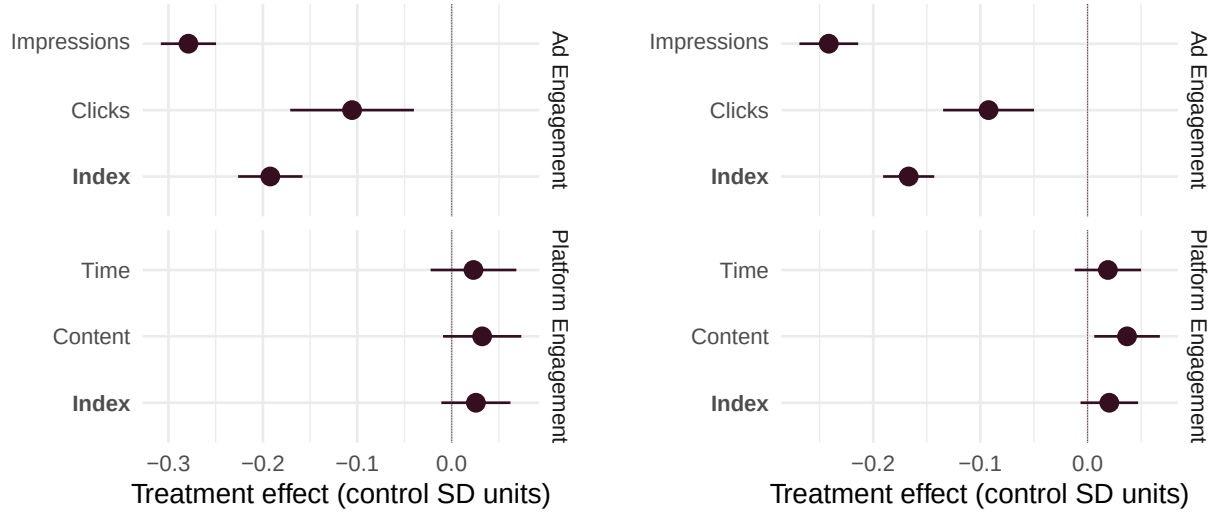
Note: The figure displays the average daily ad distance on Facebook, disaggregated by treatment condition (Replace vs. Replace Control). Ad distance is defined as the Euclidean distance between a user's characteristics and the average demographic profile targeted by that advertiser, based on baseline data. The values on the y-axis are normalized using the mean and standard deviation from the baseline period. The x-axis represents study periods, measured in days relative to each individual's intervention start date. The vertical dashed line marks the final day of the pre-intervention period (period 0).

evolved similarly across treatment groups. Immediately after the intervention, ad distance in the *Replace* condition jumped and remained elevated, confirming that the replacement intervention successfully degraded microtargeting. The difference-in-differences estimate indicates an 11% increase in average ad distance (Table B3, column 2, panel B).

Unlike the hiding intervention, replacement triggered a mild algorithmic response: Facebook adjusted the ads it supplied, raising the supplied ad distance by 1.5% (Table B3, column 4, panel B). The long-run supplied-distance coefficient is smaller and only significant at the 10% level (Table B4). The asymmetry likely reflects how the algorithm weights recent interactions—as users engaged with different ad content, the platform updated its model of their preferences. The pattern suggests that Facebook treats users' ad-load elasticities as stable but allows inferred ad preferences to evolve over time.

4.2.2 Effects of Hiding Ads

Ad Engagement. Figure 6(a) presents the short-run treatment effects; Table B5 reports full regression results. The hiding intervention sharply reduced ad engagement: impressions fell by 95% (0.28 SD) and clicks by 58% (0.11 SD). The combined ad engagement index declined by 0.19 SD. Long-run effects over 19 weeks (Figure 6(b); Table B6) are consistent in both direction and magnitude.



(a) Short-Term Effects

(b) Long-Term Effects

FIGURE 6: OWN-PLATFORM ENGAGEMENT EFFECTS OF HIDING INTERVENTION

Note: The figure displays estimated treatment effects of the hiding intervention on Facebook using Equation 3 and our main experimental sample. The dependent variables are: (i) number of ad impressions, (ii) number of ad clicks, (iii) an ad engagement index based on (i) and (ii), (iv) active time spent on the platform, (v) number of posts and comments shown to the user, and (vi) a platform engagement index based on (iv) and (v). **Panel A** shows results based on the six-week intervention period, whereas **Panel B** shows results based on a longer intervention period (134 day-periods or 19.1 weeks). The unit of observation is the individual-day, measured relative to the intervention date. We include 95% confidence intervals based on Driscoll-Kraay standard errors. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day.

Platform Engagement. While ad engagement fell, overall platform engagement increased. Using our pre-registered specification in levels, the short-term effects are modest: content consumption increased by 5.6 pieces per day (9%, or 0.03 SD) and active time by 0.65 minutes per day, although neither is statistically significant (Figure 6(a); Table B5). Long-term estimates are similar in magnitude: content consumption increased by 0.04 SD ($p < 0.05$) and active time by 0.57 minutes per day (Figure 6(b); Table B6). The overall platform engagement index rose by 0.03 SD in the short run and 0.02 SD in the long run. The effects

emerge early and persist over time.

Outliers and skewness in the levels specification reduce precision of our estimates (see Section 4.2.6). We address this with two alternative specifications: an inverse hyperbolic sine (*asinh*) transformation and a levels specification trimming the top percentile. Both yield significant positive effects on active time and content consumption in both the short and long run (Tables B7, B9, B8, and B10). Under these specifications, the platform engagement index rises by 0.05–0.07 SD ($p < 0.01$)—roughly double the baseline estimate—and active time increases by roughly 1.6 minutes per day in the long run.

Elasticities. We translate these effects into an elasticity: the ratio of the percentage change in active time to the percentage change in ad load. Using our difference-in-differences estimates (Tables B3 and B5), the short-run elasticity is:

$$\text{Elasticity} = \frac{\%\Delta\text{Active Time}}{\%\Delta\text{Ad Load}} = -0.07.$$

The *asinh* specification yields –0.16 (Table B7).[21] Both estimates are below 0.2 in absolute value, indicating highly inelastic demand. This accords with Brynjolfsson et al. (2025), who reach a similar conclusion.

4.2.3 Effects of Replacing Ads

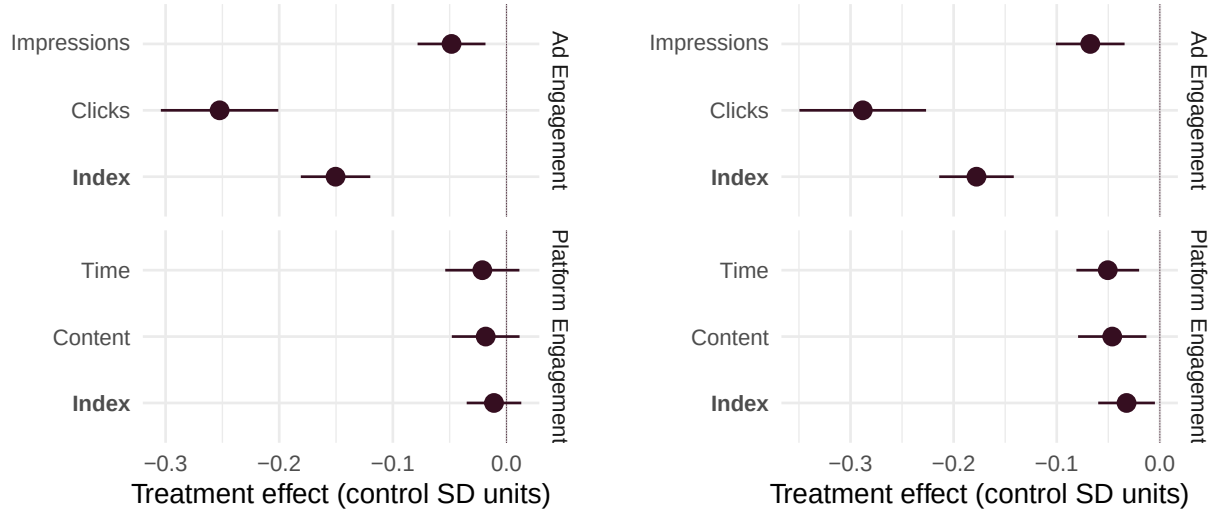
Ad Engagement. Figure 7(a) presents short-term effects of the replacement intervention; Table B5 reports full results. Replacement reduced ad engagement: impressions fell by 0.05 SD and clicks by 0.25 SD. The combined index declined by 0.15 SD. Long-term effects are similar, with the index falling by 0.18 SD (Figure 7(b); Table B6). These results indicate that microtargeting is a key driver of ad engagement.

Platform Engagement. In the short run, the replacement intervention had no statistically significant effect on platform engagement, a pattern that holds across specifications

[21] To express treatment effects estimated on *asinh*-transformed outcomes as percentage changes in the original (level) units, we use the following formula:

$$\%\Delta Y = \frac{\sinh(m+\beta) - \sinh(m)}{\sinh(m)} \times 100,$$

where β is the estimated treatment effect and m is the mean of the *asinh*-transformed outcome variable. This expression maps the estimated effect back to the original scale and expresses the change as a percentage relative to the mean.



(a) Short-Term Effects

(b) Long-Term Effects

FIGURE 7: OWN-PLATFORM ENGAGEMENT EFFECTS OF REPLACEMENT INTERVENTION

Note: The figure displays estimated treatment effects of the replacement intervention on Facebook relative to *Replace Control* using Equation 3 and our main experimental sample. The dependent variables are: (i) number of ad impressions, (ii) number of ad clicks, (iii) an ad engagement index based on (i) and (ii), (iv) active time spent on the platform, (v) number of posts and comments shown to the user, and (vi) a platform engagement index based on (iv) and (v). **Panel A** shows results based on the six-week intervention period, whereas **Panel B** shows results based on a longer intervention period (134 day-periods or 19.1 weeks). The unit of observation is the individual-day, measured relative to the intervention date. We include 95% confidence intervals based on Driscoll-Kraay standard errors. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day.

(Figure 7(a); Table B5). Over the longer run, however, clear negative effects emerged (Figure 7(b) and Table B6). Active time fell by 1.7 minutes per day (13%, 0.05 SD, $p < 0.01$) and content consumption by 9.4 posts per day (13%, 0.05 SD, $p < 0.01$), with the platform engagement index declining by 0.03 SD ($p < 0.05$). These long-run results are robust to trimming (0.04 SD, $p < 0.05$) and *asinh* specifications (0.06 SD, $p < 0.01$).

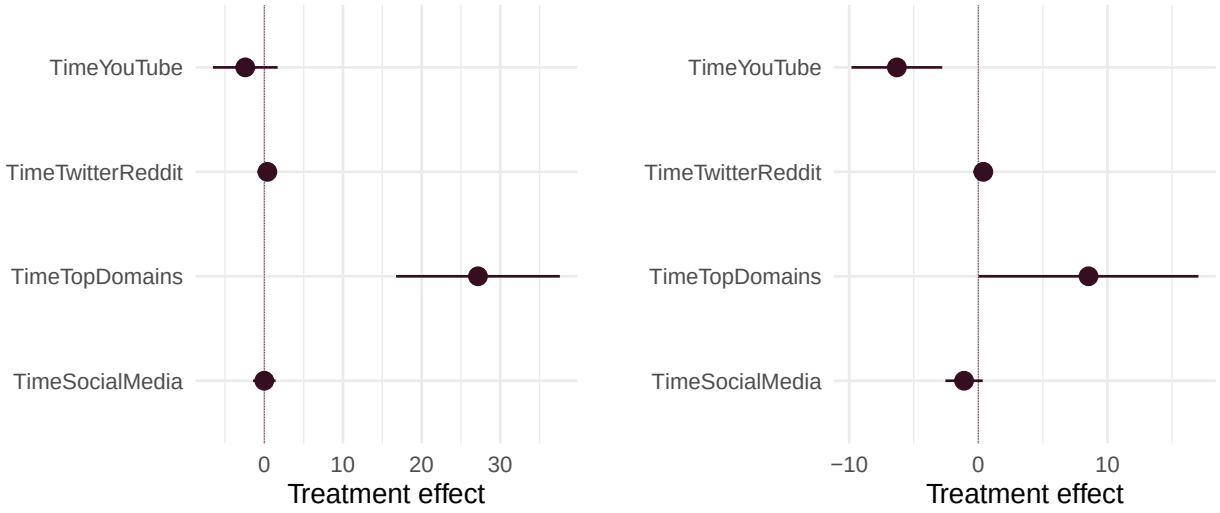
The delayed emergence of platform engagement effects suggests that users update their beliefs about the platform experience gradually as they encounter less relevant ads—an important consideration underscoring the need for long-duration interventions to inform policy. The replacement intervention thus reduced both ad and platform engagement, implying that both consumers and the platform may benefit from ad microtargeting.

4.2.4 Substitution Effects

Figure 8(a) presents effects of the hiding intervention on time spent on competing platforms. We find little evidence of substitution: the intervention does not significantly affect time

on X and Reddit, YouTube, or the combined set of pre-registered social media platforms. However, time on users' top 10 most-visited domains increased—possibly a "linking effect," whereby greater Facebook engagement drives traffic to sites frequently shared in users' feeds.

The replacement intervention yields no clear substitution pattern either (Figure 8(b)). Active time on X and Reddit increased marginally, as did time on top-visited domains, but time on YouTube fell and effects on the broader set of social media platforms are insignificant.



(a) Hiding Intervention

(b) Replacement Intervention

FIGURE 8: IMPACTS ON ENGAGEMENT WITH NON-TREATED PLATFORMS

Note: **Panel A** reports estimated treatment effects of the hiding intervention on Facebook, relative to the *Hide Control* group, based on Equation 3 and the main experimental sample. **Panel B** presents treatment effects of the replacement intervention, relative to the *Replace Control* group. The dependent variables are: (i) time spent on YouTube; (ii) combined active time on X and Reddit; (iii) time spent on the top 10 domains as determined from baseline data; and (iv) time spent on a set of non-treated social media platforms defined in Beknazar-Yuzbashev et al. (2025). All outcome variables are inverse hyperbolic sine (asinh) transformed. The unit of observation is the individual-day, indexed relative to the intervention start date. Regressions include 95% confidence intervals based on Driscoll-Kraay standard errors. All estimates are weighted using inverse probability weights reflecting group assignment probabilities conditional on recruitment day.

4.2.5 Platform Valuation

Figure 9 presents effects on platform valuation, measured as willingness to accept (WTA) payment to deactivate Facebook for four weeks; Table B18 reports full results. Neither intervention significantly affects WTA. The null result for hiding is consistent with Brynjolfsson et al. (2025).[22] We extend their finding by showing that degraded microtargeting also leaves platform valuation unchanged.

[22]Our estimates of the effects of the interventions on WTA are less precise than theirs due to a smaller sample size.

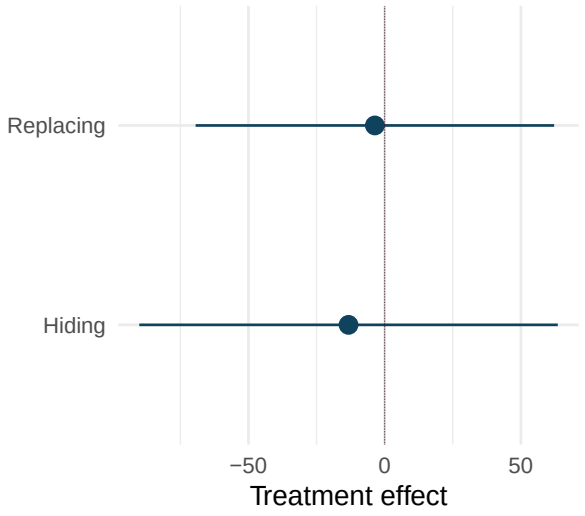


FIGURE 9: EFFECTS OF INTERVENTIONS ON PLATFORM VALUATION

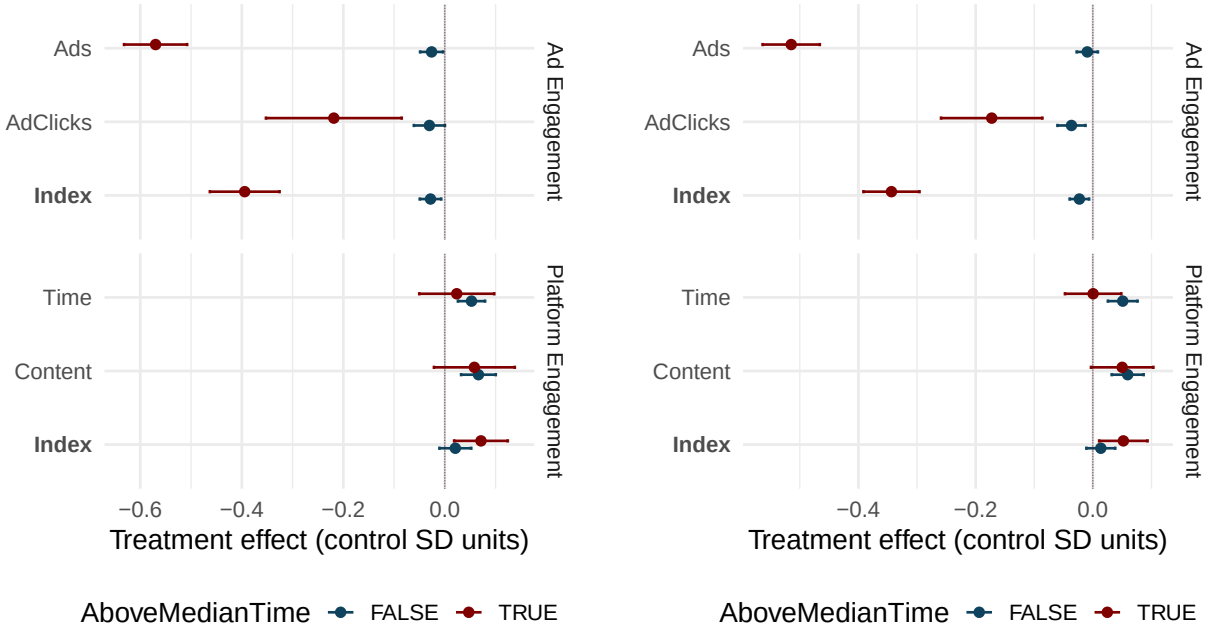
Note: The figure displays the estimated treatment effects of the hiding and replacement interventions on individuals' willingness to accept (WTA) the deactivation of Facebook for four weeks. WTA is measured in both the baseline and endline surveys. The estimates are based on a difference-in-differences regression of WTA on the endline survey indicator and its interaction with the treatment group indicator, with individual fixed effects. The unit of observation is the individual, and 95% confidence intervals are provided, based on standard errors clustered at the individual level. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day.

4.2.6 Heterogeneous Effects

Figure 10 presents treatment effects of the hiding intervention by baseline engagement (above vs. below median active time). Negative effects on ad engagement concentrate among heavy users: the ad engagement index falls by 0.35–0.40 SD for above-median users, with smaller but detectable effects for below-median users.

Platform engagement effects are less clear-cut. Confidence intervals for above-median users are wide, driven by outliers among heavy users. As before, we address this with an *asinh* specification (Figure A8). Both groups show significant positive effects on platform engagement of roughly 0.05 SD.

Figure 11 presents corresponding results for the replacement intervention. Heavy users again drive the negative effects on ad engagement: ad clicks fall by more than 0.5 SD for above-median users. Short-run effects on platform engagement are insignificant for both groups, consistent with Section 4.2.3. In the long run, however, above-median users show clear declines—active time and content consumption both fall by nearly 0.1 SD. Results are similar under the asinh specification (Figure A9).



(a) Short-Term Effects

(b) Long-Term Effects

FIGURE 10: HETEROGENEITY OF ENGAGEMENT EFFECTS: HIDING

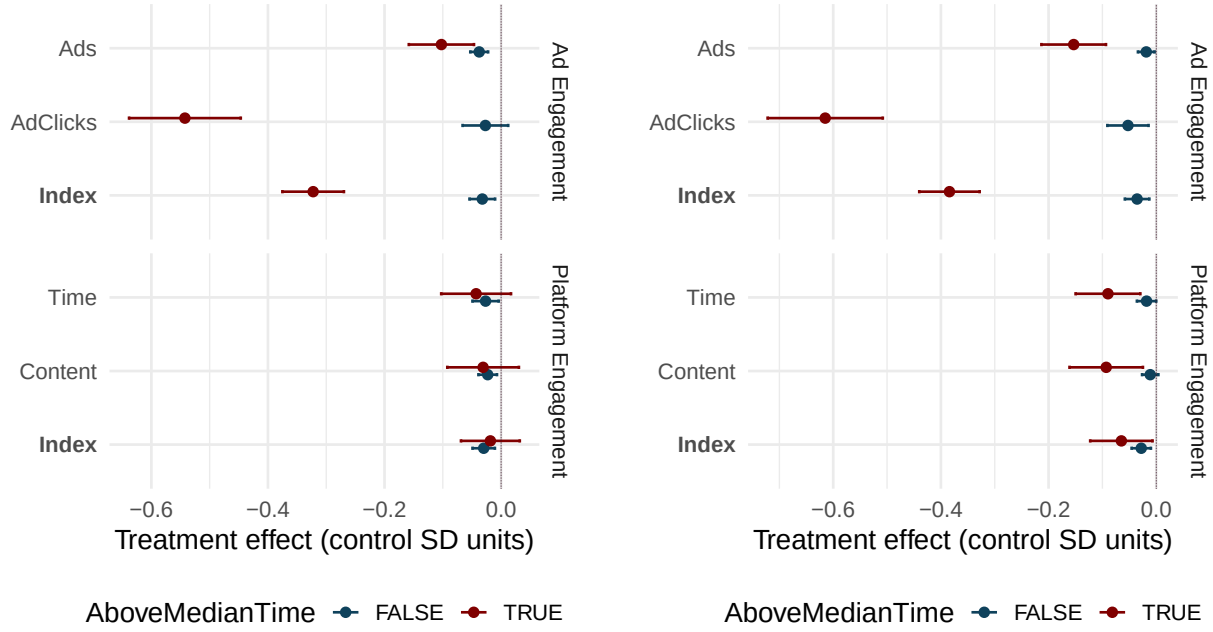
Note: The figure displays estimated treatment effects of the hiding intervention on Facebook relative to *Hide Control* separately for users whose baseline active time on Facebook was above and below the median. The estimation is based on Equation 4, with $\beta + \phi$ representing the above-median coefficient and β representing the below-median coefficient. The regression relies on the main experimental sample. The dependent variables are: (i) number of ad impressions, (ii) number of ad clicks, (iii) an ad engagement index based on (i) and (ii), (iv) active time spent on the platform, (v) number of posts and comments shown to the user, and (vi) a platform engagement index based on (iv) and (v). **Panel A** shows results based on the six-week intervention period, whereas **Panel B** shows results based on a longer intervention period (134 day-periods or 19.1 weeks). The unit of observation is the individual-day, measured relative to the intervention date. We include 95% confidence intervals based on Driscoll-Kraay standard errors. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day.

4.3 Robustness

This section discusses potential concerns and robustness checks. Section 4.3.1 shows that attrition does not vary by treatment assignment. Section 4.3.2 demonstrates robustness to alternative specifications. Appendix C confirms that results are unchanged when excluding participants who experienced browser extension errors.

4.3.1 Attrition

Figure A13 in the appendix shows survival rates by treatment group over time. We measure attrition conservatively: an individual i is classified as having survived until period t only if they were active on day t or later. Survival rates at the end of the six-week intervention



(a) Short-Term Effects

(b) Long-Term Effects

FIGURE 11: HETEROGENEITY OF ENGAGEMENT EFFECTS: REPLACEMENT

Note: The figure displays estimated treatment effects of the replacement intervention on Facebook relative to *Replace Control* separately for users whose baseline active time on Facebook was above and below the median. The estimation is based on Equation 4, with $\beta + \phi$ representing the above-median coefficient and β representing the below-median coefficient. The regression relies on the main experimental sample. The dependent variables are: (i) number of ad impressions, (ii) number of ad clicks, (iii) an ad engagement index based on (i) and (ii), (iv) active time spent on the platform, (v) number of posts and comments shown to the user, and (vi) a platform engagement index based on (iv) and (v). **Panel A** shows results based on the six-week intervention period, whereas **Panel B** shows results based on a longer intervention period (134 day-periods or 19.1 weeks). The unit of observation is the individual-day, measured relative to the intervention date. We include 95% confidence intervals based on Driscoll-Kraay standard errors. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day.

exceed 75% in all arms, ranging from 76.1% (*Replace*) to 76.9% (*Replace Control*), and the trajectories are parallel across conditions throughout the study period. Table B11 provides formal tests: regressions of survival on treatment assignment yield no significant pairwise differences, and interactions with predicted survival probability are uniformly insignificant.[23]

Because we report long-term outcomes, we also examine attrition over the full observation window. Figure A13 shows that survival trajectories remain parallel well beyond the six-week intervention (dashed vertical line). Table B12 confirms this formally: regressions of uncapped dropout timing on treatment indicators reveal no differential attrition. We conclude that attrition is unlikely to explain either the short- or long-run results.

[23]The dependent variable is capped at day 42, the final day of the six-week intervention; baseline periods are indexed as zero and below.

4.3.2 Alternative Specifications

Skewness and Outliers. As noted in Sections 4.2.2 and 4.2.3, our main specification uses untransformed outcomes, which are right-skewed. We address this with two alternatives: an inverse hyperbolic sine transformation (Tables B7 and B8) and top-1% trimming (Tables B9 and B10). Both yield improved precision and reinforce the main findings: the hiding intervention increases platform engagement in the short and long term, while the replacement intervention reduces it in the long term.

Alternative Standard Errors. Our main specification uses Driscoll-Kraay standard errors, which account for both serial and cross-sectional dependence. As a more conservative alternative, we cluster at the individual level (Tables B13 and B14). Under clustering, ad engagement effects remain significant for both interventions (1% for hiding, 10% for replacement), though platform engagement effects lose significance in the levels specification. Under the *asinh* transformation, the positive effects of hiding on platform engagement and the long-term negative effects of replacement are robust to individual-level clustering (Tables B15 and B16).

4.4 Mechanisms

Supplementary Collection. We conducted supplementary data collection with $N = 215$ regular U.S. Facebook users to characterize how they attend to and perceive advertisements during a naturalistic Facebook session.[24] We designed it to understand why users appear largely inelastic to ad loads. We collect measures to distinguish two alternative explanations: (1) an avoidance channel, in which users readily ignore or scroll past ads so that ads impose little attentional cost, and (2) a valuation channel, in which users find ads nearly as interesting or valuable as organic posts. Our data collection was preregistered on AsPredicted under the number 298326.

Study Flow. Participants are recruited via CloudResearch and screened for using Facebook at least monthly and for completing the study on a desktop Chrome browser. They then install the study Chrome extension and browse their own Facebook feed as they normally

[24]The sample includes everyone who installed the study browser extension, consumed at least one feed element, and whose survey response could be matched to their extension data; survey completion was not required.

would for five minutes. While the participant scrolls, the extension records, for each feed element, the element type (ad, post, comment) and its geometry (width, height, position). It also records measures that capture how long each element stays on screen: viewport-edge events as the element enters and leaves the screen; the time each element spends in the viewport (the timer pauses when the tab becomes inactive, e.g., on a tab switch); and a fixed-interval visibility-and-scroll tick every 250 milliseconds. All extension-based measures are computed over the window up to the end of the participant's survey session (their last survey timestamp). After the browsing session, participants complete a survey that provides the perception, recall, and valuation measures.

Mobile Session. Participants also have the option to complete a second five-minute session in the Facebook mobile app for an additional incentive: they screen-record their phone and then upload the recording to cloud storage via a QR-code link. In total, 97 participants uploaded a recording with feed content.[25] Each mobile recording is annotated by OpenAI's GPT-5.4-mini that labels every feed element as an ad or post, records the time it remains on screen, and counts the comments shown inline beneath each element. Each recording is coded in three independent passes that are reconciled within participant; Appendix E.3 details the recording, annotation, and reconciliation procedure.

Descriptives. Over the five-minute desktop session participants consume a substantial amount of content—the median person is exposed to 26 posts and 4 ads— see Panel A of Table B24 in Appendix E for more summary statistics. The mean feed ad load is 14.7% with a lot of dispersion across individuals (see Figure A19(a)). Around 20% of users have not seen any ads during the session. This is partly explained by the shortness of browsing sessions—this share goes down to 17.5% if we restrict the sample to participants who saw at least 17 pieces of content (Figure A19(b))—and partly explained by ad blocker usage: 11.7% of participants report having an ad blocker enabled while browsing Facebook during the study, and among them the mean feed ad load is lower, at 7.2% (roughly half the 15.8% of non-users; Table B25).

On the optional mobile session (Panel B of Table B24), browsing intensity is broadly comparable to desktop: the median participant again sees 26 posts. Feed ad load is higher,

[25] Of the 97 participants who uploaded a video recording, 84 have viable extension data from their desktop session (meaning that they correctly installed the extension and browsed through Facebook). The latter sample allows within-person cross-device comparisons.

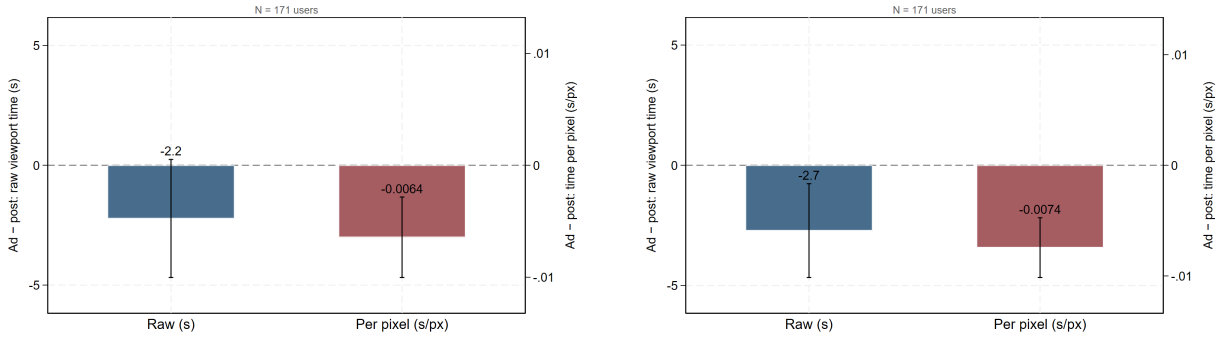
with a mean of 26.6% (median 27.3%). Figure A19 overlays the desktop and mobile distributions. The two device types also differ in where ads appear: desktop ads concentrate at the top of the feed, while mobile ads are spread more evenly across positions (Figure A20).

Relative Screen Time. We find evidence against the pure valuation channel explaining limited ad elasticity: ads receive less screen time than organic posts, suggesting that participants engage with and value ads less than organic content. Comparing each participant's own ad and post averages during the desktop session, the typical ad is held in the viewport about 2.2 seconds less than the typical post on the basic measure and 2.7 seconds less on the counter measure—the latter excludes time when the tab is inactive (Figures 12(a) and 12(b)). The gap holds after adjusting for element size: per pixel of height, ads still draw less time than posts (right axis of each panel). In a pooled, element-level regression, ads draw about 10–11% less viewport time than posts, and this gap is robust to controlling for element height and adding user fixed effects (Table B26).

Moreover, Figure A21 shows that ads and posts fill a similar share of the screen in the first two seconds after entering the viewport, but ad coverage then falls sharply relative to posts in the seconds that follow. This early similarity suggests users pay a small time cost to evaluate each element before scrolling ads out of view. The pattern is consistent with the ad avoidance channel partly explaining users' inelasticity to ad load: it takes little time for users to avoid ads they do not like.

The same pattern appears in the mobile browsing data. Comparing each participant's own ad and post averages, ads again receive less viewport time than posts on both the basic and counter measures (Figure A26). We note that mobile measures of viewport time might have classification error: they are reconstructed from screen recordings by a large language model (Appendix E.3). However, the consistency of the ad–post gap across devices reinforces the conclusion that the valuation channel cannot fully explain limited elasticity to ad loads.

Stated Interest and Value. The survey responses provide further evidence against the valuation channel. Comparing the ads participants saw this session with organic posts, about 71% of participants rate ads as less interesting (the modal response is "much less interesting") and fewer than 10% rate them more interesting (Figure A22(a) in the appendix). Ads fare no better in replies about their value: roughly 81% call the session's ads "not at all" or only "slightly" valuable (Figure A22(b) in the appendix).



(a) Basic (timestamp) measure

(b) Counter (logged) measure

FIGURE 12: INDIVIDUAL DIFFERENCE IN VIEWPORT TIME BETWEEN ADS AND POSTS

Note: This figure reports per-participant paired differences in viewport time between ads and posts. For each participant we compute the mean viewport time across their ads minus the mean across their posts; bars show the average of these differences across participants, with 95% confidence intervals. In each panel the left bar (left axis) is the raw difference in seconds and the right bar (right axis) the difference per pixel of element height; values below the dashed zero line indicate that ads are viewed less than posts. Panel (a) uses the basic timestamp measure and panel (b) the counter (logged) measure, which excludes time when the tab is inactive. The sample is participants with at least one ad and one post. The figure is based on the sample recruited during the supplementary data collection (see Section 4.4).

Ad Recall. We also ask whether users remember the ads they encounter. In an aided-recognition task, each participant is shown two ads the extension recorded in their own feed during the session immediately preceding the survey questions, alongside four placebo ads they did not see, and marks the ones they recall (Figure A23 in the appendix). They mark about 51% of their real ads but only about 7% of placebos; this gap shows that users distinguish their own ads from decoys. The modest absolute rate of real ad recall (51%) likely reflects a conservative response—participants mark an item only when highly confident they saw it. Overall, users recognize their own ads even while scrolling past them quickly, which indicates that they do not ignore ad content entirely.

Other Evidence. We provide additional results from the supplementary data collection in Appendix E.2. This includes evidence that participants are accurate about their ad exposure on average but noisy at the individual level (Figure A27), and that users find ads easy to distinguish and to have lower general value than organic posts (Figures A28 and A29).

4.5 External Validity

Our main experimental sample differs from the full Facebook population in two respects that are worth noting. The extension operates only on desktop browsers, whereas a substantial

share of Facebook usage occurs on mobile devices; to the extent that desktop and mobile users differ in their sensitivity to ads, our estimates may not extrapolate to the full user base. Our sample also consists of U.S.-based, English-speaking users who responded to a Facebook recruitment ad, introducing potential self-selection. Our results are therefore most directly informative about active Facebook users who engage with platform content. These limitations are common to browser-extension-based designs and should be weighed against the advantage of observing actual platform behavior without relying on self-reports or platform cooperation.

Independent evidence suggests that neither limitation materially affects our main conclusions. Our core finding—that engagement is highly inelastic to ad loads—is consistent with Brynjolfsson et al. (2025), who run a server-side experiment removing ads for a random 0.5% of *all* Facebook users over nine years. Their design addresses both concerns above: it captures mobile and desktop usage, and their sample is drawn randomly from the full user base rather than recruited through advertising. The agreement of elasticity estimates across two very different experimental designs suggests that our findings are not driven by the desktop or self-selection restrictions.

Because participants enrolled by clicking our recruitment ad, they might be unusually prone to clicking ads in general. We do not find evidence of that: advertising click-through rates in our data match industry benchmarks. Pooling baseline ad impressions and clicks, our participants' click-through rate is 1.73%, in line with the 1.5–1.7% reported for U.S. Facebook campaigns in 2024–2025.[26] If anything, the typical participant clicks less often—the median user-level rate is 0.86%. Overall, we find no indication that ad-based recruitment selected unusually ad-responsive users.

We can also compare the field experiment sample, recruited via social media advertising, with the supplementary sample recruited on CloudResearch (see Section 4.4). The two encounter very similar feed ad loads: the mean is 14.7% in the supplementary sample versus 12.8% in the field experiment (Figure A24). The distribution of individual feed ad loads in the supplementary sample (the desktop distribution in Figure A19) has similar shape to the field's Facebook distribution (Figure 1) but is slightly more dispersed—a difference we attribute to a shorter, five-minute measurement window, which estimates each participant's

[26] WordStream reports average click-through rates of 1.57% (February 2023–April 2024 data) and 1.71% (April 2024–June 2025) for U.S. Facebook traffic campaigns (see https://www.wordstream.com/blog/facebook-ads-benchmarks-2025, accessed July 11, 2026). Databox reports the median CTR across all industries of 1.54% as of 2026 (see https://databox.com/average-facebook-ctr, accessed July 11, 2026).

load less precisely. The mix of ad formats differs; the supplementary sample sees relatively more video and fewer image ads (Figure A25), consistent with the platform's shift toward video content between the 2024 field collection and the 2026 supplementary collection. Overall, that two samples recruited through very different channels and two years apart encounter similar ad loads is reassuring for the external validity of our results.

Finally, the mobile sessions in our supplementary collection suggest that similar forces are at work on mobile. Ad loads on mobile vary widely across users—the interquartile range runs from 22% to 34% (Table B24, Panel B)—consistent with the ad-load personalization we document on desktop. Average ad loads are also markedly higher on mobile: 26.6% versus 14.7% on desktop (Figure A19). Through the lens of our model, higher ad loads are the platform's optimal response if mobile users are less elastic to ads—a plausible reading, but one our desktop-only experiment cannot test, so we leave measuring mobile ad-load elasticity to future work. We note that we should read the magnitudes of mobile ad loads with some caution, since the mobile figures are reconstructed from screen recordings by a large language model (Appendix E.3).

5 Counterfactual Monetization

We now use the experimental estimates to simulate counterfactual monetization regimes and quantify the decomposition in (1). To do so, we first estimate a structural model of engagement decisions based on the model presented in Section 2.

5.1 Structural Estimation

5.1.1 Functional Forms

We specialize the user's problem of Section 2 to a parametric form. A user i of observable type θ first decides whether to join the platform for T periods or not. A user who joins receives membership benefit ξ_i and a network benefit from aggregate participation (explained below), and accumulates per-period utility flows over the T periods.

Daily engagement. On each day t, the user takes the ad load a_{it} and ad match quality q_{it} as given and chooses engagement $e_{it} \geq 0$ to maximize quadratic utility:

$$v_{it}^{\theta}(a_{it}, q_{it}) = \max_{e_{it} \geq 0} \underbrace{\delta_{it}^{\theta} e_{it} - \frac{\beta^{\theta}}{2} e_{it}^2}_{\text{Utility from content}} + \underbrace{(\varphi^{\theta} q_{it} - \psi^{\theta})\, a_{it}\, e_{it}}_{\text{Utility from consumed ads}} + \underbrace{\gamma^{\theta}(q_{it})\, a_{it}}_{\text{Intrinsic ad utility}}. \tag{5}$$

Engagement is measured as the number of organic posts and comments consumed in the feed, and the ad load as the ratio of ads to organic content in the feed. Ad match quality is the share of a user's ads that are micro-targeted: it equals one under status-quo targeting and falls in the replacement arm (by the share of each user's ads that were replaced with untargeted ones).

We allow δ_{it}^{θ} to vary across individuals and time:

$$\delta_{it}^{\theta} = \delta^{\theta} + \alpha_i + \lambda_t + \nu_{it},$$

where α_i captures persistent unobserved differences in platform attachment beyond the observable type θ, while λ_t captures common shocks such as seasonality and platform-wide changes affecting all users on a given day. The intrinsic ad utility

$$\gamma^{\theta}(q_{it}) = \gamma_0^{\theta} + \gamma_1^{\theta} q_{it}$$

captures the value of ad exposure that is independent of engagement, for example reflecting a privacy disutility from being tracked.

The first-order condition of (5) results in a linear functional form for engagement,

$$e_{it} = \frac{\delta^{\theta} + \alpha_i + \lambda_t + \nu_{it}}{\beta^{\theta}} - \frac{\psi^{\theta}}{\beta^{\theta}} a_{it} + \frac{\varphi^{\theta}}{\beta^{\theta}} a_{it}\, q_{it}. \tag{6}$$

This is the regression of Section 4: the individual and calendar shifters enter as two-way fixed effects, and the treatment-induced variation in a_{it} and $a_{it}q_{it}$ identifies the engagement slopes. Substituting in Equation (5) gives the indirect (flow) utility

$$v_{it}^{\theta}(a_{it}, q_{it}) = \frac{\beta^{\theta}}{2} e_{it}^2 + \gamma^{\theta}(q_{it})\, a_{it}.$$

Participation. A user joins when the value of access over the T periods exceeds the outside option. Summing the daily flows and adding the network and membership benefits, the value

of joining (which we measure using the willingness to accept to deactivate) is:

$$V_i^\theta = \sum_{t=1}^{T} v_{it}^\theta + \eta \log x + \xi_i$$

where x is the platform's market share and $\eta \log x$ is the network benefit from aggregate participation.

5.1.2 External Parameters

Three sets of objects are not identified by our experiment and are estimated or calibrated from outside sources. We describe these sources below and discuss sensitivity to these parameters in Section 5.3.3.

Network effects. The network term $\eta \log x$ governs how the value of joining responds to aggregate participation. Our experiment holds the platform's overall size fixed and so cannot identify η; we estimate it from the replication data of Bursztyn et al. (2025), comparing each respondent's valuation when the network is intact to the scenario in which two-thirds of it is removed.[27] A non-trivial share of respondents hold negative valuations (they would pay to deactivate), so the relationship is estimated on the inverse-hyperbolic-sine scale. The resulting slope with respondent fixed effects is 1.777 (SE 0.234).

Ad prices. We obtain advertiser prices p^θ per demographic segment (by age, gender, and income) and untargeted ($\overline{p}$) from the Meta Ads Library. This data provides the cross-sectional variation across user types; however, the price level falls well below the platform's realized ad revenue, so we rescale all prices by a common factor that anchors the level to Facebook's 2024 revenue per user while preserving the cross-sectional variation. This factor is $\rho = 4.36$ and is computed as the ratio of average revenue per user to the baseline ad revenue per user implied by the raw prices.[28]

[27] Note that, while we incorporate network effects in our estimates, our measure of user welfare is consumer surplus, and not product market surplus, as defined in Bursztyn et al. (2025).

[28] The per-person ad revenue implied by multiplying baseline ad impressions in our experiment by the raw Ads-Library prices is roughly \$54 per user-year (\$0.15 per user-day). By comparison, Facebook's reported advertising revenue per user in US and Canada summed to \$229.6 in 2023, the last year Meta reported the metric (https://s21.q4cdn.com/399680738/files/doc_financials/2023/q4/Earnings-Presentation-Q4-2023.pdf). Meta discontinued reporting it in 2024; third-party estimates place the 2024 figure near \$233 per user-year (https://businesstats.com/facebooks-average-advertising-revenue-per-user/).

WTA-WTP wedge. Finally, an empirical regularity is that willingness to accept tends to exceed willingness to pay. When mapping valuations to the monetary prices that govern participation under a subscription, we multiply willingness to accept by $\kappa = 0.05$, in line with the median WTP/WTA ratio of Facebook in Sunstein (2020), and apply this only at the counterfactual step.[29] Because willingness to accept is elicited for deactivation over four weeks, we set $T = 28$.

5.1.3 Estimation

The model's structural objects are the two engagement slopes ψ^θ and φ^θ, the welfare curvature β^θ, and the two intrinsic ad-utility parameters γ_0^θ and γ_1^θ. These define utility and are carried into the counterfactuals, together with the membership benefit ξ_i.[30]

We estimate the model sequentially, recovering each parameter from the experimental moment that identifies it.[31] Our baseline specification holds the structural coefficients common across users and user heterogeneity is captured by the membership benefit ξ_i, which comes from the empirical willingness-to-accept distribution. As a robustness check we also estimate a version that allows the engagement slopes to be heterogeneous based on demographics: each user's number of Facebook friends and age, together with desktop usage for the ad-load slope, selected by generalized random-forest variable importance (Athey et al. 2019).

The five identifying moments are the *Hide* and *Replace* intention-to-treat effects on engagement, the cross-sectional relationship between baseline willingness to accept and engagement, and the *Hide* and *Replace* treatment effects on the change in willingness to accept from baseline to endline. We recover the parameters in two steps.

First, the slopes $\psi^\theta/\beta^\theta, \varphi^\theta/\beta^\theta$ come from the engagement equation (6). We estimate it by two-stage least-squares: the *Hide* treatment shifts ad loads, while the *Replace* treatment

[29] Sunstein (2020) documents a median WTP and WTA for a month of Facebook of \$5 and \$87.50, respectively.

[30] The time shifter λ_t, the type constant δ^θ, and the transitory shock ν_{it} are nuisance terms not used in our counterfactuals. The person intercept α_i is absorbed by the individual fixed effects in the slope regression, and enters the counterfactuals only through each user's observed baseline engagement, so it does not need to be separately recovered.

[31] Because the parametric system is just-identified (with five coefficients matched to five moments), the choice of the GMM weighting matrix is irrelevant and the point estimates coincide with those of a minimum-distance estimator. Estimating sequentially makes the mapping from each experimental contrast to each parameter transparent.

shifts quality-weighted exposure holding ad load fixed.[32]

The curvature β^θ is calibrated to the cross-sectional relationship between baseline willingness to accept and engagement, under the assumption that the membership benefit is uncorrelated with engagement. Because high-engagement users may also hold higher membership benefits, this is the one identifying assumption we cannot test directly, and we report sensitivity to β^θ in Section 5.3.3. The intrinsic ad-utility terms $\gamma_0^\theta, \gamma_1^\theta$ are identified from the *Hide* and *Replace* treatment effects on the change in willingness to accept. The membership benefit ξ_i is then the residual of each user's baseline willingness to accept.

We compute standard errors by a user-level block bootstrap with 1,000 repetitions that resamples within treatment arm.

5.2 Parameter Estimates

Table 1 reports the estimated structural parameters in our baseline specification without demographic heterogeneity. The engagement slope point estimates have the expected signs: a higher ad load reduces engagement ($\psi/\beta > 0$), and microtargeting attenuates this crowding-out ($\varphi/\beta > 0$), so targeted ads are less disengaging than untargeted ones. The welfare curvature β is positive, as required for an interior engagement choice. The slopes and the intrinsic ad-utility terms have wide confidence intervals that span zero; the latter imply that ad exposure carries a direct disutility ($\gamma_0 < 0$) that microtargeting partly offsets ($\gamma_1 > 0$). As we show in Section 5.3.3, the counterfactual rankings below are robust to this imprecision, because they are governed by the willingness-to-accept distribution and the curvature rather than by the slopes or the intrinsic ad-utility terms. Appendix Table B17 reports the sensitivity of each estimate to each moment in the sense of Andrews et al. (2017). The engagement slopes and curvature load on the engagement and willingness-to-accept moments that identify them; the intrinsic ad-utility terms load on the endline willingness-to-accept contrasts, with γ_0 identified as the difference between the *Hide* and *Replace* effects. Note that, since φ^θ multiplies both a_{it} and q_{it} in the first-order condition, it is not surprising that the ad-load slope also responds to the *Replace* moments.

[32] Both engagement slopes are estimated on the six-week intervention window, matching the horizon of the endline willingness-to-accept elicitation. Because the engagement response to de-targeting builds over time, we consider robustness to re-estimating both slopes on the full nineteen-week panel.

Table 1: Structural Parameter Estimates

Parameter	Description	Estimate	95% CI	Units
Estimated parameters				
ψ/β	Ad-load engagement slope	58	$[-268, 451]$	items / ad load
φ/β	Targeting engagement slope	21	$[-226, 346]$	items / ad load
β	Welfare curvature	1.74	$[0.9, 3.0]$	\$/day/item^2 ($\times 10^{-5}$)
γ_0	Intrinsic ad disutility	-9.5	$[-30, 12]$	\$/day per ad load
γ_1	Targeting ad utility	4.9	$[-16, 25]$	\$/day per ad load
Externally estimated / calibrated parameters				
η	Network effect	1.777	$[1.318, 2.236]$	asinh, platform size
κ	WTP–WTA ratio	0.05	—	ratio
ρ	Ad-price rescaling	4.36	—	ratio

Note: This table presents structural parameter estimates and calibrated parameters. Brackets are 95% percentile confidence intervals from a user-level block bootstrap with 1,000 draws.

5.3 Counterfactual Simulations

5.3.1 Description

We use the estimated model to compare three monetization regimes: the observed advertising model with personalized ad loads and microtargeting, a subscription model with a uniform fee and no ads, and a hybrid "opt-in" regime in which users choose between an ad-supported tier and a paid ad-free tier. For each regime we evaluate profits, engagement, consumer surplus, and platform demand (the share of users with positive consumer surplus).

The counterfactual construction proceeds in three steps. First, for each user we hold the membership benefit ξ_i fixed—absorbed into the observed baseline willingness to accept—and add the model-predicted change in flow utility implied by the counterfactual ad load and match quality, evaluating the join value V_i^θ at the new regime. Ad revenue is priced on impressions, $p^\theta \, a \, e_i(a, q)$, where $e_i(a, q)$ is the model-predicted engagement from (6).

Second, for consumer surplus and platform demand, we allow willingness to accept to depend on aggregate participation to capture network effects. We map each user's valuation across platform sizes on the inverse-hyperbolic-sine scale using the calibrated η, and convert valuations to monetary prices under a subscription using the willingness-to-pay to willingness-to-accept ratio of $\kappa = 0.05$. Platform demand in each regime is determined as a fixed point: individual participation depends on counterfactual consumer surplus, while aggregate participation feeds back into consumer surplus.

Finally, we aggregate outcomes across individuals and repeat the procedure across 1,000

user-level block bootstrap resamples to obtain confidence intervals for the decomposition in Equation (1).

5.3.2 Results

Business Model Decomposition. Figure 13 reports the effect of switching from the status-quo business model (personalized ad loads with microtargeting) to a subscription-based model. The figure decomposes this total effect into four components: eliminating microtargeting, eliminating ad price differentiation, eliminating ad-load discrimination, and replacing ads with an optimal subscription.

The ad-based business model performs at least as well as the subscription-based alternative across the outcomes we consider. Profits, consumer surplus, and participation are all higher under advertising, and the paywall reduces each across virtually all bootstrap draws; engagement is directionally higher under advertising but not statistically distinguishable.

The mechanisms differ across outcomes, but reduced ad match quality and the switch to monetary pricing are the two dominant channels. Because targeted ads are less disengaging than generic ones, de-targeting is the main source of the engagement decline and the largest single reduction in consumer surplus. The business-model change—replacing advertising with a subscription—is the largest component for profits and for participation: because users are more sensitive to monetary prices than to ad loads, the subscription drives users out. The two channels are comparable in size for consumer surplus and participation, and their split is imprecisely estimated. For engagement, the business-model change compounds rather than offsets the decline: although the users who remain engage more without ads, enough users leave that aggregate engagement falls. Ad-price differentiation is a smaller, profit-specific component, and eliminating ad-load discrimination has little effect anywhere.

We can further decompose the fourth component, in the case of profits, into an intensive-margin term related to the relative revenue per user and an extensive-margin term related to the participation response due to the relative sensitivity to monetary prices compared to advertising loads:

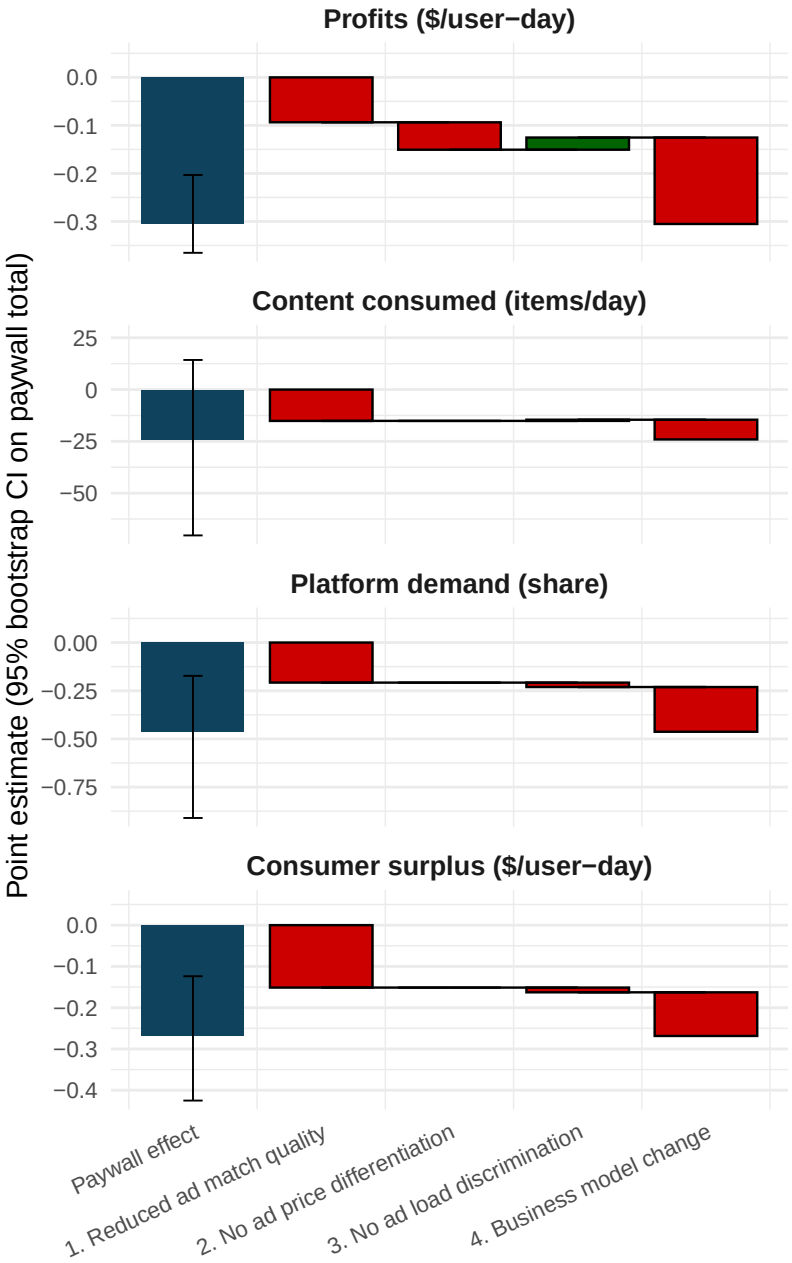


FIGURE 13: COUNTERFACTUAL MONETIZATION DECOMPOSITION

Note: The blue bar shows the point estimate of the total effect of switching from the status-quo business model (microtargeted ads with personalized ad loads and prices) to a subscription-based model, with a 95% percentile bootstrap confidence interval from 1,000 replications. The figure decomposes this total effect into four components: (i) eliminating microtargeting while holding each user's ad load fixed; (ii) eliminating ad-price differentiation by setting all ad prices to their untargeted level; (iii) eliminating ad-load discrimination by moving all users to a common ad load; and (iv) replacing ads with an optimally chosen subscription fee. Positive components are shown in green, negative components in red. Ad-price differentiation affects profits only; it is a transfer on the revenue side and does not enter consumer surplus, demand, or engagement.

$$\underbrace{\pi(\bar{p}, a{=}0, \bar{q}, s^*) - \pi(\bar{p}, \bar{a}, \bar{q}, s{=}0)}_{\text{4. Business model change}} = \underbrace{\left[s^* - \bar{p}\,\bar{a}\,e(\bar{a}, \bar{q})\right] x(\bar{a}, \bar{q}, s{=}0)}_{\text{4a. Intensive margin: Relative revenue per user}}$$

$$+ \underbrace{s^* \left\{ \underbrace{\left[x(0, \bar{q}, s^*) - x(0, \bar{q}, s{=}0)\right]}_{\text{4b. Sensitivity to monetary prices}} - \underbrace{\left[x(\bar{a}, \bar{q}, s{=}0) - x(0, \bar{q}, s{=}0)\right]}_{\text{4c. Sensitivity to ad loads}} \right\}}_{\text{Extensive margin: Participation response}}.$$

Appendix Figure A11 shows that the intensive margin (revenue-per-user) term (4a) accounts for almost the entire business-model change. This is where users' greater sensitivity to monetary prices than to ad loads manifests: because demand falls steeply in monetary price, the profit-maximizing fee is only about $2.75 per four weeks. This is a small fraction of the advertising revenue the same users generate, so each subscriber is worth far less than an ad-supported user. Network effects amplify this, pulling the optimal fee down further. The extensive-margin participation terms are second order at this fee.

Opt-In Policy Lastly, Appendix Figure A12 reports an additional counterfactual in which we consider a hybrid "opt-in" business model, where users choose between the status-quo advertising version of the platform and an ad-free subscription. This option has recently been implemented by Meta in Europe, following regulatory changes that require platforms to offer users a choice between personalized advertising and a paid, ad-free alternative. The figure shows that introducing an opt-in subscription has negligible effects on profits, engagement, platform demand, and consumer surplus relative to the status quo. This follows directly from our decomposition: because users are more sensitive to monetary prices than to ad loads, only a small fraction of users (approximately 1% in our simulations) choose to pay $2 per four weeks for the ad-free version, so the opt-in model leaves aggregate outcomes largely unchanged relative to a purely advertising-based business model.[33]

Overall, these counterfactuals suggest that advertising is a relatively efficient monetization strategy on social media. It generates higher profits and higher consumer surplus, because microtargeted ads are engaging and because users appear more willing to pay with

[33]For reference, at the end of 2024 Meta set the monthly ad-free subscription for *both* Facebook and Instagram to €5.99/month on the web; see https://about.fb.com/news/2024/11/facebook-and-instagram-to-offer-subscription-for-no-ads-in-europe/. In the UK the subscription is £2.99/month for the first Meta account; see https://about.fb.com/news/2025/09/facebook-and-instagram-to-offer-subscription-for-no-ads-in-the-uk/?ref=1.

attention than with money. This helps rationalize the prevalence of ad-based business models on social media platforms.

5.3.3 Robustness

Several structural parameters are imprecisely estimated and others are calibrated from outside sources, so we check whether the counterfactual conclusions depend on them. We vary each parameter one at a time over a plausible range: the 5th–95th percentile of its bootstrap distribution for the estimated parameters (ψ/β, φ/β, β, γ_0, γ_1); $\pm 50\%$ for the estimated network effect η and "content quality" δ; and from the raw Ads-Library price level ($\rho = 1$) to above Facebook's realized revenue per user ($\rho = 6$) for the ad-price rescaling factor ρ. We also vary the WTP/WTA ratio from our baseline $\kappa = 0.05$ to 0.5. For reference, the average WTA/WTP ratio reported in meta-analyses from the contingent valuation literature corresponds to $\kappa = 0.1394$ for public, non-market, and private goods, and $\kappa = 0.3424$ for private goods (Horowitz and McConnell 2002).

Appendix Figure A14 reports the results. The profit ranking is stable across almost the entire exercise: a paywall reduces profits over the full swept range of every estimated parameter, the network effect, and the ad-price factor. Varying the estimated parameters and the network effect, the effect stays between -0.34 and -0.21 per user-day. The intrinsic ad-utility terms γ_0, γ_1 move the magnitude the most, while the engagement slopes and curvature barely move it—consistent with the sensitivity matrix of Table B17. Naturally, the profit level scales linearly with the ad-price factor ρ, from -0.03 at the raw Ads-Library level ($\rho = 1$) to -0.44 at $\rho = 6$, but advertising remains more profitable across the entire plausible range of ad prices.

The one exception is the willingness-to-pay to willingness-to-accept ratio κ, which enters only the subscription counterfactual: ad-regime profit does not depend on it, while subscription profit rises with it, so the profit comparison reverses at a break-even of $\kappa \approx 0.33$. This value lies well above the WTP/WTA ratios measured for Facebook by Sunstein (2020), which were as high as 0.23 (the ratio of average WTP to average WTA among users). The consumer-surplus and participation rankings never reverse at any κ: a higher ratio raises the optimal subscription fee, so consumer surplus under the paywall falls further, and advertising delivers higher consumer surplus by a widening margin.

Next, Appendix Figures A15, A16, and A17 present the counterfactual results allowing for heterogeneity in the engagement slopes. We let the ad-load and targeting slopes ψ^θ/β^θ

and $\varphi^\theta/\beta^\theta$ vary with each user's number of Facebook friends and age, together with desktop usage for the ad-load slope. The main decomposition, its profit sub-decomposition, and the opt-in counterfactual are all essentially unchanged. The conclusions are therefore not an artifact of assuming common coefficients.

Additionally, our main results remain robust if we use the longer-run data (19 weeks) to estimate the engagement slopes: both ψ/β and φ/β are larger, consistent with the engagement response building over time, but every counterfactual total is unchanged, and only the split between the match-quality and business-model components shifts modestly (Appendix Figure A18).

6 Conclusion

As new privacy regulations are rolled out, social media will have to re-evaluate and potentially adjust their business models. In this paper, we evaluate the ad-funded business model of social media platforms under alternative ad targeting ability. We find that ad loads act as prices—personalized ones, with two-thirds of their variation across users—and that engagement is highly inelastic to them, reflecting not inattention but users' ability to scroll past ads at little cost. Users are more sensitive to which ads they see than to how many ads they see: reducing targeting impacts engagement more than hiding ads. Counterfactual simulations suggest that an ad-based business model gives higher short-term profits and consumer surplus relative to a subscription-based one, which helps rationalize the prevalence of ads on social media. Key mechanisms are that microtargeted ads are engaging and that users are much less sensitive to ad loads than to monetary prices, making advertising a relatively efficient revenue source.

References

Acemoglu, Daron, Daniel Huttenlocher, Asuman Ozdaglar, and James Siderius (2024). *Online Business Models, Digital Ads, and User Welfare*. Tech. rep. National Bureau of Economic Research.

Acquisti, Alessandro and Jens Grossklags (2005). "Privacy and rationality in individual decision making". In: *IEEE security & privacy* 3.1, pp. 26–33.

Alcobendas, Miguel, Shunto Kobayashi, Ke Shi, and Matthew Shum (2023). "The impact of privacy protection on online advertising markets". In: *Available at SSRN 3782889*.

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow (2020). "The welfare effects of social media". In: *American economic review* 110.3, pp. 629–676.

Ambrus, Attila, Emilio Calvano, and Markus Reisinger (2016). "Either or both competition: A "two-sided" theory of advertising with overlapping viewerships". In: *American Economic Journal: Microeconomics* 8.3, pp. 189–222.

Anderson, Simon P and Stephen Coate (2005). "Market provision of broadcasting: A welfare analysis". In: *The Review of Economic Studies* 72.4, pp. 947–972.

Anderson, Simon P and André De Palma (2012). "Competition for attention in the information (overload) age". In: *The RAND Journal of Economics* 43.1, pp. 1–25.

Anderson, Simon P. and Martin Peitz (May 2023). "Ad Clutter, Time Use, and Media Diversity". In: *American Economic Journal: Microeconomics* 15.2, pp. 227–70. DOI: 10.1257/mic.20210139. URL: https://www.aeaweb.org/articles?id=10.1257/mic.20210139.

Andrews, Isaiah, Matthew Gentzkow, and Jesse M Shapiro (2017). "Measuring the sensitivity of parameter estimates to estimation moments". In: *The Quarterly Journal of Economics* 132.4, pp. 1553–1592.

Argentesi, Elena, Paolo Buccirossi, Emilio Calvano, Tomaso Duso, Alessia Marrazzo, and Salvatore Nava (2021). "Merger policy in digital markets: An ex post assessment". In: *Journal of Competition Law & Economics* 17.1, pp. 95–140.

Aridor, Guy (2025). "Measuring substitution patterns in the attention economy: An experimental approach". In: *The RAND Journal of Economics* 56.3, pp. 302–324.

Aridor, Guy, Yeon–Koo Che, and Tobias Salz (2023). "The Effect of Privacy Regulation on the Data Industry: Empirical Evidence from GDPR". In: *RAND Journal of Economics* 54.4, pp. 695–730. DOI: 10.1111/1756-2171.12455. URL: https://doi.org/10.1111/1756-2171.12455.

Aridor, Guy, Yeon-Koo Che, Brett Hollenbeck, Maximilian Kaiser, and Daniel McCarthy (2025a). "Evaluating the impact of privacy regulation on e-commerce firms: Evidence from apple's app tracking transparency". In: *Management Science*.

Aridor, Guy, Rafael Jiménez-Durán, Ro'ee Levy, and Lena Song (2025b). "Experiments on social media". In: *Handbook of Experimental Methods in the Social Sciences*. Edward Elgar Publishing.

Aridor, Guy, Rafael Jiménez-Durán, Ro'ee Levy, and Lena Song (2024). "The economics of social media". In: *Journal of Economic Literature* 62.4, pp. 1422–1474.

Athey, Susan, Emilio Calvano, and Joshua S Gans (2018). "The impact of consumer multi-homing on advertising markets and media competition". In: *Management science* 64.4, pp. 1574–1590.

Athey, Susan, Julie Tibshirani, Stefan Wager, et al. (2019). "Generalized random forests". In: *Annals of Statistics* 47.2, pp. 1148–1178.

Banchio, Martino, Francesco Decarolis, Carl-Christian Groh, Rafael Jiménez-Durán, and Miguel Risco (2026). "Addictive Platform Design: Competition, Awareness, and Regulation". Working paper.

Beknazar-Yuzbashev, George, Rafael Jiménez-Durán, Jesse McCrosky, and Mateusz Stalinski (2025). "Toxic content and user engagement on social media: Evidence from a field experiment". In.

Beknazar-Yuzbashev, George, Rafael Jiménez-Durán, and Mateusz Stalinski (2024). "A model of harmful yet engaging content on social media". In: *AEA Papers and Proceedings*. Vol. 114. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, pp. 678–683.

Brynjolfsson, Erik, Avinash Collis, Daniel Deisenroth, Haritz Garro, Daley Kutzman, Asad Liaqat, and Nils Wernerfelt (Dec. 2025). "The Consumer Welfare Effects of Online Ads: Evidence from a Nine-Year Experiment". In: *American Economic Review: Insights* 7.4, pp. 447–62. DOI: 10.1257/aeri.20240452. URL: https://www.aeaweb.org/articles?id=10.1257/aeri.20240452.

Bursztyn, Leonardo, Benjamin Handel, Rafael Jiménez-Durán, and Christopher Roth (Dec. 2025). "When Product Markets Become Collective Traps: The Case of Social Media". In: *American Economic Review* 115.12, pp. 4105–36. DOI: 10.1257/aer.20231468. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20231468.

Calvano, Emilio and Michele Polo (2020). "Strategic differentiation by business models: Free-to-air and pay-TV". In: *The Economic Journal* 130.625, pp. 50–64.

— (2021). "Market power, competition and innovation in digital markets: A survey". In: *Information Economics and Policy* 54, p. 100853.

Cameron, A Colin and Douglas L Miller (2015). "A practitioner's guide to cluster-robust inference". In: *Journal of human resources* 50.2, pp. 317–372.

Candogan, Ozan, Kostas Bimpikis, and Asuman Ozdaglar (2012). "Optimal pricing in networks with externalities". In: *Operations Research* 60.4, pp. 883–905.

Cengiz, Doruk, Arindrajit Dube, Attila Lindner, and Ben Zipperer (2019). "The effect of minimum wages on low-wage jobs". In: *The Quarterly Journal of Economics* 134.3, pp. 1405–1454.

Center, Stigler (2019). *Stigler Committee on Digital Platforms*. Stigler Center.

Driscoll, John C and Aart C Kraay (1998). "Consistent covariance matrix estimation with spatially dependent panel data". In: *Review of economics and statistics* 80.4, pp. 549–560.

Dubé, Jean-Pierre, John G Lynch, Dirk Bergemann, Mert Demirer, Avi Goldfarb, Garrett Johnson, Anja Lambrecht, Tesary Lin, Anna Tuchman, and Catherine E Tucker (2025).

"The intended and unintended consequences of privacy regulation for consumer marketing". In.

European Commission (2025). *Commission finds Apple and Meta in breach of the Digital Markets Act*. https://ec.europa.eu/commission/presscorner/detail/en/ip_25_1085. Accessed: 2026-01-14.

Farronato, Chiara, Andrey Fradkin, and Chris Karr (2024). *Webmunk: A new tool for studying online behavior and digital platforms*. Tech. rep. National Bureau of Economic Research.

Gentzkow, Matthew, Jesse M Shapiro, Frank Yang, and Ali Yurukoglu (2024). "Pricing power in advertising markets: Theory and evidence". In: *American Economic Review* 114.2, pp. 500–533.

Goldfarb, Avi and Verina F Que (2023). "The economics of digital privacy". In: *Annual Review of Economics* 15.1, pp. 267–286.

Goldfarb, Avi and Catherine Tucker (2012). "Shifts in privacy concerns". In: *American Economic Review* 102.3, pp. 349–353.

Goldfarb, Avi and Catherine E Tucker (2011). "Privacy regulation and online advertising". In: *Management science* 57.1, pp. 57–71.

Goli, Ali, Jason Huang, David Reiley, and Nickolai M Riabov (2025a). "Measuring consumer sensitivity to audio advertising: a long-run field experiment on Pandora internet radio". In: *Quantitative Marketing and Economics*, pp. 1–31.

Goli, Ali, David H Reiley, and Hongkai Zhang (2025b). "Personalizing ad load to optimize subscription and ad revenues: Product strategies constructed from experiments on pandora". In: *Marketing Science* 44.2, pp. 327–352.

Goodman, Joseph, Lancelot Henry de Frahan, Justin E. Holz, John A. List, Niall MacMenamin, Evan C. McKay, Magne Mogstad, Sally Sadoff, and Hal Sider (Jan. 2026). *Consumer Demand and Market Competition with Time-Intensive Goods*. Working Paper 34743. National Bureau of Economic Research. DOI: 10.3386/w34743. URL: https://www.nber.org/papers/w34743.

Hoechle, Daniel (2007). "Robust standard errors for panel regressions with cross-sectional dependence". In: *The stata journal* 7.3, pp. 281–312.

Horowitz, John K and Kenneth E McConnell (2002). "A review of WTA/WTP studies". In: *Journal of environmental economics and Management* 44.3, pp. 426–447.

Jimenez-Duran, Rafael (2022). "The Economics of Content Moderation: Theory and Experimental Evidence from Hate Speech on Twitter". In.

Johnson, Garrett A (2024). "4. Economic Research on Privacy Regulation: Lessons from the GDPR and Beyond". In: *The Economics of Privacy*. University of Chicago Press, pp. 97–126.

Katz, Justin and Hunt Allcott (2025). *Digital Media Mergers: Theory and Application to Facebook-Instagram*. Tech. rep. Working paper.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz (2007). "Experimental Analysis of Neighborhood Effects". In: *Econometrica* 75.1, pp. 83–119.

Lin, Tesary (2022). "Valuing intrinsic and instrumental preferences for privacy". In: *Marketing Science* 41.4, pp. 663–681.

Liu, Yi, Pinar Yildirim, and Z John Zhang (2022). "Implications of revenue models and technology for content moderation strategies". In: *Marketing Science* 41.4, pp. 831–847.

Moshary, Sarah (2021). "Sponsored search in equilibrium: evidence from two experiments". In: *Available at SSRN 3903602*.

Pigou, Arthur (1920). *The economics of welfare*. Routledge.

Sunstein, Cass R (2020). "Valuing facebook". In: *Behavioural Public Policy* 4.3, pp. 370–381.

Wernerfelt, Nils, Anna Tuchman, Bradley T Shapiro, and Robert Moakler (2025). "Estimating the value of offsite tracking data to advertisers: Evidence from meta". In: *Marketing Science* 44.2, pp. 268–286.

Wilbur, Kenneth C (2008). "A two-sided, empirical model of television advertising and viewing markets". In: *Marketing science* 27.3, pp. 356–378.

A Figures

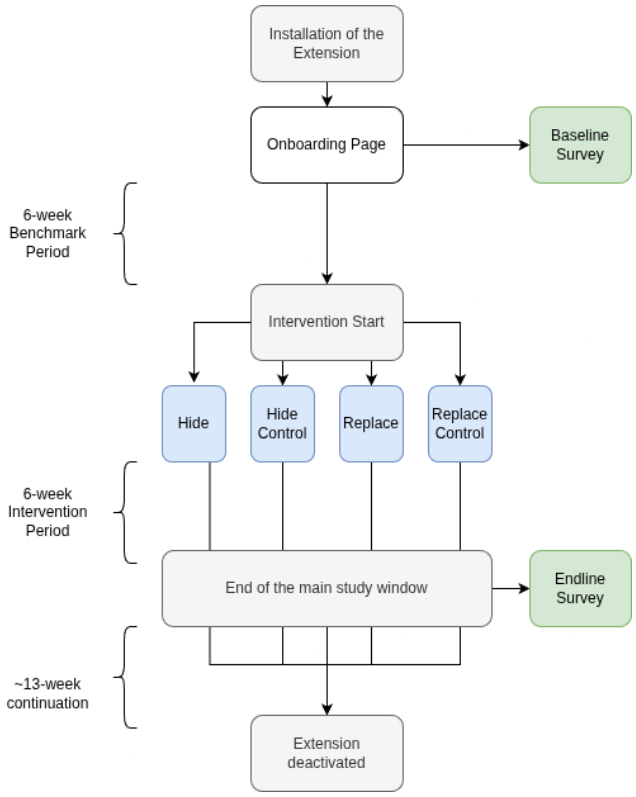
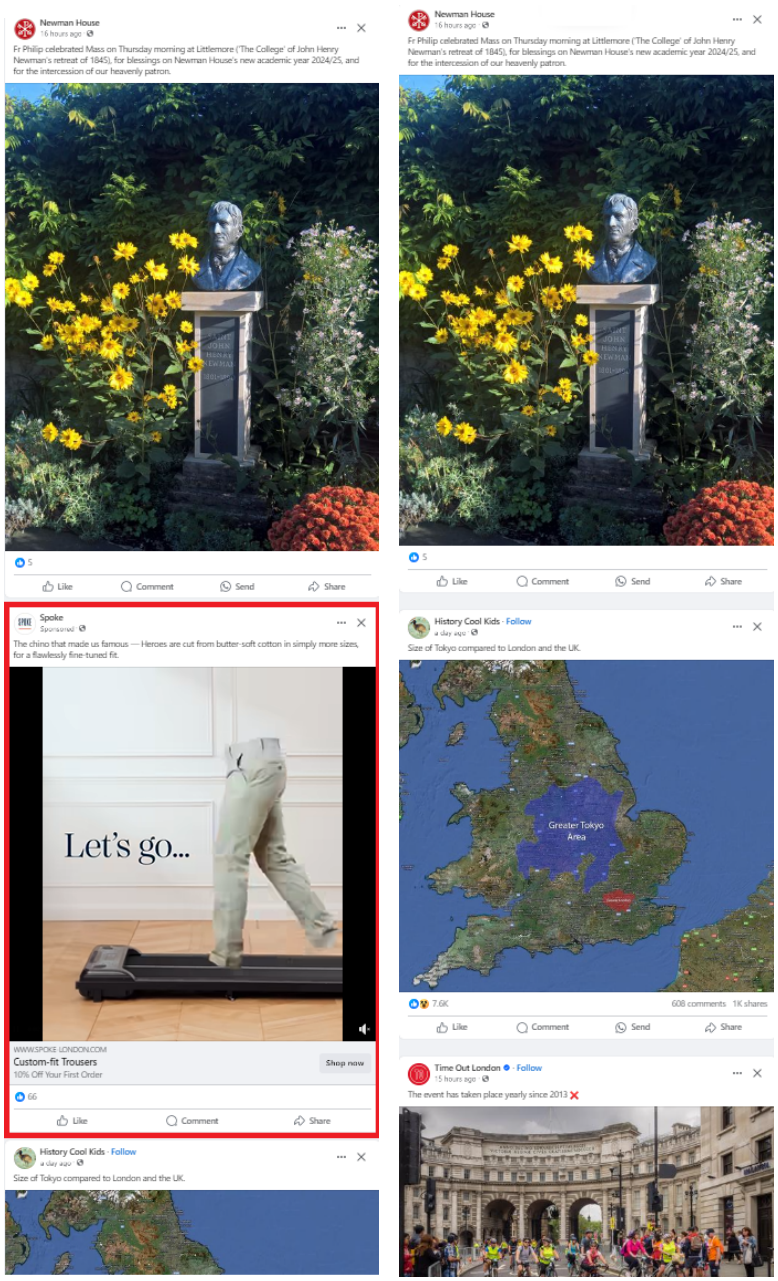


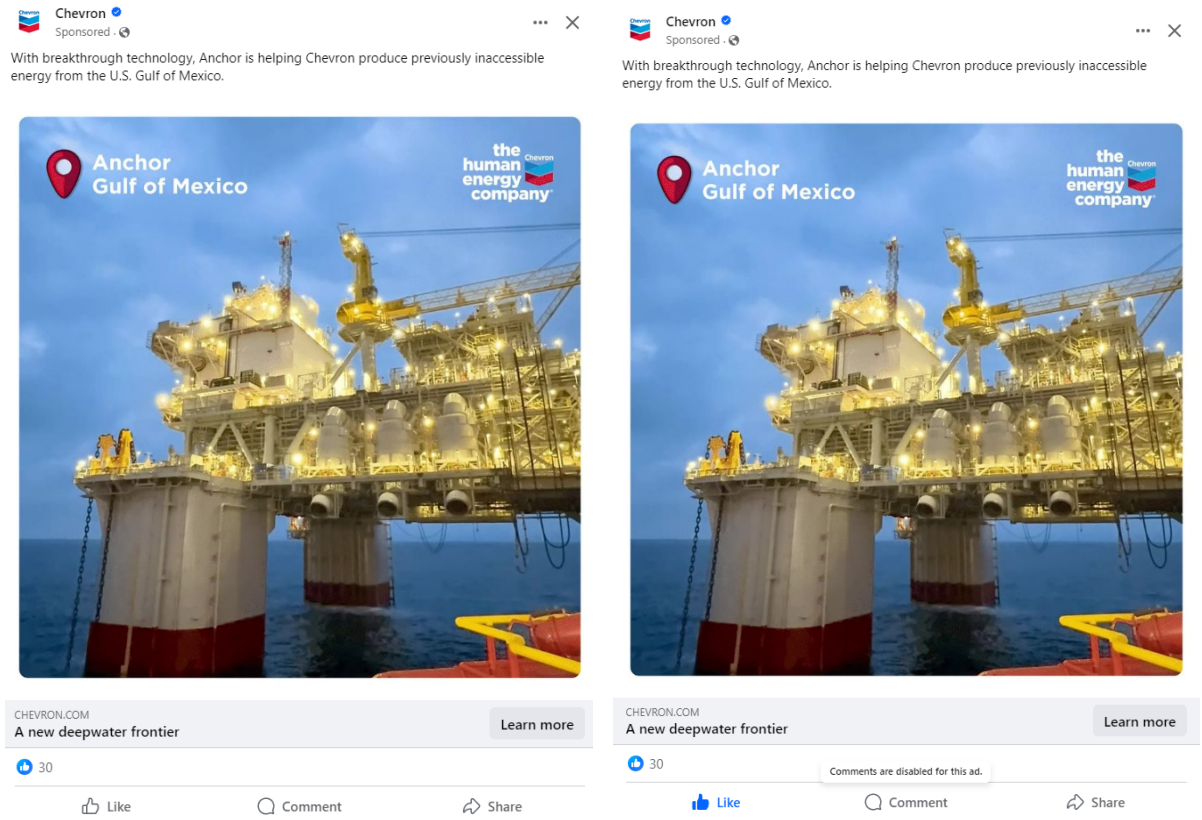
Figure A1: Study Flow



(a) Original Feed (b) Moderated Feed

FIGURE A2: HIDING INTERVENTION

Note: The **left panel** shows an unmoderated feed with a non-ad post, followed by an ad from *Spoke*, a company offering custom-fit garments, and another non-ad post. The ad is framed in red for demonstration purposes. In the **right panel**, we display the moderated feed after the ad-hiding action has been applied. The *Spoke* ad is removed, and the second non-ad post is pulled upward, along with the content below it. As a result, a third non-ad post, from *Time Out London*, becomes visible.



(a) Replacement Ad (b) Possible Interactions

FIGURE A3: REPLACEMENT INTERVENTION

Note: The **left panel** of the figure shows an example of a replacement ad used in the replacement intervention. The **right panel** illustrates potential interactions with the replacement ad. Users can like the ad or react to it using any available emoji, and their reaction will be displayed naturally, as shown in the figure. If someone attempts to comment on or share the ad, a small pop-up appears, informing them that commenting and sharing are disabled (as depicted in the figure).

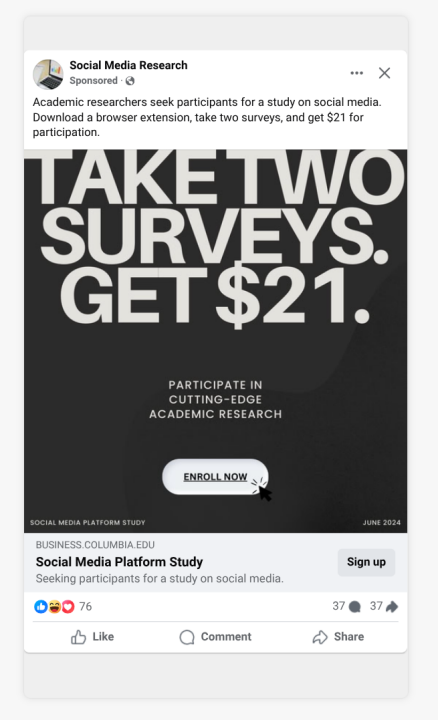
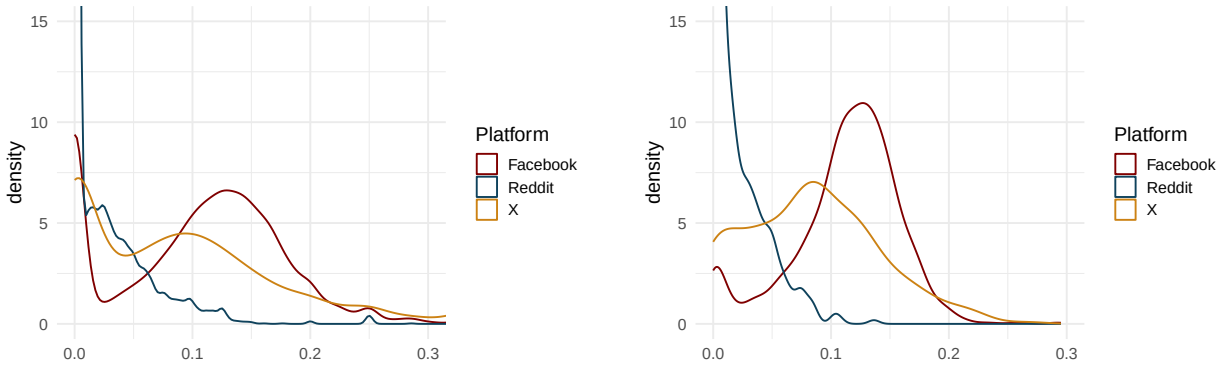


FIGURE A4: EXAMPLE OF A RECRUITMENT AD



(a) Daily Experienced Ad Load

(b) Individual Experienced Ad Load

FIGURE A5: DISTRIBUTION OF AD LOAD BY PLATFORM

Note: **Panel A** presents the distributions of the experienced advertising load per person-day on Facebook, Reddit, and X, based on the baseline period data. **Panel B** presents the equivalent distributions per person. The ad load is defined as the number of ads as a share of total content (ads, posts, and comments) shown to the user on each platform during periods outside of any treatment. The x-axis for Panel A is truncated at 0.3 and the y-axis is truncated at 30. Panel B relies on the subsample of users who viewed at least 50 posts and comments throughout the relevant period.

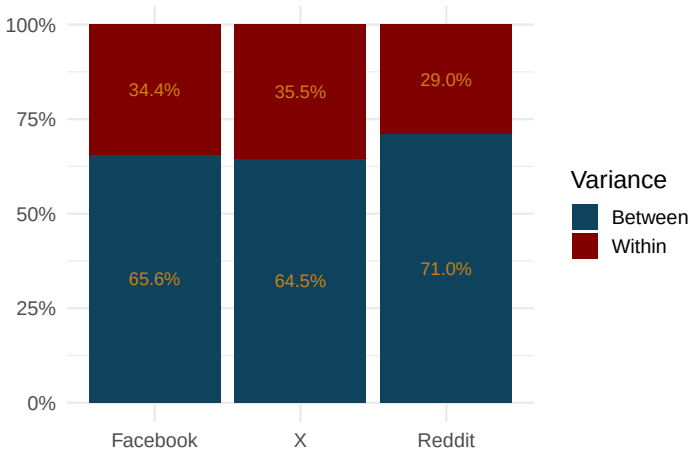


FIGURE A6: AD LOADS: VARIANCE DECOMPOSITION

Note: The figure reports the results of a variance decomposition of daily feed ad load into a persistent between-user component and a within-user day-to-day component for each social media platform (Facebook, X, and Reddit). Estimates come from the random-intercept specification $\text{FeedAdLoad}_{it} = \alpha + u_i + \varepsilon_{it}$, estimated by restricted maximum likelihood, with independent $u_i \sim N(0, \sigma_u^2)$ and $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$; the between-user share is $\sigma_u^2/(\sigma_u^2+\sigma_\varepsilon^2)$. The figure is based on untreated user-day observations with at least 50 feed elements observed on that day.

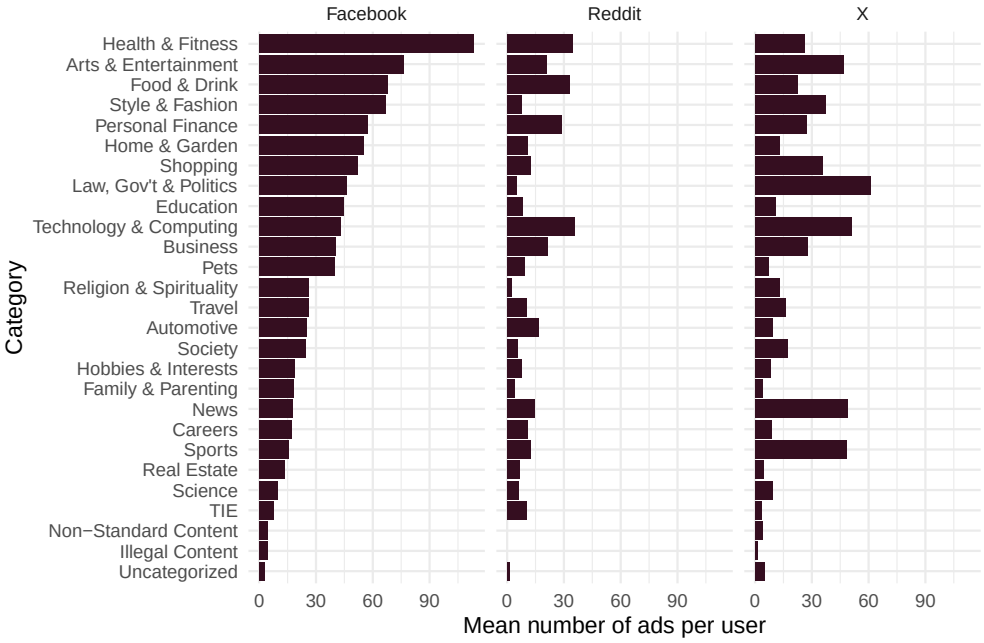
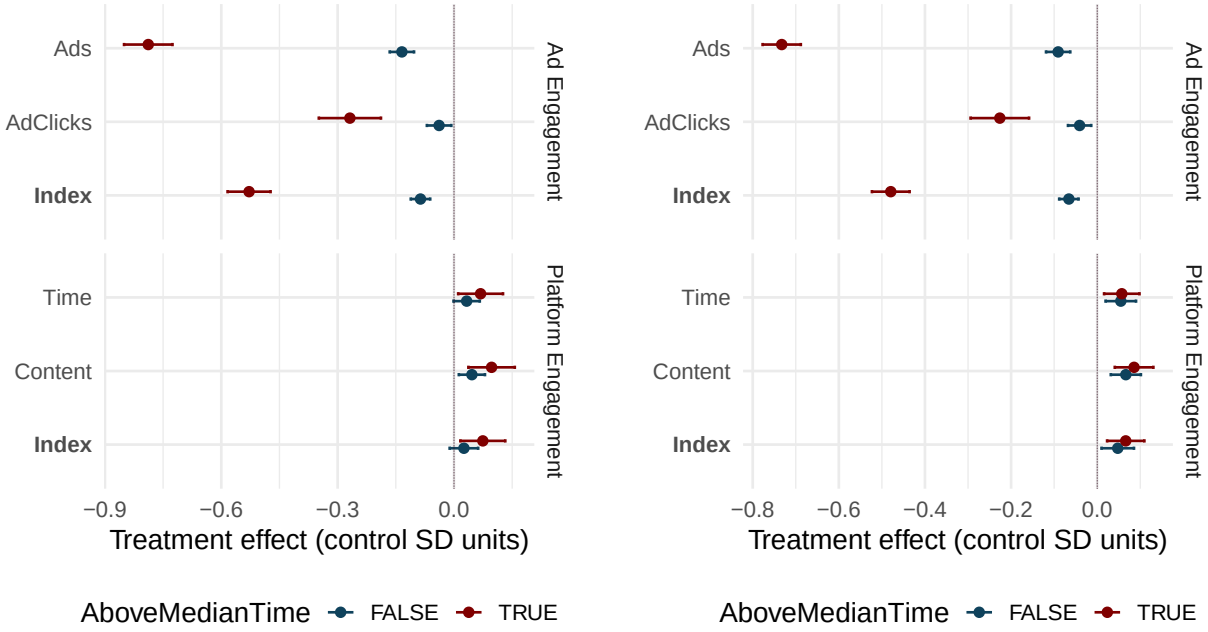


FIGURE A7: AD CATEGORIES BY PLATFORM

Note: The figure reports the mean number of ads by topic category per user, separately for Facebook, Reddit, and X. The categories are sorted by their frequency in our Facebook data. The categorization was performed using generative AI. The computations are based on subsamples of people who watched at least 50 ads on each of the relevant platforms.

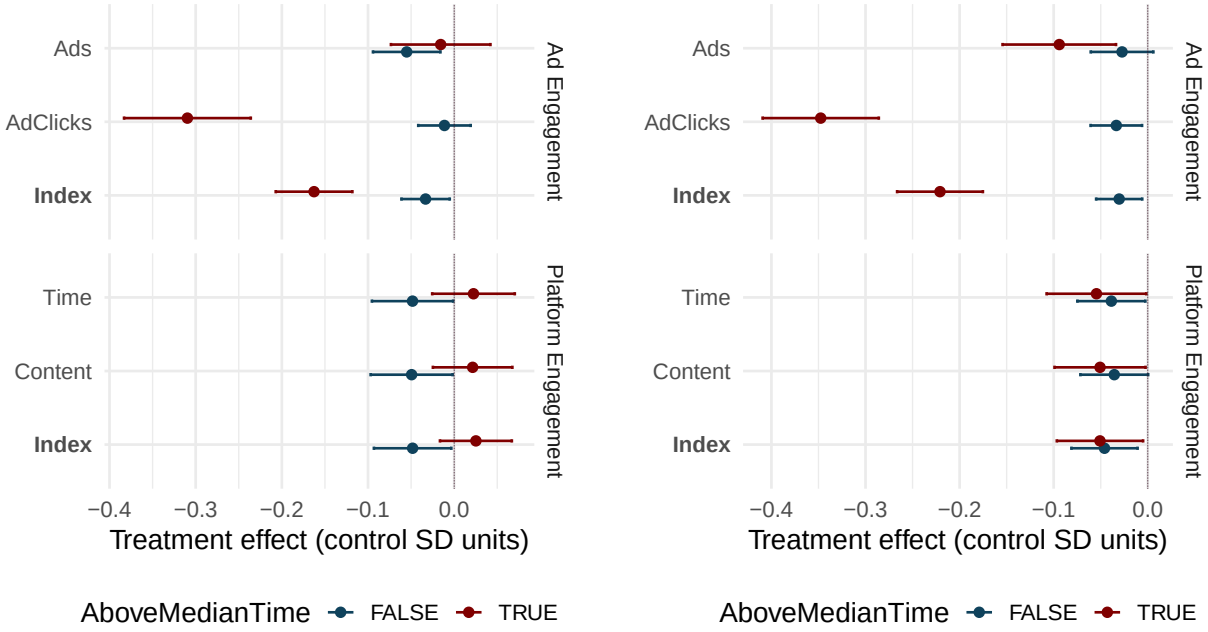


(a) Short-Term Effects

(b) Long-Term Effects

FIGURE A8: HETEROGENEITY OF ENGAGEMENT EFFECTS: HIDING (ASINH)

Note: The figure displays estimated treatment effects of the hiding intervention on Facebook relative to *Hide Control* separately for users whose baseline active time on Facebook was above and below the median. The estimation is based on Equation 4, with $\beta + \phi$ representing the above-median coefficient and β representing the below-median coefficient. The regression relies on the main experimental sample. The dependent variables are: (i) number of ad impressions, (ii) number of ad clicks, (iii) an ad engagement index based on (i) and (ii), (iv) active time spent on the platform, (v) number of posts and comments shown to the user, and (vi) a platform engagement index based on (iv) and (v). **Panel A** shows results based on the six-week intervention period, whereas **Panel B** shows results based on a longer intervention period (134 day-periods or 19.1 weeks). All outcomes are transformed using the inverse hyperbolic sine (asinh) function; for index variables, the transformation is applied prior to standardization. The unit of observation is the individual-day, measured relative to the intervention date. We include 95% confidence intervals based on Driscoll-Kraay standard errors. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day.



(a) Short-Term Effects

(b) Long-Term Effects

FIGURE A9: HETEROGENEITY OF ENGAGEMENT EFFECTS: REPLACEMENT (ASINH)

Note: The figure displays estimated treatment effects of the replacement intervention on Facebook relative to *Replace Control* separately for users whose baseline active time on Facebook was above and below the median. The estimation is based on Equation 4, with $\beta + \phi$ representing the above-median coefficient and β representing the below-median coefficient. The regression relies on the main experimental sample. The dependent variables are: (i) number of ad impressions, (ii) number of ad clicks, (iii) an ad engagement index based on (i) and (ii), (iv) active time spent on the platform, (v) number of posts and comments shown to the user, and (vi) a platform engagement index based on (iv) and (v). **Panel A** shows results based on the six-week intervention period, whereas **Panel B** shows results based on a longer intervention period (134 day-periods or 19.1 weeks). All outcomes are transformed using the inverse hyperbolic sine (asinh) function; for index variables, the transformation is applied prior to standardization. The unit of observation is the individual-day, measured relative to the intervention date. We include 95% confidence intervals based on Driscoll-Kraay standard errors. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day.

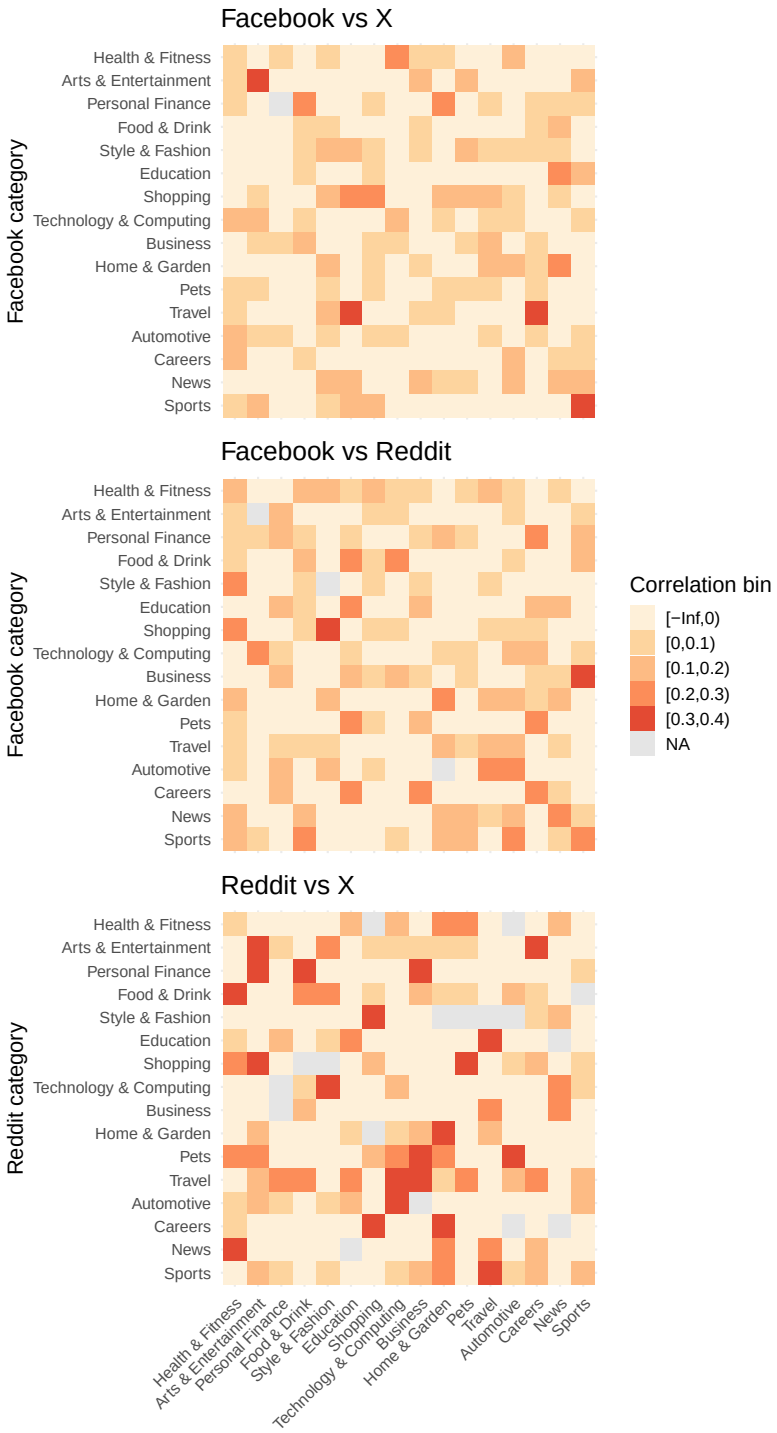


FIGURE A10: CORRELATIONS IN AD CATEGORIES BETWEEN PLATFORMS

Note: This figure presents heatmaps of correlations in topic-based ad targeting across platforms. Ads are first classified into topic categories from the IAB Content Taxonomy (see Section 4.1.3). For each user and platform, we compute the share of ads shown in each topic category and divide by the platform-level category share to obtain relative topic shares. Each cell reports the correlation between relative topic shares for a given pair of categories across two platforms. Lighter colors indicate lower correlations and darker colors indicate higher correlations. The computations are based on subsamples of people who watched at least 50 ads on each of the relevant platforms, and we limit the comparison to ad categories that appear at least 300 times on each of the platforms.

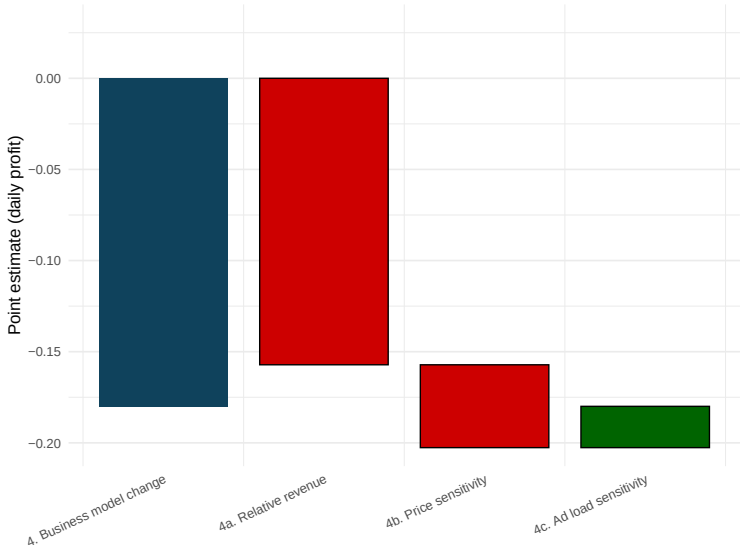


FIGURE A11: DECOMPOSITION OF THE BUSINESS MODEL CHANGE CHANNEL

Note: This figure decomposes the business-model change (blue) into relative revenue, price sensitivity, and ad-load sensitivity components. Mean effects across 1000 bootstrap replications are shown. Positive components are shown in green, and negative components in red.

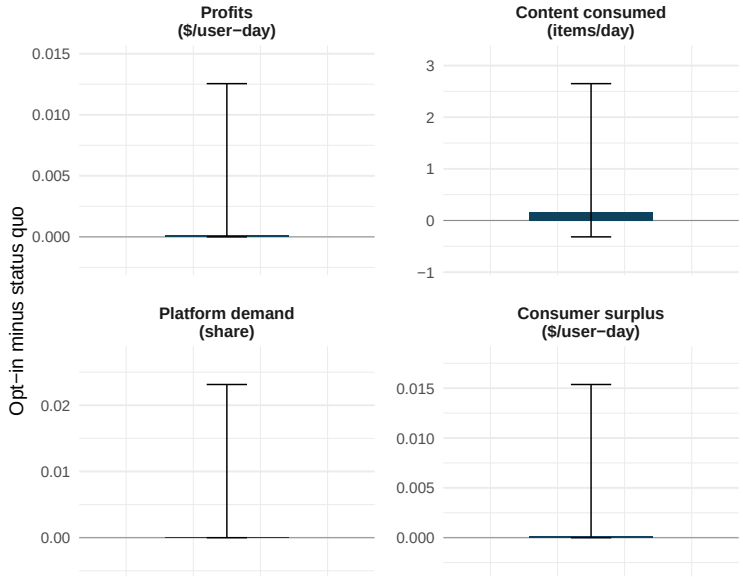


FIGURE A12: COMPARISON OF PAYWALL AND OPT-IN EFFECTS

Note: This figure compares the status-quo advertising-based business model to 1) a subscription-based model and 2) a hybrid "opt-in" model in which users may choose between the ad-supported platform and an ad-free subscription. Bars report mean effects across 1000 bootstrap replications, with 95% confidence intervals.

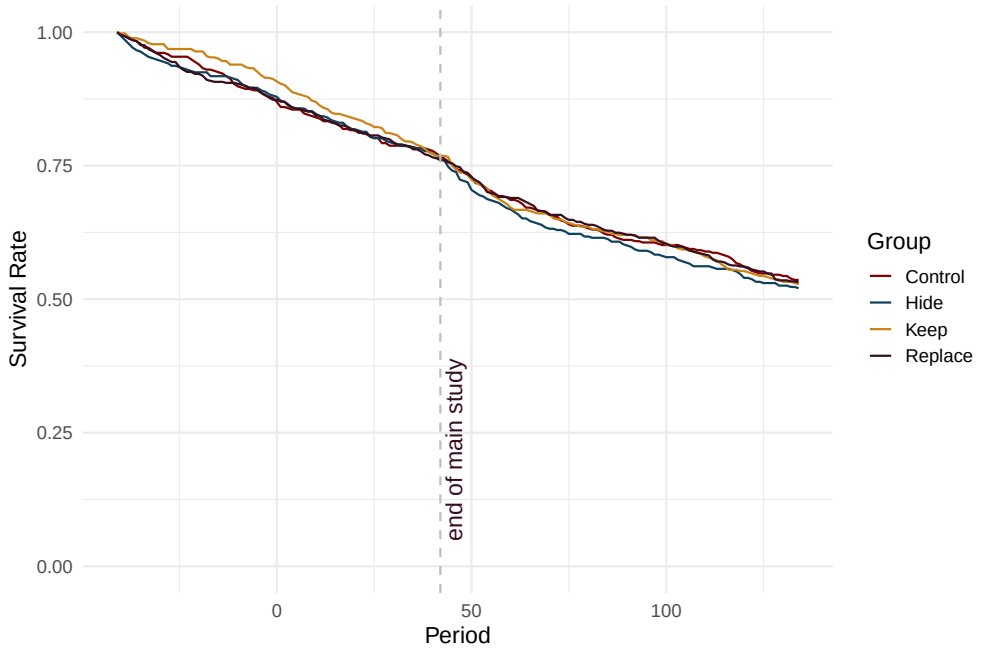


FIGURE A13: SURVIVAL RATE IN THE STUDY

Note: The figure presents the share of users enrolled in the study who survived until a particular study period, measured as the number of days relative to the intervention start date (which starts in period 1). We say that a particular user survived until period t if they were active on their browser on the day corresponding to period t or later. The survival shares are split by treatment conditions.

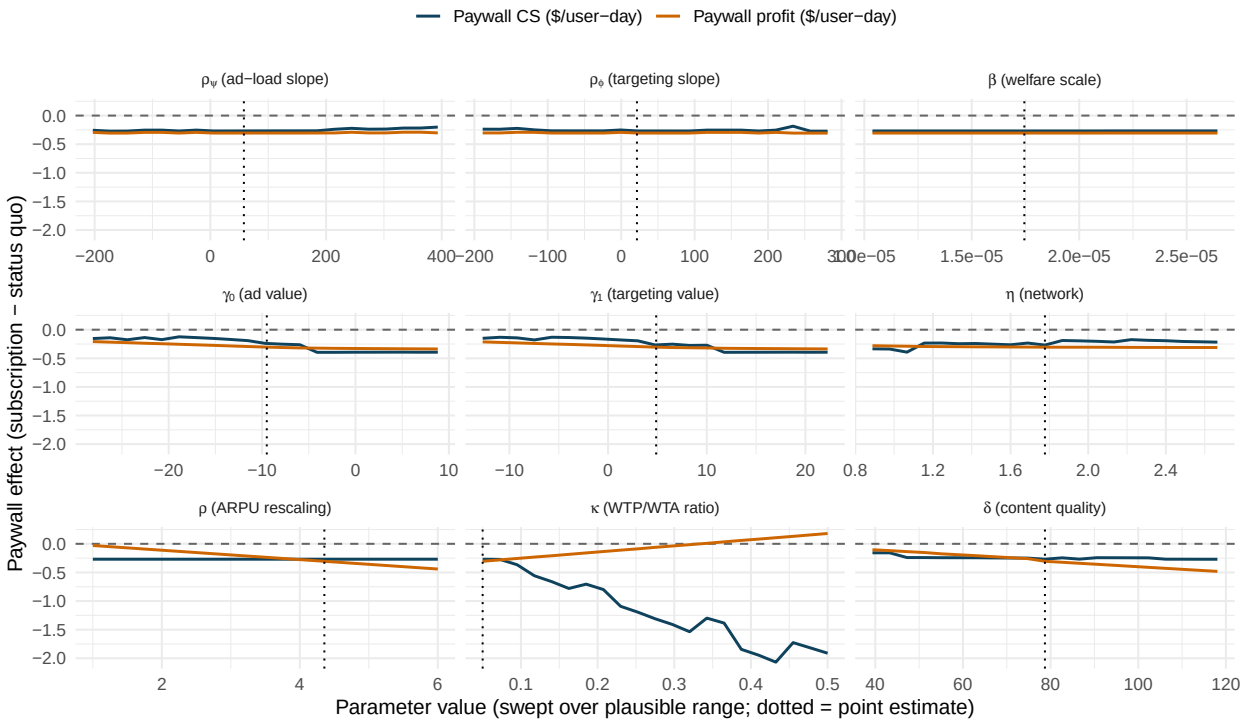


FIGURE A14: SENSITIVITY OF THE COUNTERFACTUAL TO PARAMETERS

Note: Each panel varies one structural or calibrated parameter over its plausible range (5th–95th bootstrap percentile for the estimated parameters; ±50% for the network effect η and content quality δ), holding the others at their point estimates, and reports the paywall effect on profits (and consumer surplus).

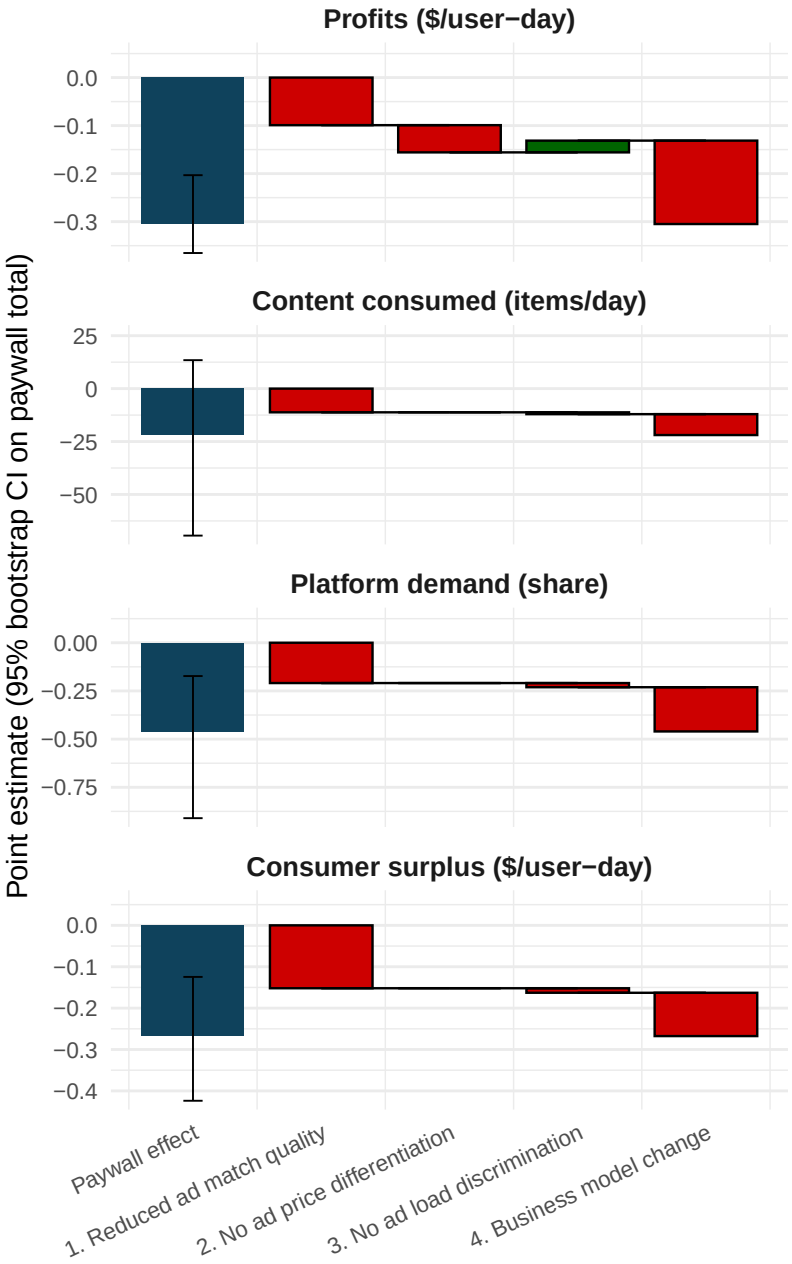


FIGURE A15: COUNTERFACTUAL MONETIZATION DECOMPOSITION (HETEROGENEOUS SLOPES)

Note: As Figure 13, but allowing the engagement slopes $\psi^\theta/\beta^\theta, \varphi^\theta/\beta^\theta$ to vary with observable demographics (number of Facebook friends, age, and—for the ad-load slope—desktop usage), selected by generalized random-forest variable importance. $B = 1{,}000$ bootstrap replications.

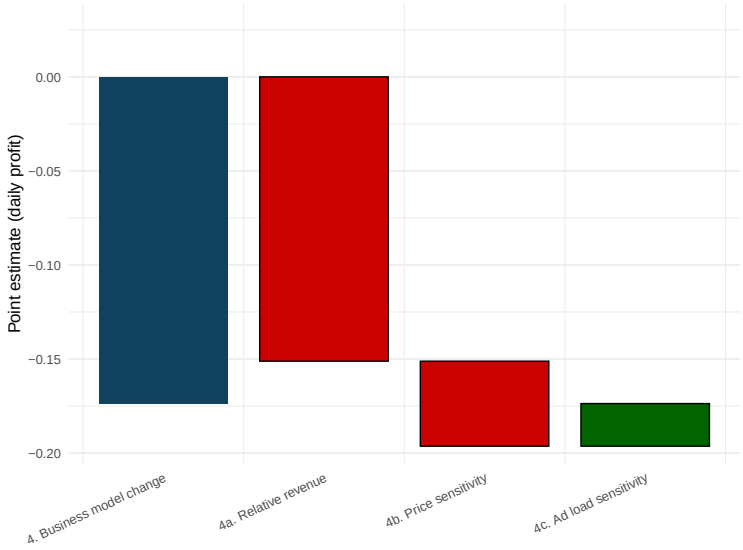


FIGURE A16: PROFIT SUB-DECOMPOSITION OF THE BUSINESS-MODEL CHANGE (HETEROGENEOUS SLOPES)

Note: As Appendix Figure A11, with heterogeneous engagement slopes.

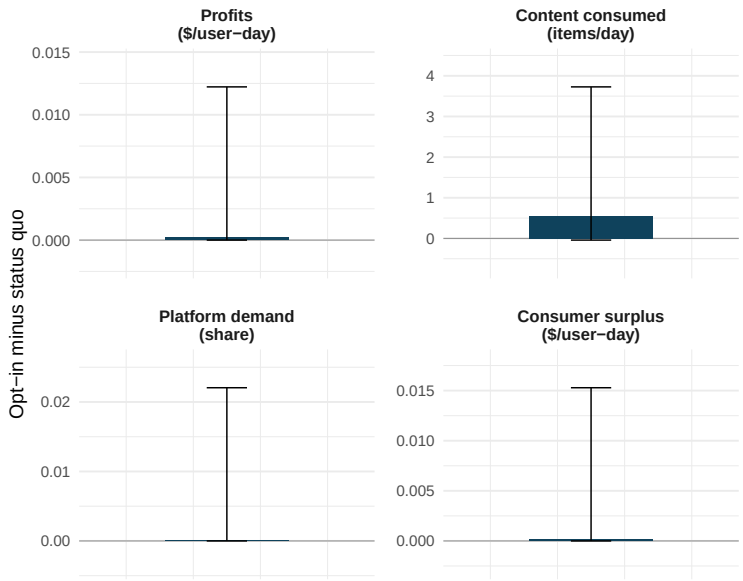


FIGURE A17: OPT-IN AD-FREE TIER VERSUS STATUS QUO (HETEROGENEOUS SLOPES)

Note: As Appendix Figure A12, with the engagement slopes allowed to vary with demographics. $B = 1{,}000$ bootstrap replications.

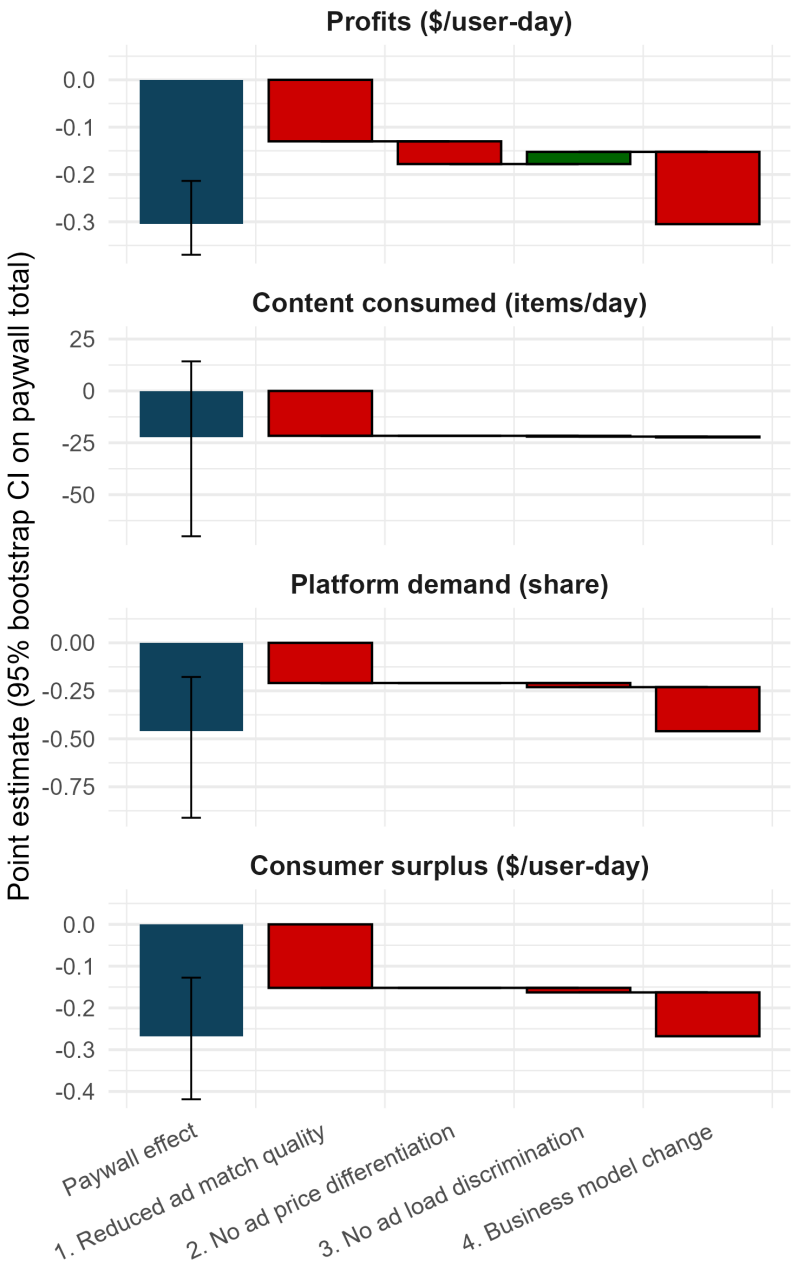


FIGURE A18: COUNTERFACTUAL DECOMPOSITION: NINETEEN-WEEK DATA

Note: Point estimates of the paywall effect (subscription minus status quo) and its ordered decomposition, with both engagement slopes estimated on the full nineteen-week panel instead of the six-week intervention window. 200 bootstrap replications; 95% percentile confidence interval on the total.

B Tables

Table B1: Balance Table: Hide vs. Hide Control

	Control (N=414)		Treatment (N=413)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	p
Age	43.9	13.0	44.6	13.4	0.7	0.446
Male	0.4	0.5	0.3	0.5	0.0	0.829
White	0.7	0.5	0.7	0.5	0.0	0.913
Bachelors	0.4	0.5	0.3	0.5	0.0	0.261
Income 50k+	0.4	0.5	0.4	0.5	0.0	0.399
Democrat	0.6	0.5	0.6	0.5	0.0	0.926
Facebook Desktop Usage	52.0	23.2	49.4	23.1	−2.6	0.121
Baseline Time on FB	14.3	22.5	15.8	34.4	1.4	0.474
Baseline Content on FB	73.4	132.7	80.9	185.0	7.4	0.507
Baseline Ads on FB	8.6	17.2	8.1	16.6	−0.4	0.715
Baseline Ad Clicks on FB	0.2	0.4	0.1	0.3	0.0	0.191

Note: This table compares user characteristics between the Hide treatment group and the Hide Control group in the main experimental sample. It reports group means, standard deviations, differences in means, and p-values from two-sided tests of differences in means. The variables are reported in the following order: (1) age (in years), (2) an indicator for male, (3) an indicator for identifying as white/Caucasian, (4) an indicator for having at least a bachelor's degree, (5) an indicator for having household income above $50,000, (6) an indicator for being a Democrat, (7) desktop share of Facebook usage (from the baseline survey), (8) number of Facebook friends (from the baseline survey), (9) baseline active time on Facebook (minutes per day), (10) baseline number of posts and comments consumed per day on Facebook, (11) baseline number of ad impressions per day on Facebook, and (12) baseline number of ad clicks per day on Facebook. The p-values are based on regressions estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day.

Table B2: Balance Table: Replace vs. Replace Control

	Control (N=445)		Treatment (N=538)			
	Mean	Std. Dev.	Mean	Std. Dev.	Diff. in Means	p
Age	44.0	12.8	45.5	13.2	1.5	0.073
Male	0.3	0.5	0.3	0.5	0.0	0.374
White	0.7	0.4	0.8	0.4	0.0	0.756
Bachelors	0.4	0.5	0.4	0.5	0.0	0.414
Income 50k+	0.4	0.5	0.4	0.5	0.0	0.560
Democrat	0.6	0.5	0.6	0.5	0.0	0.428
Facebook Desktop Usage	50.8	22.7	51.3	24.1	0.5	0.764
Baseline Time on FB	15.7	27.3	15.6	29.6	−0.1	0.961
Baseline Content on FB	83.3	180.9	82.3	165.1	−1.0	0.927
Baseline Ads on FB	9.8	24.1	8.9	19.3	−0.9	0.524
Baseline Ad Clicks on FB	0.1	0.5	0.2	1.0	0.1	0.253

Note: This table compares user characteristics between the Replace treatment group and the Replace Control group in the main experimental sample. It reports group means, standard deviations, differences in means, and p-values from two-sided tests of differences in means. The variables are reported in the following order: (1) age (in years), (2) an indicator for male, (3) an indicator for identifying as white/Caucasian, (4) an indicator for having at least a bachelor's degree, (5) an indicator for having household income above $50,000, (6) an indicator for being a Democrat, (7) desktop share of Facebook usage (from the baseline survey), (8) number of Facebook friends (from the baseline survey), (9) baseline active time on Facebook (minutes per day), (10) baseline number of posts and comments consumed per day on Facebook, (11) baseline number of ad impressions per day on Facebook, and (12) baseline number of ad clicks per day on Facebook. The p-values are based on regressions estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day.

Table B3: Short-Term Effects of Interventions on Advertising Loads

	User-Side Outcomes		Platform-Side Outcomes	
	Experienced Ad Load	Experienced Advertiser Distance	Supplied Ad Load	Supplied Advertiser Distance
Panel A: Hide vs Pure Control				
Treated	-0.095***	-0.070***	0.003	-0.015
	(0.003)	(0.013)	(0.002)	(0.009)
Mean	0.114	0.997	0.114	0.997
SD	0.0794	0.233	0.0794	0.233
N	22720	12552	22721	14775
Panel B: Replace vs Replace Control				
Treated	0.000	0.108***	-0.000	0.014*
	(0.003)	(0.008)	(0.003)	(0.007)
Mean	0.112	0.999	0.112	0.999
SD	0.0848	0.234	0.0848	0.234
N	28302	16919	28302	16907

Note: **Panel A** presents estimated treatment effects of the hiding intervention on Facebook (relative to Hide Control), based on Equation 3 and our main experimental sample. **Panel B** reports corresponding estimates for the replacement intervention (relative to Replace Control). The dependent variables are: (i) experienced ad load, advertisement as a share of all content viewed by the user (ads, posts, and comments), (ii) experienced advertising distance, (iii) supplied ad load, advertisement as a share of all content the user was targeted with (ads, posts, and comments), before any filtering was applied, and (iv) supplied advertiser distance, based on ads that the user was targeted with, before any filtering was applied. Ad distance is defined as the Euclidean distance between a user's characteristics and the average demographic profile targeted by that advertiser, based on baseline data. The unit of observation is the individual-day, where day is measured relative to the intervention date. Driscoll-Kraay standard errors are parenthesized. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day. $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table B4: Long-Term Effects of Interventions on Advertising Loads

	User-Side Outcomes		Platform-Side Outcomes	
	Experienced Ad Load	Experienced Advertiser Distance	Supplied Ad Load	Supplied Advertiser Distance
Panel A: Hide vs Pure Control				
Treated	-0.096***	-0.063***	0.003	-0.019**
	(0.002)	(0.011)	(0.002)	(0.009)
Mean	0.116	1.00	0.116	1.00
SD	0.0837	0.237	0.0837	0.237
N	36183	17100	36184	22329
Panel B: Replace vs Replace Control				
Treated	0.001	0.100***	0.001	0.012*
	(0.003)	(0.006)	(0.002)	(0.006)
Mean	0.116	1.01	0.116	1.01
SD	0.0871	0.236	0.0871	0.236
N	44328	24505	44328	24435

Note: **Panel A** presents estimated treatment effects of the hiding intervention on Facebook (relative to Hide Control), based on Equation 3 and our main experimental sample. **Panel B** reports corresponding estimates for the replacement intervention (relative to Replace Control). The dependent variables are: (i) experienced ad load, advertisement as a share of all content viewed by the user (ads, posts, and comments), (ii) experienced advertising distance, (iii) supplied ad load, advertisement as a share of all content the user was targeted with (ads, posts, and comments), before any filtering was applied, and (iv) supplied advertiser distance, based on ads that the user was targeted with, before any filtering was applied. Ad distance is defined as the Euclidean distance between a user's characteristics and the average demographic profile targeted by that advertiser, based on baseline data. The unit of observation is the individual-day, where day is measured relative to the intervention date. Driscoll-Kraay standard errors are parenthesized. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day. $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table B5: Short-Term Effects of Interventions on Engagement

	Platform Engagement			Ad Engagement		
	Index	Time	Content	Index	Impressions	Clicks
Panel A: Hide vs Pure Control						
Treated	0.026	0.646	5.641	-0.192***	-7.049***	-0.038***
	(0.018)	(0.641)	(3.648)	(0.017)	(0.372)	(0.012)
Mean	0.00	11.26	59.93	0.00	7.44	0.066
SD	0.85	28.08	175.10	0.808	25.28	0.357
N	68712	68712	68712	68712	68712	68712
Panel B: Replace vs Replace Control						
Treated	-0.011	-0.674	-3.738	-0.150***	-1.586***	-0.093***
	(0.012)	(0.521)	(3.063)	(0.015)	(0.493)	(0.010)
Mean	0.00	12.37	64.52	0.00	8.30	0.067
SD	0.852	31.74	204.40	0.796	32.80	0.37
N	82320	82320	82320	82320	82320	82320

Note: **Panel A** presents estimated treatment effects of the hiding intervention on Facebook (relative to Hide Control), based on Equation 3 and our main experimental sample. **Panel B** reports corresponding estimates for the replacement intervention (relative to Replace Control). The dependent variables are: (1) a platform engagement index, constructed from active time spent on the platform and the number of posts and comments shown to the user; (2) active time spent on the platform; (3) number of posts and comments shown to the user; (4) an ad engagement index, constructed from the number of ad impressions and the number of ad clicks; (5) number of ad impressions; and (6) number of ad clicks. The unit of observation is the individual-day, measured relative to the intervention date. Driscoll-Kraay standard errors are reported in parentheses. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day. $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table B6: Long-Term Effects of Interventions on Engagement

	Platform Engagement			Ad Engagement		
	Index	Time	Content	Index	Impressions	Clicks
Panel A: Hide vs Pure Control						
Treated	0.020	0.572	6.632**	-0.167***	-6.397***	-0.039***
	(0.014)	(0.474)	(2.791)	(0.012)	(0.368)	(0.009)
Mean	0.00	11.90	63.08	0.00	7.89	0.073
SD	0.85	30.16	180.03	0.813	26.50	0.422
N	113308	113308	113308	113308	113308	113308
Panel B: Replace vs Replace Control						
Treated	-0.032**	-1.701***	-9.401***	-0.178***	-2.033***	-0.109***
	(0.014)	(0.519)	(3.413)	(0.018)	(0.508)	(0.012)
Mean	0.00	13.09	69.67	0.00	8.90	0.067
SD	0.864	33.66	203.49	0.808	30.11	0.38
N	136924	136924	136924	136924	136924	136924

Note: **Panel A** presents the estimated treatment effects of the hiding intervention on Facebook, relative to the Hide Control group, based on Equation 3. **Panel B** reports the corresponding estimates for the replacement intervention, relative to the Replace Control group. All regressions are based on the extended sample, which involves 19.1 weeks of the intervention period (long-term effects) as opposed to 6 weeks specified in the pre-registration (short-term effects). The dependent variables are: (1) a platform engagement index, constructed from active time spent on the platform and the number of posts and comments shown to the user; (2) active time spent on the platform; (3) number of posts and comments shown to the user; (4) an ad engagement index, constructed from the number of ad impressions and the number of ad clicks; (5) number of ad impressions; and (6) number of ad clicks. The unit of observation is the individual-day, measured relative to the intervention date. Driscoll-Kraay standard errors are reported in parentheses. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day. $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table B7: Short-Term Engagement Effects (Asinh)

	Platform Engagement			Ad Engagement		
	Index	Time	Content	Index	Impressions	Clicks
Panel A: Hide vs Pure Control						
Treated	0.066***	0.117***	0.202***	-0.285***	-0.687***	-0.033***
	(0.024)	(0.042)	(0.057)	(0.014)	(0.032)	(0.005)
Mean	0.00	1.20	1.63	0.00	0.904	0.05
SD	0.976	1.82	2.48	0.841	1.59	0.241
N	68712	68712	68712	68712	68712	68712
Panel B: Replace vs Replace Control						
Treated	-0.021	-0.043	-0.069*	-0.087***	-0.048*	-0.035***
	(0.014)	(0.029)	(0.038)	(0.013)	(0.024)	(0.005)
Mean	0.00	1.22	1.64	0.00	0.883	0.05
SD	0.978	1.85	2.49	0.844	1.59	0.242
N	82320	82320	82320	82320	82320	82320

Note: **Panel A** presents estimated treatment effects of the hiding intervention on Facebook (relative to Hide Control), based on Equation 3 and our main experimental sample. **Panel B** reports corresponding estimates for the replacement intervention (relative to Replace Control). The dependent variables are: (1) a platform engagement index, constructed from active time spent on the platform and the number of posts and comments shown to the user; (2) active time spent on the platform; (3) number of posts and comments shown to the user; (4) an ad engagement index, constructed from the number of ad impressions and the number of ad clicks; (5) number of ad impressions; and (6) number of ad clicks. All outcomes are transformed using the inverse hyperbolic sine (asinh) function; for index variables, the transformation is applied prior to standardization. The unit of observation is the individual-day, measured relative to the intervention date. Driscoll-Kraay standard errors are reported in parentheses. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day. $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table B8: Long-Term Engagement Effects (Asinh)

	Platform Engagement			Ad Engagement		
	Index	Time	Content	Index	Impressions	Clicks
Panel A: Hide vs Pure Control						
Treated	0.054***	0.103***	0.175***	-0.260***	-0.640***	-0.032***
	(0.019)	(0.033)	(0.048)	(0.012)	(0.026)	(0.004)
Mean	0.00	1.22	1.72	0.00	0.951	0.053
SD	0.976	1.84	2.52	0.841	1.62	0.256
N	113308	113308	113308	113308	113308	113308
Panel B: Replace vs Replace Control						
Treated	-0.057***	-0.107***	-0.143***	-0.110***	-0.089***	-0.040***
	(0.013)	(0.026)	(0.032)	(0.015)	(0.025)	(0.005)
Mean	0.00	1.25	1.74	0.00	0.959	0.05
SD	0.978	1.87	2.53	0.839	1.64	0.243
N	136924	136924	136924	136924	136924	136924

Note: **Panel A** presents estimated treatment effects of the hiding intervention on Facebook (relative to Hide Control), based on Equation 3 and our main experimental sample. **Panel B** reports corresponding estimates for the replacement intervention (relative to Replace Control). All regressions are based on the extended sample, which involves 19.1 weeks of the intervention period (long-term effects) as opposed to 6 weeks specified in the pre-registration (short-term effects). The dependent variables are: (1) a platform engagement index, constructed from active time spent on the platform and the number of posts and comments shown to the user; (2) active time spent on the platform; (3) number of posts and comments shown to the user; (4) an ad engagement index, constructed from the number of ad impressions and the number of ad clicks; (5) number of ad impressions; and (6) number of ad clicks. All outcomes are transformed using the inverse hyperbolic sine (asinh) function; for index variables, the transformation is applied prior to standardization. The unit of observation is the individual-day, measured relative to the intervention date. Driscoll-Kraay standard errors are reported in parentheses. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day. $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table B9: Short-Term Engagement Effects (Trimming)

	Platform Engagement			Ad Engagement		
	Index	Time	Content	Index	Impressions	Clicks
Panel A: Hide vs Pure Control						
Treated	0.057***	1.604**	11.155***	-0.277***	-6.190***	-0.055***
	(0.020)	(0.622)	(3.598)	(0.014)	(0.373)	(0.006)
Mean	0.00	10.90	54.47	0.00	6.34	0.051
SD	0.86	27.43	151.92	0.811	18.19	0.297
N	67200	67956	67956	67200	67956	67956
Panel B: Replace vs Replace Control						
Treated	0.007	-0.275	-3.326	-0.081***	-1.140**	-0.037***
	(0.013)	(0.558)	(2.704)	(0.018)	(0.476)	(0.007)
Mean	0.00	11.46	54.47	0.00	6.62	0.064
SD	0.854	29.64	159.06	0.83	22.31	0.352
N	80556	81480	81480	80640	81480	81480

Note: **Panel A** presents estimated treatment effects of the hiding intervention on Facebook (relative to Hide Control), based on Equation 3 and our main experimental sample. **Panel B** reports corresponding estimates for the replacement intervention (relative to Replace Control). All regressions are based on the sample excluding the top 1 percentile of observations—the exclusion is performed for each outcome separately. The dependent variables are: (1) a platform engagement index, constructed from active time spent on the platform and the number of posts and comments shown to the user; (2) active time spent on the platform; (3) number of posts and comments shown to the user; (4) an ad engagement index, constructed from the number of ad impressions and the number of ad clicks; (5) number of ad impressions; and (6) number of ad clicks. The unit of observation is the individual-day, measured relative to the intervention date. Driscoll-Kraay standard errors are reported in parentheses. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day. ∗, ∗∗, and ∗∗∗ denote significance at the 10%, 5%, and 1% levels, respectively.

Table B10: Long-Term Engagement Effects (Trimming)

	Platform Engagement			Ad Engagement		
	Index	Time	Content	Index	Impressions	Clicks
Panel A: Hide vs Pure Control						
Treated	0.047***	1.631***	10.893***	-0.235***	-5.468***	-0.049***
	(0.015)	(0.434)	(2.575)	(0.012)	(0.291)	(0.005)
Mean	0.00	11.30	56.10	0.00	6.57	0.052
SD	0.855	28.93	152.01	0.819	18.74	0.307
N	110508	112106	111863	110377	111837	111808
Panel B: Replace vs Replace Control						
Treated	-0.038**	-1.199**	-9.019***	-0.090***	-1.651***	-0.037***
	(0.015)	(0.551)	(2.950)	(0.018)	(0.495)	(0.008)
Mean	0.00	12.27	63.09	0.00	7.50	0.065
SD	0.865	31.60	177.12	0.824	23.01	0.362
N	134188	135431	135448	134147	135523	135492

Note: **Panel A** presents estimated treatment effects of the hiding intervention on Facebook (relative to Hide Control), based on Equation 3 and our main experimental sample. **Panel B** reports corresponding estimates for the replacement intervention (relative to Replace Control). All regressions are based on the extended sample, which involves 19.1 weeks of the intervention period (long-term effects) as opposed to 6 weeks specified in the pre-registration (short-term effects). In addition, we exclude the top 1 percentile of observations for each outcome. The dependent variables are: (1) a platform engagement index, constructed from active time spent on the platform and the number of posts and comments shown to the user; (2) active time spent on the platform; (3) number of posts and comments shown to the user; (4) an ad engagement index, constructed from the number of ad impressions and the number of ad clicks; (5) number of ad impressions; and (6) number of ad clicks. The unit of observation is the individual-day, measured relative to the intervention date. Driscoll-Kraay standard errors are reported in parentheses. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day. ∗, ∗∗, and ∗∗∗ denote significance at the 10%, 5%, and 1% levels, respectively.

Table B11: Main Attrition Analysis

	Survived						Last Period					
	Hiding	Replacing	Controls	Hiding	Replacing	Controls	Hiding	Replacing	Controls	Hiding	Replacing	Controls
(Intercept)	0.768***	0.769***	0.768***	−0.065	0.115	−0.065	31.502***	33.252***	31.502***	−13.681	1.813	−13.681
	(0.021)	(0.020)	(0.021)	(0.164)	(0.160)	(0.164)	(1.088)	(0.917)	(1.088)	(9.088)	(7.514)	(9.088)
GroupTreatment	−0.005	−0.008	0.000	0.010	−0.177	0.180	−0.168	−1.964	1.749	2.682	−11.561	15.494
	(0.030)	(0.027)	(0.029)	(0.233)	(0.219)	(0.229)	(1.565)	(1.341)	(1.423)	(12.552)	(11.080)	(11.793)
SurvivalScore				1.099***	0.846***	1.099***				59.655***	40.697***	59.655***
				(0.203)	(0.198)	(0.203)				(11.069)	(9.073)	(11.068)
GroupTreatment×SurvivalScore				−0.023	0.224	−0.253				−3.604	12.725	−18.959
				(0.289)	(0.270)	(0.283)				(15.240)	(13.388)	(14.313)
N	827	983	859	777	918	807	827	983	859	777	918	807

Note: Columns 1-6 report individual-level regressions of an indicator for survival until the end of the intervention period (period 42, i.e., 42 days after the start of the intervention) on the treatment indicator. Columns titled "Hiding" pertain to the comparison between Hiding and Hide Control. Columns titled "Replacing" pertain to the comparison between Replacement and Replace Control. Columns titled "Controls" pertain to the comparison between Replace Control and Hide Control. In columns 4-6, we additionally include the survival propensity based on user characteristics—computed by regressing survival on age, male indicator, white indicator, indicator for obtaining a bachelor's degree, indicator for being a Democrat, share of desktop usage of Facebook, active time spent on Facebook during the baseline period, content consumption on Facebook during the baseline, number of ad impressions during the baseline, and baseline number of ad clicks. Columns 7-12 report analogous regressions, with the dependent variable being the last period of recorded activity, which we use as a proxy for the last period in the study. The dependent variable is capped at 42, which indicates the last period of the six-week intervention. Note that the intervention starts in period 1, with baseline periods represented by negative numbers and zero. Standard errors clustered at the individual level are reported in parentheses. ∗, ∗∗, and ∗∗∗ denote significance at the 10%, 5%, and 1% levels, respectively.

Table B12: Long-Term Attrition Analysis

	Hiding	Replacing	Controls	Hiding	Replacing	Controls
(Intercept)	137.836***	141.915***	137.836***	−50.663	−86.308**	−50.663
	(5.262)	(5.092)	(5.262)	(40.514)	(42.084)	(40.510)
GroupTreatment	0.075	−0.820	4.079	−11.330	−1.230	−35.645
	(7.558)	(6.950)	(7.323)	(57.852)	(56.995)	(58.423)
SurvivalScore				248.872***	296.450***	248.872***
				(52.214)	(54.681)	(52.209)
GroupTreatment×SurvivalScore				14.965	0.554	47.578
				(74.512)	(73.573)	(75.614)
N	827	983	859	777	918	807

Note: Columns 1-3 report individual-level regressions of the last period of recorded activity, which we use as a proxy for the last period in the study, on the treatment indicator. Columns titled "Hiding" pertain to the comparison between Hiding and Hide Control. Columns titled "Replacing" pertain to the comparison between Replacement and Replace Control. Columns titled "Controls" pertain to the comparison between Replace Control and Hide Control. In columns 4-6, we additionally include the survival propensity based on user characteristics—computed by regressing survival on age, male indicator, white indicator, indicator for obtaining a bachelor's degree, indicator for being a Democrat, share of desktop usage of Facebook, active time spent on Facebook during the baseline period, content consumption on Facebook during the baseline, number of ad impressions during the baseline, and baseline number of ad clicks. The dependent variable is not capped at any value. Note that the intervention starts in period 1, with baseline periods represented by negative numbers and zero. Standard errors clustered at the individual level are reported in parentheses. ∗, ∗∗, and ∗∗∗ denote significance at the 10%, 5%, and 1% levels, respectively.

Table B13: Short-Term Engagement Effects: Clustered Errors

	Platform Engagement			Ad Engagement		
	Index	Time	Content	Index	Impressions	Clicks
Panel A: Hide vs Pure Control						
Treated	0.026	0.646	5.641	-0.192***	-7.049***	-0.038
	(0.036)	(1.276)	(7.298)	(0.051)	(1.100)	(0.027)
Mean	0.00	11.26	59.93	0.00	7.44	0.066
SD	0.85	28.08	175.10	0.808	25.28	0.357
N	68712	68712	68712	68712	68712	68712
Panel B: Replace vs Replace Control						
Treated	-0.011	-0.674	-3.738	-0.150*	-1.586	-0.093
	(0.032)	(1.277)	(7.147)	(0.082)	(1.096)	(0.057)
Mean	0.00	12.37	64.52	0.00	8.30	0.067
SD	0.852	31.74	204.40	0.796	32.80	0.37
N	82320	82320	82320	82320	82320	82320

Note: **Panel A** presents estimated treatment effects of the hiding intervention on Facebook (relative to Hide Control), based on Equation 3 and our main experimental sample. **Panel B** reports corresponding estimates for the replacement intervention (relative to Replace Control). The dependent variables are: (1) a platform engagement index, constructed from active time spent on the platform and the number of posts and comments shown to the user; (2) active time spent on the platform; (3) number of posts and comments shown to the user; (4) an ad engagement index, constructed from the number of ad impressions and the number of ad clicks; (5) number of ad impressions; and (6) number of ad clicks. The unit of observation is the individual-day, measured relative to the intervention date. Standard errors clustered at the individual level are reported in parentheses. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day. $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table B14: Long-Term Engagement Effects: Clustered Errors

	Platform Engagement			Ad Engagement		
	Index	Time	Content	Index	Impressions	Clicks
Panel A: Hide vs Pure Control						
Treated	0.020	0.572	6.632	-0.167***	-6.397***	-0.039
	(0.030)	(1.092)	(6.156)	(0.041)	(1.005)	(0.024)
Mean	0.00	11.90	63.08	0.00	7.89	0.073
SD	0.85	30.16	180.03	0.813	26.50	0.422
N	113308	113308	113308	113308	113308	113308
Panel B: Replace vs Replace Control						
Treated	-0.032	-1.701	-9.401	-0.178*	-2.033**	-0.109*
	(0.031)	(1.152)	(6.391)	(0.093)	(0.950)	(0.066)
Mean	0.00	13.09	69.67	0.00	8.90	0.067
SD	0.864	33.66	203.49	0.808	30.11	0.38
N	136924	136924	136924	136924	136924	136924

Note: **Panel A** presents estimated treatment effects of the hiding intervention on Facebook (relative to Hide Control), based on Equation 3 and our main experimental sample. **Panel B** reports corresponding estimates for the replacement intervention (relative to Replace Control). All regressions are based on the extended sample, which involves 19.1 weeks of the intervention period (long-term effects) as opposed to 6 weeks specified in the pre-registration (short-term effects). The dependent variables are: (1) a platform engagement index, constructed from active time spent on the platform and the number of posts and comments shown to the user; (2) active time spent on the platform; (3) number of posts and comments shown to the user; (4) an ad engagement index, constructed from the number of ad impressions and the number of ad clicks; (5) number of ad impressions; and (6) number of ad clicks. The unit of observation is the individual-day, measured relative to the intervention date. Standard errors clustered at the individual level are reported in parentheses. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day. $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table B15: Short-Term Engagement Effects: Clustered Errors (Asinh)

	Platform Engagement			Ad Engagement		
	Index	Time	Content	Index	Impressions	Clicks
Panel A: Hide vs Pure Control						
Treated	0.066*	0.117*	0.202**	-0.285***	-0.687***	-0.033**
	(0.037)	(0.070)	(0.091)	(0.045)	(0.076)	(0.013)
Mean	0.00	1.20	1.63	0.00	0.904	0.05
SD	0.976	1.82	2.48	0.841	1.59	0.241
N	68712	68712	68712	68712	68712	68712
Panel B: Replace vs Replace Control						
Treated	-0.021	-0.043	-0.069	-0.087**	-0.048	-0.035**
	(0.032)	(0.060)	(0.081)	(0.037)	(0.050)	(0.014)
Mean	0.00	1.22	1.64	0.00	0.883	0.05
SD	0.978	1.85	2.49	0.844	1.59	0.242
N	82320	82320	82320	82320	82320	82320

Note: **Panel A** presents estimated treatment effects of the hiding intervention on Facebook (relative to Hide Control), based on Equation 3 and our main experimental sample. **Panel B** reports corresponding estimates for the replacement intervention (relative to Replace Control). The dependent variables are: (1) a platform engagement index, constructed from active time spent on the platform and the number of posts and comments shown to the user; (2) active time spent on the platform; (3) number of posts and comments shown to the user; (4) an ad engagement index, constructed from the number of ad impressions and the number of ad clicks; (5) number of ad impressions; and (6) number of ad clicks. All outcomes are transformed using the inverse hyperbolic sine (asinh) function; for index variables, the transformation is applied prior to standardization. The unit of observation is the individual-day, measured relative to the intervention date. Standard errors clustered at the individual level are reported in parentheses. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day. $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table B16: Long-Term Engagement Effects: Clustered Errors (Asinh)

	Platform Engagement			Ad Engagement		
	Index	Time	Content	Index	Impressions	Clicks
Panel A: Hide vs Pure Control						
Treated	0.054	0.103	0.175**	-0.260***	-0.640***	-0.032***
	(0.034)	(0.063)	(0.084)	(0.039)	(0.068)	(0.012)
Mean	0.00	1.22	1.72	0.00	0.951	0.053
SD	0.976	1.84	2.52	0.841	1.62	0.256
N	113308	113308	113308	113308	113308	113308
Panel B: Replace vs Replace Control						
Treated	-0.057*	-0.107*	-0.143*	-0.110***	-0.089*	-0.040***
	(0.031)	(0.058)	(0.077)	(0.037)	(0.051)	(0.014)
Mean	0.00	1.25	1.74	0.00	0.959	0.05
SD	0.978	1.87	2.53	0.839	1.64	0.243
N	136924	136924	136924	136924	136924	136924

Note: **Panel A** presents estimated treatment effects of the hiding intervention on Facebook (relative to Hide Control), based on Equation 3 and our main experimental sample. **Panel B** reports corresponding estimates for the replacement intervention (relative to Replace Control). All regressions are based on the extended sample, which involves 19.1 weeks of the intervention period (long-term effects) as opposed to 6 weeks specified in the pre-registration (short-term effects). The dependent variables are: (1) a platform engagement index, constructed from active time spent on the platform and the number of posts and comments shown to the user; (2) active time spent on the platform; (3) number of posts and comments shown to the user; (4) an ad engagement index, constructed from the number of ad impressions and the number of ad clicks; (5) number of ad impressions; and (6) number of ad clicks. All outcomes are transformed using the inverse hyperbolic sine (asinh) function; for index variables, the transformation is applied prior to standardization. The unit of observation is the individual-day, measured relative to the intervention date. Standard errors clustered at the individual level are reported in parentheses. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day. $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table B17: Sensitivity of Structural Estimates to Identifying Moments

Parameter	*Hide* Δe	*Rep.* Δe	Cov (WTA, $\overline{e^2}$)	*Hide* ΔWTA	*Rep.* ΔWTA
ψ/β	0.29	**-0.37**	-0.03	0.00	0.03
φ/β	0.04	**-0.46**	-0.09	-0.03	0.03
β	-0.02	-0.17	**0.31**	-0.02	0.05
γ_0	-0.06	-0.11	-0.07	-0.35	**0.46**
γ_1	0.04	0.09	0.05	-0.26	**-0.69**

Note: Standardized sensitivity of each estimate to each identifying moment (Andrews et al. 2017), from the bootstrap (per-draw parameters regressed on the five standardized identifying moments). Each row's largest-magnitude entry appears in bold. Columns: Δe and ΔWTA are the *Hide* and *Replace* treatment effects on engagement e and willingness to accept; Cov(WTA, $\overline{e^2}$) is the cross-user covariance between baseline WTA and the mean of squared daily baseline engagement.

Table B18: Effects of Interventions on Platform Valuation

	Hiding	Replacing
Post	26.646	12.283
	(27.197)	(25.751)
Treatment × Post	-13.262	-3.599
	(39.044)	(33.470)
Mean	258.64	217.72
SD	336.19	288.29
N	678	814

Note:
The table reports the estimated treatment effects of the hiding and replacement interventions—in the *Hiding* and *Replacing* columns, respectively—on individuals' willingness to accept (WTA) the deactivation of Facebook for four weeks. WTA is measured in both the baseline and endline surveys. The estimates are based on a difference-in-differences regression of WTA on the endline survey indicator (*Post*), its interaction with treatment assignment, and individual fixed effects. The unit of observation is the individual, and standard errors clustered at the individual level are reported in parentheses. $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

C Extension Errors

Browser extensions are complex pieces of software tailored to work with the HTML structure of specific websites. In our case, most of the functionality relates to user experience on Facebook, which the extension can seamlessly alter (via a hiding or a replacement intervention). Any change that Facebook makes to the HTML structure of its website may result in browser extension errors, invalidating parts of its functionality. This is because the extension relies on the specific structure of the website to perform its functions. In such an environment, some errors are typically inevitable. This appendix describes the errors that affected some of our participants during the study and offers robustness analysis to show that they did not impact the main conclusions of the paper.

C.1 Website Update

This error occurred as a result of a sudden change of the HTML code of the Facebook page. We are aware that it caused an issue with displaying advertiser names in replacement ads (in the replacement intervention)—the advertiser name sometimes contained wrong text. Since the HTML change potentially affected the extension's behavior for users in all treatment groups, we assign one to the error dummy for all users in the study on the days from the beginning of the problem until the fix was implemented. The share of user-day observations in which participants experienced at least one instance of this error was 4.9%; in total, 684 users were affected at some point during the pre-registered study period.

Given that the error occurred for all people on the same days, the error dummy is absorbed by period $\times$ cohort fixed effects. The error enters the regression through the dummy's interaction with treatment. Note that we have users in our sample who already finished the pre-registered intervention period by the time that the error occurred—we use that subsample for robustness analysis.

C.2 Faulty Replacement

This error only affects some users in the replacement group. Other groups are not affected. For a period of time, the replacement intervention replaced ads not with a randomly selected ad from the pool of ads but with a default ad. This still reduced the degree of ad microtargeting but did not maintain the natural ad variety. The share of user-day observations with at least one instance of this error was 0.5%, with 133 affected users.

Since the error does not affect users in groups other than the replacement group, for the hiding regression, both the error dummy and its treatment interaction get dropped. We have users in our sample who were recruited or transitioned into the intervention period after the error was fixed—we use that subsample for robustness analysis.

C.3 Faulty Hiding

This error affected a random subset of users enrolled at the time of its occurrence. During an extension update, some users who were in the baseline period temporarily experienced the hiding intervention. This was caused by the extension overriding values specifying the extension's behavior for a random subset of users. The share of user-day observations with at least one instance of this error was 0.4%, with 261 affected users.

Faulty hiding occurs only during the baseline period, so the error dummy's treatment interaction gets dropped from the regressions. We have users in our sample who were recruited after the error was fixed. We also have users who were never affected, as the extension did not randomly turn on the hiding intervention for them during the baseline. All of these users constitute a subsample that we use for robustness analysis.

C.4 Robustness

For each of the three error types, we assess robustness by excluding all users affected by the respective error. All robustness exercises use specifications with an asinh-transformed outcome variable, and results are compared to the baseline asinh regressions. Because skewness in the outcome variable limits the precision of estimates in levels regressions, comparing asinh regressions across samples provides a more meaningful robustness check.

Table B19 presents short-term results excluding individuals who experienced the website update error. All of the treatment effects reported in the original regression (Table B7) are robust to this exclusion. We do not have enough observations to perform this check for the long-term effects since all users are affected by this error unless they dropped out of the study before the error occurred.[34]

Tables B20 and B21 present short-term and long-term estimates excluding individuals who experienced the faulty replacement error. Treatment effects of the hiding intervention are unaffected by this exclusion. All long-term treatment effects of the replacement inter-

[34]The update causing this error happened close to the end of our study period.

vention reported in Table B8 remain robust, as do the short-term negative effects on content consumption and the ad engagement index relative to Table B7.

Lastly, Tables B22 and B23 present short-term and long-term estimates excluding participants who experienced the faulty hiding error. All short- and long-term treatment effects of the hiding intervention are robust to this exclusion. The same holds for all long-term treatment effects of the replacement intervention. The short-term negative effect of the replacement intervention on the ad engagement index also remains robust. The only platform-engagement effect that remains statistically insignificant is the short-term effect of the replacement intervention, consistent with the main results.

Based on the results of our robustness analysis, we conclude that the extension errors do not meaningfully impact the treatment effects and the paper's experimental conclusions.

Table B19: Short-Term Effects Website Update Error (Asinh)

	Platform Engagement			Ad Engagement		
	Index	Time	Content	Index	Impressions	Clicks
Panel A: Hide vs Pure Control						
Treated	0.058**	0.102**	0.190***	-0.287***	-0.712***	-0.033***
	(0.023)	(0.041)	(0.058)	(0.017)	(0.034)	(0.005)
Mean	0.00	1.28	1.76	0.00	0.963	0.052
SD	0.974	1.85	2.55	0.839	1.62	0.245
N	41160	41160	41160	41160	41160	41160
Panel B: Replace vs Replace Control						
Treated	-0.022*	-0.051*	-0.078**	-0.088***	-0.052**	-0.037***
	(0.012)	(0.027)	(0.035)	(0.012)	(0.020)	(0.006)
Mean	0.00	1.31	1.80	0.00	0.959	0.055
SD	0.977	1.90	2.60	0.845	1.68	0.257
N	53004	53004	53004	53004	53004	53004

Note: **Panel A** presents estimated treatment effects of the hiding intervention on Facebook (relative to Hide Control), based on Equation 3 and our main experimental sample, excluding individuals who experienced the website update error (Section C.1). **Panel B** reports corresponding estimates for the replacement intervention (relative to Replace Control). The dependent variables are: (1) a platform engagement index, constructed from active time spent on the platform and the number of posts and comments shown to the user; (2) active time spent on the platform; (3) number of posts and comments shown to the user; (4) an ad engagement index, constructed from the number of ad impressions and the number of ad clicks; (5) number of ad impressions; and (6) number of ad clicks. All outcomes are transformed using the inverse hyperbolic sine (asinh) function; for index variables, the transformation is applied prior to standardization. The unit of observation is the individual-day, measured relative to the intervention date. Driscoll-Kraay standard errors are reported in parentheses. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day. $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table B20: Short-Term Effects Faulty Replacement Error (Asinh)

	Platform Engagement			Ad Engagement		
	Index	Time	Content	Index	Impressions	Clicks
Panel A: Hide vs Pure Control						
Treated	0.066***	0.117***	0.202***	-0.285***	-0.687***	-0.033***
	(0.024)	(0.042)	(0.057)	(0.014)	(0.032)	(0.005)
Mean	0.00	1.20	1.63	0.00	0.904	0.05
SD	0.976	1.82	2.48	0.841	1.59	0.241
N	68712	68712	68712	68712	68712	68712
Panel B: Replace vs Replace Control						
Treated	-0.025	-0.055	-0.085*	-0.075***	-0.028	-0.032***
	(0.019)	(0.037)	(0.049)	(0.016)	(0.031)	(0.005)
Mean	0.00	1.22	1.64	0.00	0.883	0.05
SD	0.978	1.85	2.49	0.844	1.59	0.242
N	71148	71148	71148	71148	71148	71148

Note: **Panel A** presents estimated treatment effects of the hiding intervention on Facebook (relative to Hide Control), based on Equation 3 and our main experimental sample, excluding individuals who experienced the faulty replacement error (Section C.2). **Panel B** reports corresponding estimates for the replacement intervention (relative to Replace Control). The dependent variables are: (1) a platform engagement index, constructed from active time spent on the platform and the number of posts and comments shown to the user; (2) active time spent on the platform; (3) number of posts and comments shown to the user; (4) an ad engagement index, constructed from the number of ad impressions and the number of ad clicks; (5) number of ad impressions; and (6) number of ad clicks. All outcomes are transformed using the inverse hyperbolic sine (asinh) function; for index variables, the transformation is applied prior to standardization. The unit of observation is the individual-day, measured relative to the intervention date. Driscoll-Kraay standard errors are reported in parentheses. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day. $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table B21: Long-Term Effects Faulty Replacement Error (Asinh)

	Platform Engagement			Ad Engagement		
	Index	Time	Content	Index	Impressions	Clicks
Panel A: Hide vs Pure Control						
Treated	0.054***	0.103***	0.175***	-0.260***	-0.640***	-0.032***
	(0.019)	(0.033)	(0.048)	(0.012)	(0.026)	(0.004)
Mean	0.00	1.22	1.72	0.00	0.951	0.053
SD	0.976	1.84	2.52	0.841	1.62	0.256
N	113308	113308	113308	113308	113308	113308
Panel B: Replace vs Replace Control						
Treated	-0.073***	-0.138***	-0.187***	-0.093***	-0.066**	-0.036***
	(0.018)	(0.036)	(0.046)	(0.017)	(0.033)	(0.005)
Mean	0.00	1.25	1.74	0.00	0.959	0.05
SD	0.978	1.87	2.53	0.839	1.64	0.243
N	117541	117541	117541	117541	117541	117541

Note: **Panel A** presents estimated treatment effects of the hiding intervention on Facebook (relative to Hide Control), based on Equation 3 and our main experimental sample, excluding individuals who experienced the faulty replacement error (Section C.2). **Panel B** reports corresponding estimates for the replacement intervention (relative to Replace Control). All regressions are based on the extended sample, which involves 19.1 weeks of the intervention period (long-term effects) as opposed to 6 weeks specified in the pre-registration (short-term effects). The dependent variables are: (1) a platform engagement index, constructed from active time spent on the platform and the number of posts and comments shown to the user; (2) active time spent on the platform; (3) number of posts and comments shown to the user; (4) an ad engagement index, constructed from the number of ad impressions and the number of ad clicks; (5) number of ad impressions; and (6) number of ad clicks. All outcomes are transformed using the inverse hyperbolic sine (asinh) function; for index variables, the transformation is applied prior to standardization. The unit of observation is the individual-day, measured relative to the intervention date. Driscoll-Kraay standard errors are reported in parentheses. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day. $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table B22: Short-Term Effects Faulty Hiding Error (Asinh)

	Platform Engagement			Ad Engagement		
	Index	Time	Content	Index	Impressions	Clicks
Panel A: Hide vs Pure Control						
Treated	0.072**	0.132**	0.230***	-0.269***	-0.682***	-0.026***
	(0.029)	(0.051)	(0.071)	(0.019)	(0.036)	(0.006)
Mean	0.00	1.19	1.62	0.00	0.902	0.048
SD	0.977	1.81	2.48	0.84	1.60	0.237
N	58716	58716	58716	58716	58716	58716
Panel B: Replace vs Replace Control						
Treated	-0.007	-0.007	-0.020	-0.083***	-0.042	-0.034***
	(0.018)	(0.036)	(0.047)	(0.017)	(0.033)	(0.005)
Mean	0.00	1.26	1.70	0.00	0.926	0.052
SD	0.98	1.86	2.51	0.844	1.61	0.247
N	69468	69468	69468	69468	69468	69468

Note: **Panel A** presents estimated treatment effects of the hiding intervention on Facebook (relative to Hide Control), based on Equation 3 and our main experimental sample, excluding individuals who experienced the faulty hiding error (Section C.3). **Panel B** reports corresponding estimates for the replacement intervention (relative to Replace Control). The dependent variables are: (1) a platform engagement index, constructed from active time spent on the platform and the number of posts and comments shown to the user; (2) active time spent on the platform; (3) number of posts and comments shown to the user; (4) an ad engagement index, constructed from the number of ad impressions and the number of ad clicks; (5) number of ad impressions; and (6) number of ad clicks. All outcomes are transformed using the inverse hyperbolic sine (asinh) function; for index variables, the transformation is applied prior to standardization. The unit of observation is the individual-day, measured relative to the intervention date. Driscoll-Kraay standard errors are reported in parentheses. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day. $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

Table B23: Long-Term Effects Faulty Hiding Error (Asinh)

	Platform Engagement			Ad Engagement		
	Index	Time	Content	Index	Impressions	Clicks
Panel A: Hide vs Pure Control						
Treated	0.053**	0.107***	0.188***	-0.249***	-0.632***	-0.026***
	(0.022)	(0.038)	(0.057)	(0.015)	(0.029)	(0.005)
Mean	0.00	1.19	1.69	0.00	0.934	0.05
SD	0.977	1.82	2.50	0.84	1.61	0.249
N	96705	96705	96705	96705	96705	96705
Panel B: Replace vs Replace Control						
Treated	-0.049***	-0.085**	-0.110***	-0.110***	-0.088***	-0.041***
	(0.016)	(0.033)	(0.040)	(0.018)	(0.031)	(0.005)
Mean	0.00	1.29	1.80	0.00	1.01	0.051
SD	0.98	1.89	2.56	0.837	1.67	0.245
N	114601	114601	114601	114601	114601	114601

Note: **Panel A** presents estimated treatment effects of the hiding intervention on Facebook (relative to Hide Control), based on Equation 3 and our main experimental sample, excluding individuals who experienced the faulty hiding error (Section C.3). **Panel B** reports corresponding estimates for the replacement intervention (relative to Replace Control). All regressions are based on the extended sample, which involves 19.1 weeks of the intervention period (long-term effects) as opposed to 6 weeks specified in the pre-registration (short-term effects). The dependent variables are: (1) a platform engagement index, constructed from active time spent on the platform and the number of posts and comments shown to the user; (2) active time spent on the platform; (3) number of posts and comments shown to the user; (4) an ad engagement index, constructed from the number of ad impressions and the number of ad clicks; (5) number of ad impressions; and (6) number of ad clicks. All outcomes are transformed using the inverse hyperbolic sine (asinh) function; for index variables, the transformation is applied prior to standardization. The unit of observation is the individual-day, measured relative to the intervention date. Driscoll-Kraay standard errors are reported in parentheses. All regressions were estimated using inverse probability weights reflecting the probability of group assignment for individuals recruited on a given day. $*$, $**$, and $***$ denote significance at the 10%, 5%, and 1% levels, respectively.

D Recruitment and Sample

D.1 Recruitment Funnel

Our Facebook recruitment advertisements reached 175,044 unique users and generated 2,364 direct link clicks. We recorded 2,427 installations of the browser extension, and 2,279 participants completed the baseline survey.[35]

D.2 Sample Selection Criteria

We preregistered the following exclusion criteria, applied to all users who installed the extension (baseline completion was not a prerequisite to participation):

- English not set as the interface language on Facebook—this may interfere with content and ad detection that were designed for and tested on English language accounts;
- Using a browser other than Chrome or Edge (both popular Chromium browsers supported by our extension)—for example, Opera or Firefox. The extension was not designed or tested for other browsers;
- Creating multiple accounts or using bots in violation of the consent form. We identified bots by flagging cases where multiple survey responses arrived simultaneously from distinct IPs with identical answers, or where extension-recorded outcomes exceeded 30 standard deviations from the mean;
- Uninstalling the extension upon installation, before the treatment period begins;[36]
- Using Facebook for less than 1 minute per week on average during the baseline period.

In total, 617 of the 2,427 users who installed the extension are excluded from the analysis. The counts of users who failed individual exclusion criteria (which may overlap) are: insufficient Facebook activity (421), uninstallation event or failure to update the extension (239), browser other than Chrome or Edge (40), bot-like behavior (13), and unsupported language (1). After applying these criteria, we arrive at the main experimental sample of 1,810 users. Of these, 1,384 remained active at the end of the full six-week intervention period, and 818 completed the endline survey.

[35] The number of installations exceeds the number of link clicks because some users returned to the landing page some time after they visited it originally, others searched for the study to verify its legitimacy, etc.

[36] We use two methods to identify cases like this. The earliest version of the extension was not able to track uninstallations, so we rely on the user never updating the extension (which happens automatically as they load the browser). Later we added functionality to track uninstallations directly.

E Supplementary Data Collection

E.1 Primary Results: Figures and Tables

This section provides additional tables and figures supporting analysis of primary outcomes from the supplementary data collection described in Section 4.4 and external validity described in Section 4.5.

Table B24: Feed Content per Individual: Desktop versus Mobile

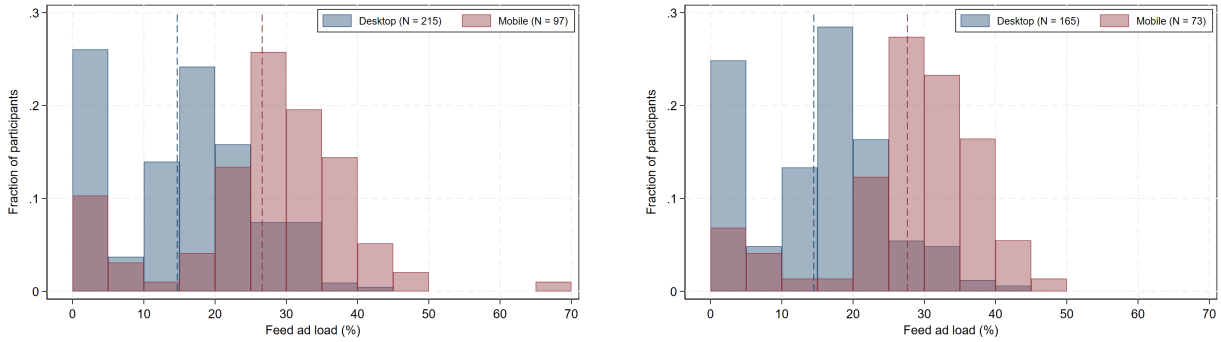
	Mean	Q1	Median	Q3
Panel A: Desktop (N = 215)				
Posts	30.64	12.00	26.00	42.00
Ads	5.62	1.00	4.00	8.00
Feed Comments	3.42	0.00	1.00	4.00
Total Content	39.68	17.00	34.00	57.00
Ad load (%)	14.69	3.92	16.67	20.83
Ad load, excl. ad blockers (%)	15.84	10.30	17.48	21.62
Panel B: Mobile (N = 97)				
Posts	32.51	15.00	26.00	40.00
Ads	12.86	5.00	10.00	17.00
Feed Comments	1.92	0.00	1.00	3.00
Total Content	47.28	23.00	39.00	60.00
Ad load (%)	26.61	21.57	27.27	34.43

Note: This table reports per-individual counts of distinct feed elements encountered during the browsing session—posts, ads, and feed comments—with *Total Content* their sum and *Ad load* the share of ads in total content. Columns report the across-participant mean and quartiles (Q1, median, Q3). Panel A covers the desktop browser-extension data; Panel B covers the mobile screen-recording LLM annotations, where per-individual counts are reconciled within participant across three independent coding passes by the median. The final row of the desktop panel recomputes ad load after excluding participants who reported an active ad blocker during the session; participants who did not answer the ad-blocker question are excluded. The table is based on the sample recruited during the supplementary data collection (see Section 4.4).

Table B25: Feed Ad Load and Posts Consumed by Ad-Blocker Use

	No Ad Blocker	Ad Blocker	Total
Share of the sample (%)	88.26	11.74	100.00
Mean feed ad load (%)	15.84	7.23	14.83
Mean posts consumed	30.11	35.72	30.77

Note: Participants are grouped by self-reported ad-blocker use. *Ad Blocker* denotes participants reporting a blocker that was active during the scrolling session; *No Ad Blocker* pools all other responses (including participants who use a blocker but disabled it for the study, and those who were unsure); *Total* comprises all responses. Rows report each group's share of the sample, its mean feed ad load (the share of feed elements—posts, ads, and comments—that are ads), and its mean number of distinct posts consumed, all computed at the individual level. The table is based on the sample recruited during the supplementary data collection (see Section 4.4).



(a) Full sample

(b) ≥ first-quartile total content

FIGURE A19: DISTRIBUTION OF FEED AD LOAD ACROSS INDIVIDUALS

Note: Each panel overlays the per-participant distribution of *feed ad load*—the share of a participant's feed elements (posts, ads, and comments) that are ads—for the desktop browser-extension sample and the mobile screen-recording sample; the y-axis is the fraction of participants within each device sample, and dashed lines mark the device means. Panel (a) uses all participants (desktop $N = 215$; mobile $N = 97$). Panel (b) restricts each device to participants who encountered at least that device's own first-quartile amount of total content (desktop $N = 165$; mobile $N = 73$). The figure is based on the sample recruited during the supplementary data collection (see Section 4.4).

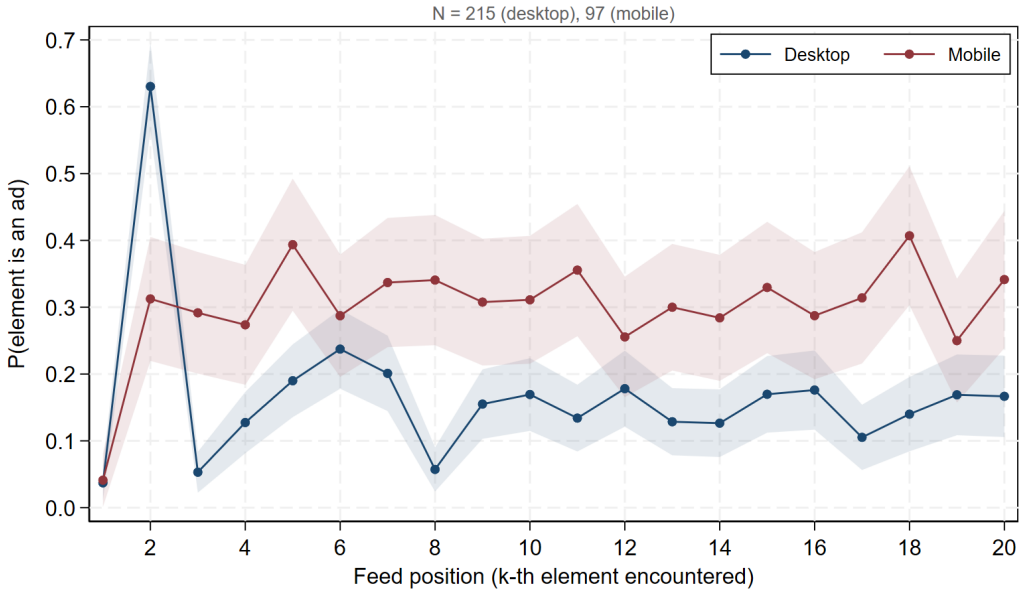


FIGURE A20: AD PROBABILITY BY FEED POSITION: DESKTOP VERSUS MOBILE

Note: This figure reports the empirical probability that the k-th feed element a participant encounters is an ad, by feed position, separately for desktop and mobile, over the first twenty positions. Feed position is the order in which elements first enter the screen within a participant's session. Desktop elements are recorded by the browser extension; mobile elements are annotated from each participant's screen recording by a vision language model (OpenAI's GPT-5.4-mini; see Appendix E.3). Shaded bands are 95% confidence intervals. The figure is based on the sample recruited during the supplementary data collection (see Section 4.4).

Table B26: Viewport Time on Ads versus Organic Posts

	Basic (timestamp)					Counter (logged)				
	(1) No ctrl	(2) OLS	(3) FE	(4) OLS+md	(5) FE+md	(6) No ctrl	(7) OLS	(8) FE	(9) OLS+md	(10) FE+md
ad	-1.30	-1.70*	-1.59*			-1.50**	-2.01***	-2.09***		
	(1.16)	(1.01)	(0.94)			(0.75)	(0.75)	(0.54)		
height		0.01	0.01**	0.01	0.01**		0.01***	0.01***	0.01***	0.01***
		(0.00)	(0.00)	(0.00)	(0.00)		(0.00)	(0.00)	(0.00)	(0.00)
mt_image				-0.97	-1.28				-0.95	-1.64**
				(1.12)	(1.20)				(1.07)	(0.74)
mt_video				-3.55***	-2.54**				-4.17***	-3.13***
				(1.06)	(0.99)				(0.90)	(0.82)
mt_carousel				-0.22	0.95				-0.18	0.04
				(1.99)	(1.40)				(2.04)	(1.54)
_cons	12.53***	9.11**	6.92***	9.06**	6.86***	13.63***	9.22***	10.16***	9.16***	10.10***
	(1.14)	(3.92)	(2.52)	(3.94)	(2.54)	(0.74)	(1.68)	(1.23)	(1.68)	(1.24)
N	9506	9506	9506	9506	9506	9506	9506	9506	9506	9506
R^2	0.00	0.00	0.07	0.00	0.07	0.00	0.01	0.18	0.01	0.18
Adj. R^2	-0.00	0.00	0.05	0.00	0.05	0.00	0.01	0.16	0.01	0.16

Note: This table reports estimates from regressions of per-element viewport time on an indicator for ads, relative to organic posts; the unit of observation is a post or an ad. Columns 1–5 use the basic timestamp measure—viewport time reconstructed by summing the intervals between an element's logged events during which it overlapped the viewport—while columns 6–10 use the logged counter measure—the extension's on-screen timer, recorded continuously in the browser and paused whenever the element leaves the viewport or the tab goes inactive. Within each measure, the first column includes no controls (its constant is the mean post viewport time), and the remaining columns successively add element height, user fixed effects, and a split of the ad indicator into ad type dummies (image, video, carousel). Standard errors clustered at the individual level are in parentheses; $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. The table is based on the sample recruited during the supplementary data collection (see Section 4.4).

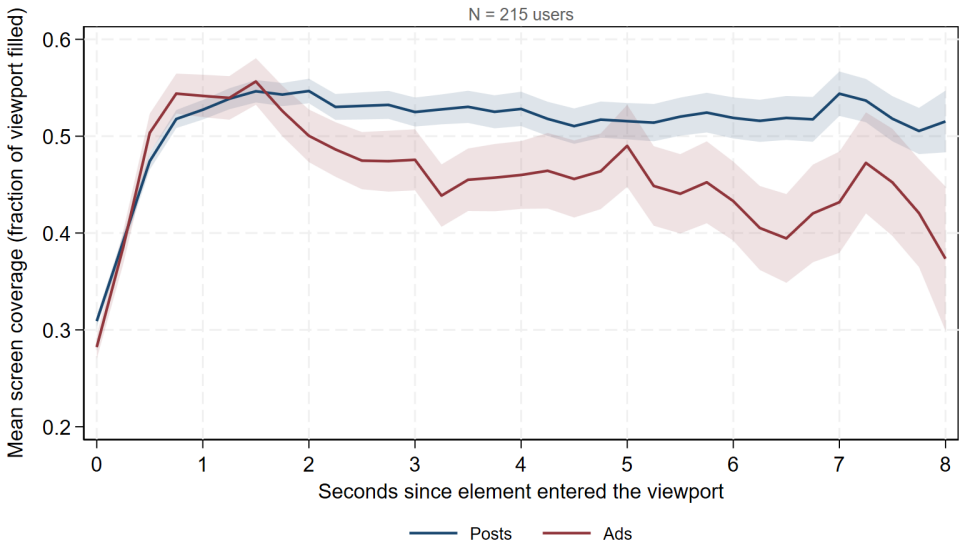
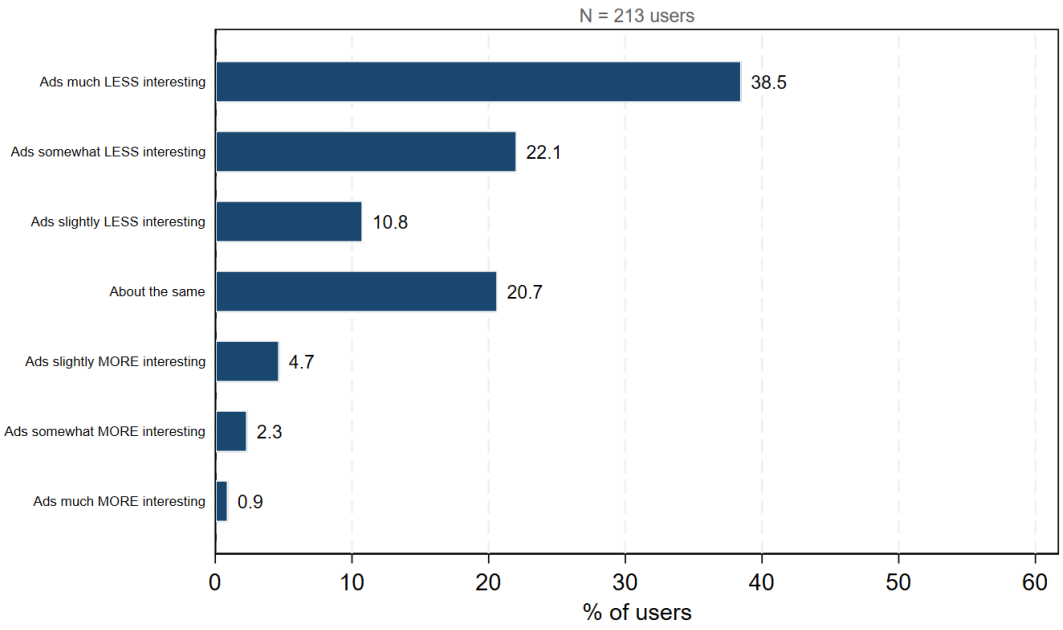
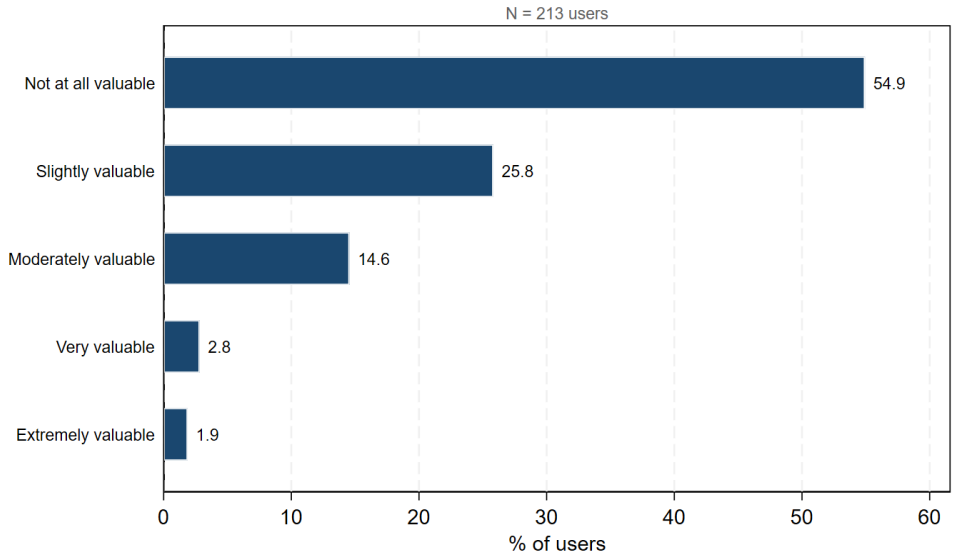


FIGURE A21: SCREEN COVERAGE OF ADS/POSTS AFTER ENTERING VIEWPORT

Note: This figure reports the average on-screen presence of ads versus organic posts in the seconds after an element enters the viewport. For each post or ad impression we measure its screen coverage—the fraction of the viewport the element fills (its visible share times its height relative to the viewport height)—every 250 milliseconds, aligned to the moment it first enters the viewport, then average these trajectories across impressions with 95% confidence bands. Sampling stops once an element scrolls out of view or the tab goes idle, so later seconds average fewer impressions. The figure is based on the sample recruited during the supplementary data collection (see Section 4.4).



(a) Interest in ads vs. organic posts



(b) Value of ads seen this session

FIGURE A22: STATED INTEREST IN AND VALUE OF ADS

Note: This figure reports the distribution of participants' self-reported attitudes toward the ads they saw during the browsing session. Panel (a) shows how interesting they found the session's ads relative to organic posts, on a seven-point scale from "much less" to "much more interesting"; panel (b) shows how valuable they found those ads, on a five-point scale from "not at all" to "extremely valuable." Bars give the share of participants choosing each response (213 participants answered each item). The figure is based on the sample recruited during the supplementary data collection (see Section 4.4).

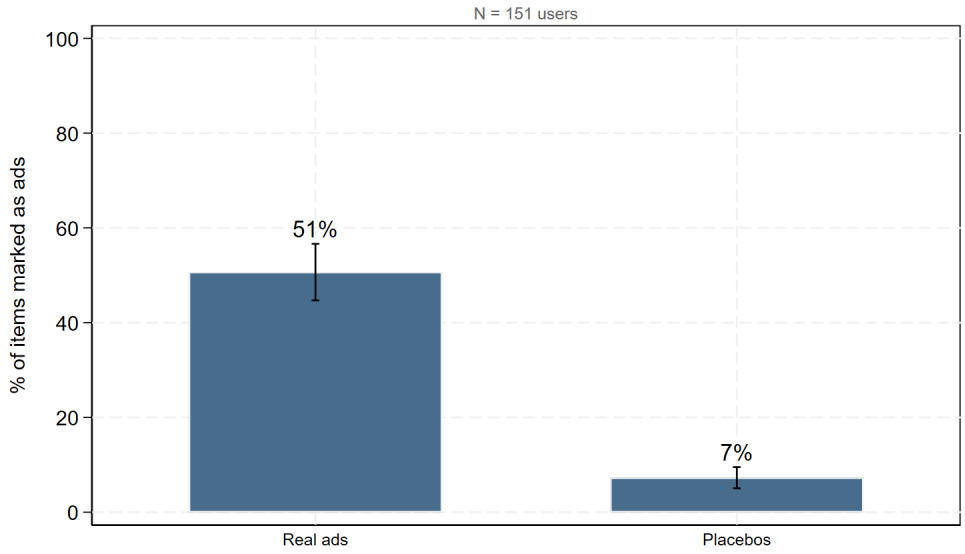


FIGURE A23: RECALL RATE FOR REAL ADS VERSUS PLACEBOS

Note: This figure reports the probability that a participant marks an item as an ad they remember seeing, conditional on the item being one of their real ads or a placebo, with 95% confidence intervals across participants. The aided-recognition task pairs two of a participant's real ads against four placebos, so it can be run only for participants for whom the extension recorded at least two distinct ads during the session; the sample is the 151 participants who meet this criterion. The figure is based on the pool of participants recruited during the supplementary data collection (see Section 4.4).

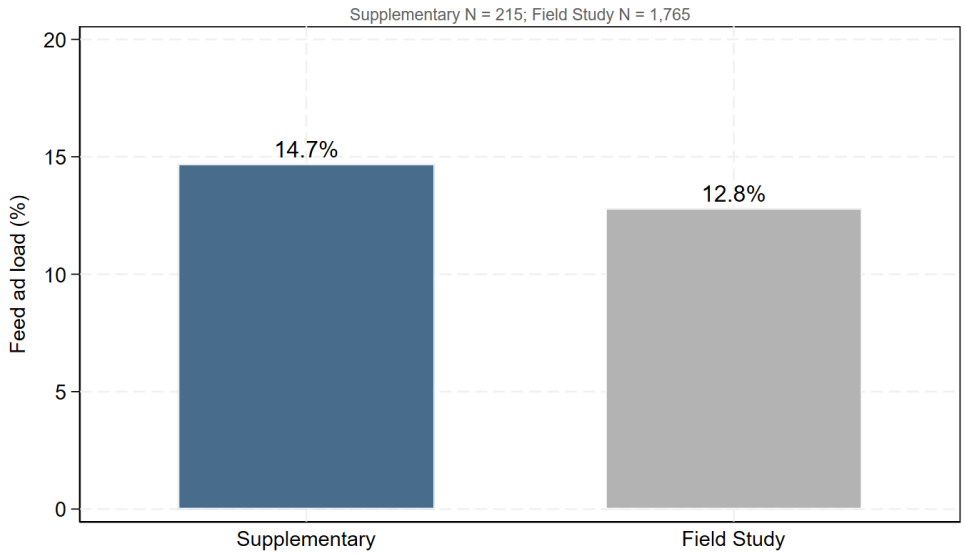


FIGURE A24: FEED AD LOAD: SUPPLEMENTARY SAMPLE VS. FIELD EXPERIMENT

Note: This figure reports the mean feed ad load—the share of feed elements that are ads, averaged across users—for our supplementary sample (see Section 4.4) and for the field experiment, which recruited participants through CloudResearch and social media advertising, respectively.

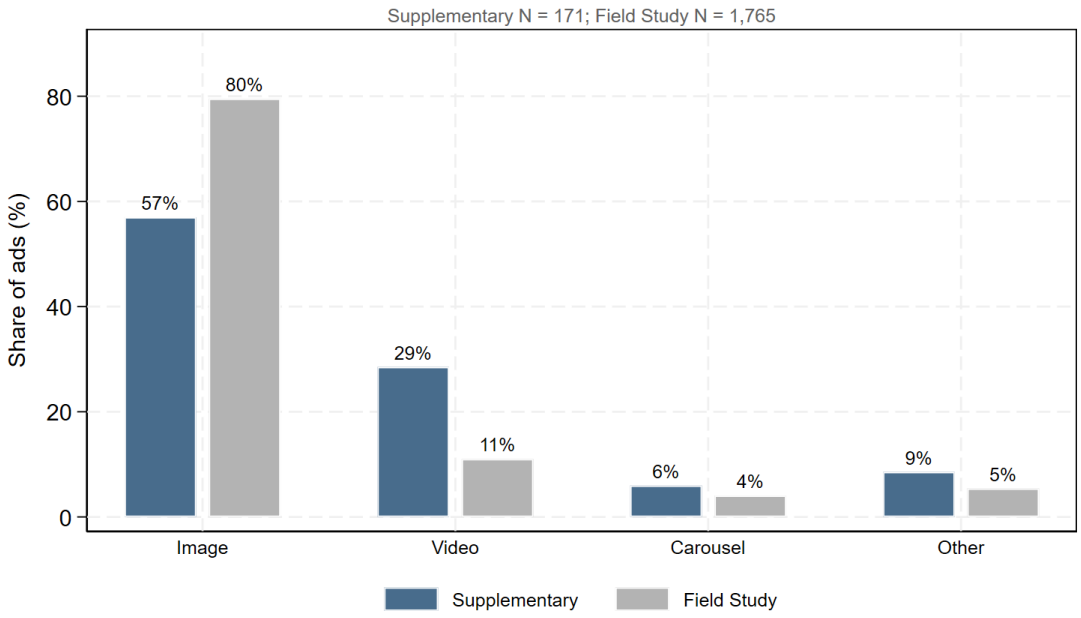


FIGURE A25: AD TYPES: SUPPLEMENTARY SAMPLE VS. FIELD EXPERIMENT

Note: This figure reports the share of ads by format—image, video, carousel, and other—averaged within users, for our supplementary sample (see Section 4.4) and for the field experiment. "Other" collects unclassified ads in the supplementary sample and lead-form ads in the field experiment.

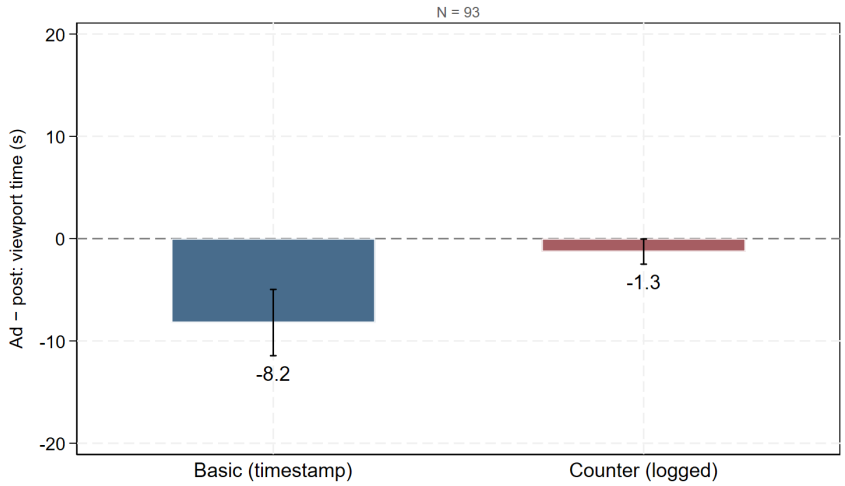


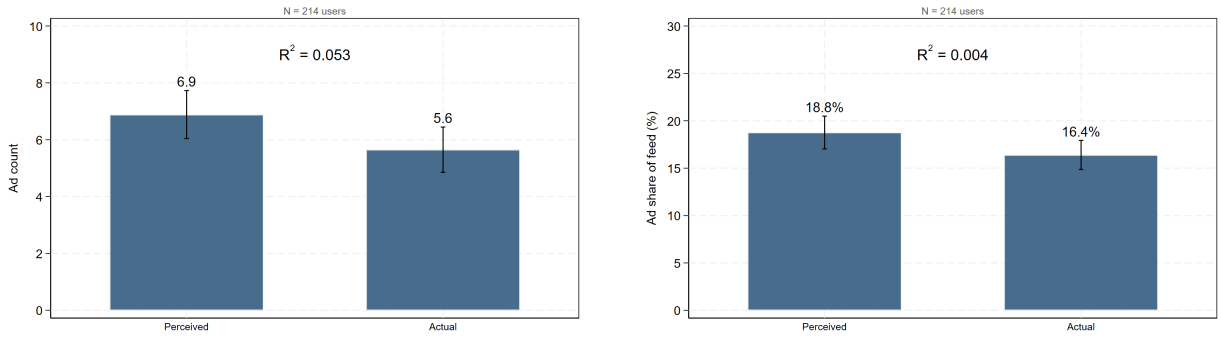
FIGURE A26: SCREEN TIME ON ADS VERSUS POSTS (MOBILE)

Note: This figure reports the within-participant difference in mean viewport time between ads and posts on mobile. On-screen times are measured from each participant's mobile screen recording, annotated element by element by a vision language model (OpenAI's GPT-5.4-mini; see Appendix E.3), under two measures of time on screen. *Basic (timestamp)* reconstructs each element's time on screen from its viewport enter and exit event times; *Counter (logged)* uses the model's running visible-time counter, averaged over the three codings. For each participant we take the mean viewport time of ads minus that of posts; bars are the across-participant means and whiskers are 95% confidence intervals. The sample is mobile participants with at least one ad and one post. The figure is based on the sample recruited during the supplementary data collection (see Section 4.4).

E.2 Other Results

This section presents analysis of additional secondary outcomes from the supplementary data collection described in Section 4.4.

Ad-Count and Ad-Load Blindness. We compare each participant's perceived ad exposure, self-reported in the survey, with the actual exposure recorded by the extension, both for the number of ads and for the ad share of the feed (as a fraction of total posts and ads).[37] On average, participants are roughly accurate: mean perceived exposure is close to mean actual exposure for both measures, with only a slight tendency to over-report (Figure A27). This average accuracy, however, masks substantial individual variation. Regressing each participant's actual exposure on their self-reported exposure yields a very low R^2—0.053 for the ad count and 0.004 for the ad share. This is consistent with participants holding only a vague, low-confidence impression of their ad exposure.



(a) Number of ads

(b) Ad share of the feed

FIGURE A27: PERCEIVED VERSUS ACTUAL AD EXPOSURE

Note: This figure reports the mean self-reported (perceived) and mean extension-recorded (actual) ad exposure across participants, with 95% confidence intervals: panel (a) the number of ads and panel (b) the ad share of the feed. The R^2 printed in each panel is from a participant-level regression of the actual measure on the self-reported one. The figure is based on the sample recruited during the supplementary data collection (see Section 4.4).

Other Ad Perceptions. Three further survey items capture participants' broader views of Facebook ads. About 78% report that ads are easy to tell apart from organic posts (Figure A28). The perceived value of Facebook ads in general—beyond this session—is low,

[37] Throughout this comparison the ad share excludes feed comments from the denominator, to match the survey question, which asked participants about the share of ads among posts. This differs from the feed ad load used elsewhere in the paper.

with 27% of participants never finding them valuable and 69% rating them valuable only rarely or sometimes (Figure A29). Together, these two findings provide further support for the avoidance channel and against the valuation channel: users' inelasticity to ad load reflects their ability to easily identify and skip ads rather than their valuing ads as highly as organic content.

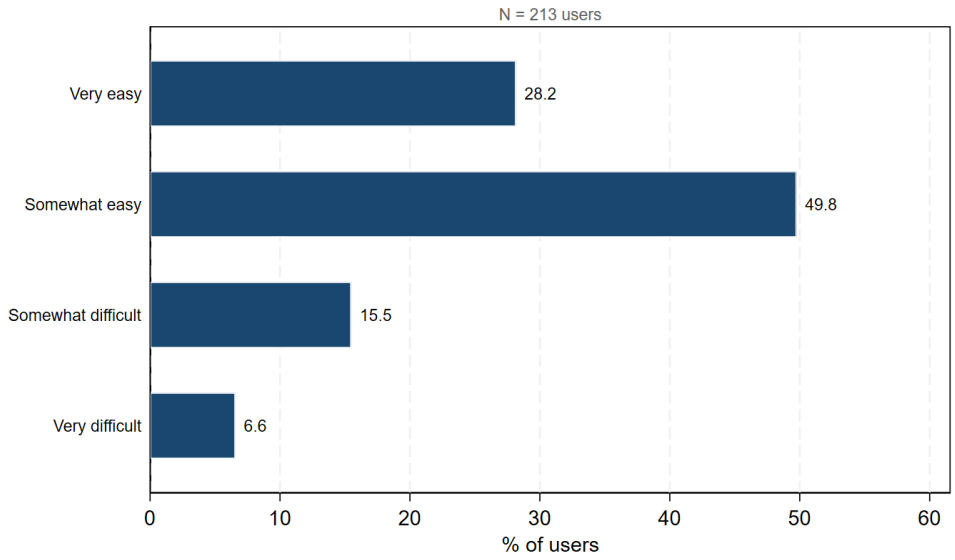


FIGURE A28: EASE OF DISTINGUISHING ADS FROM POSTS

Note: This figure reports the distribution of participants' self-reported ease of telling ads apart from organic posts. Bars give the share of participants choosing each response. The figure is based on the sample recruited during the supplementary data collection (see Section 4.4).

Separately, the perceived noticeability of Facebook ads relative to other platforms (YouTube, Instagram, TikTok) is centered on "about the same," with a lean toward Facebook ads being no more noticeable than ads elsewhere (Figure A30).

Willingness to Accept to Deactivate Facebook. Finally, we elicit each participant's willingness to accept—the minimum amount they would take to deactivate their Facebook account for four weeks—a consequential measure of how much they value the platform. The distribution is right-skewed (Figure A31): the median participant asks for $25, while a long upper tail lifts the mean to roughly $128.

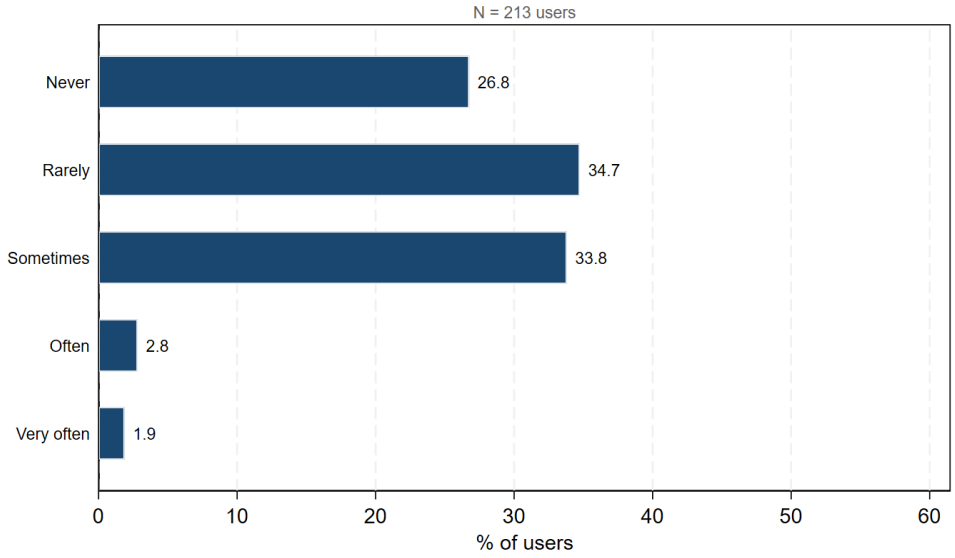


FIGURE A29: GENERAL PERCEIVED VALUE OF FACEBOOK ADS

Note: This figure reports the distribution of participants' self-reported view of how valuable Facebook ads are in general, beyond the ads seen in this session. Bars give the share of participants choosing each response. The figure is based on the sample recruited during the supplementary data collection (see Section 4.4).

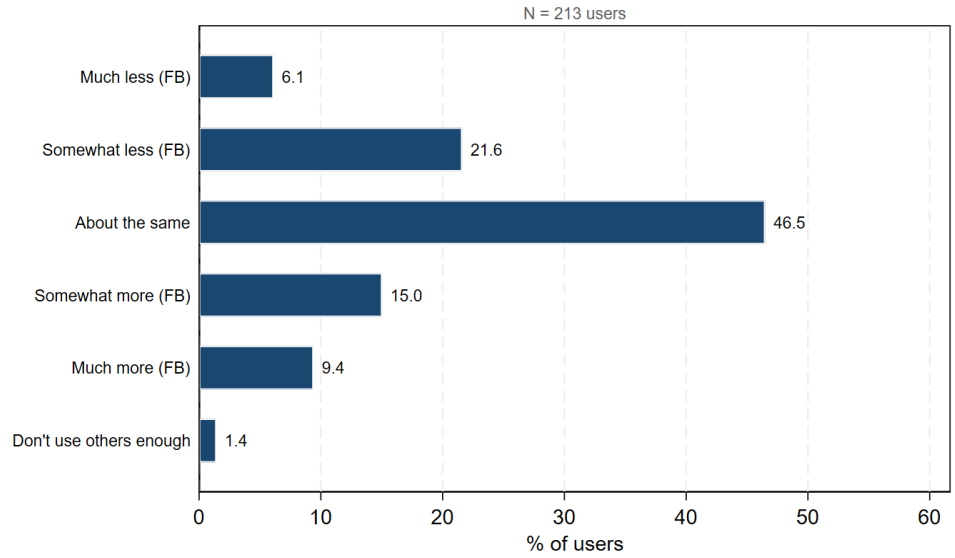


FIGURE A30: NOTICEABILITY OF FACEBOOK ADS VERSUS OTHER PLATFORMS

Note: This figure reports the distribution of participants' perceived noticeability of Facebook ads relative to other platforms (YouTube, Instagram, TikTok). Bars give the share of participants choosing each response. The figure is based on the sample recruited during the supplementary data collection (see Section 4.4).

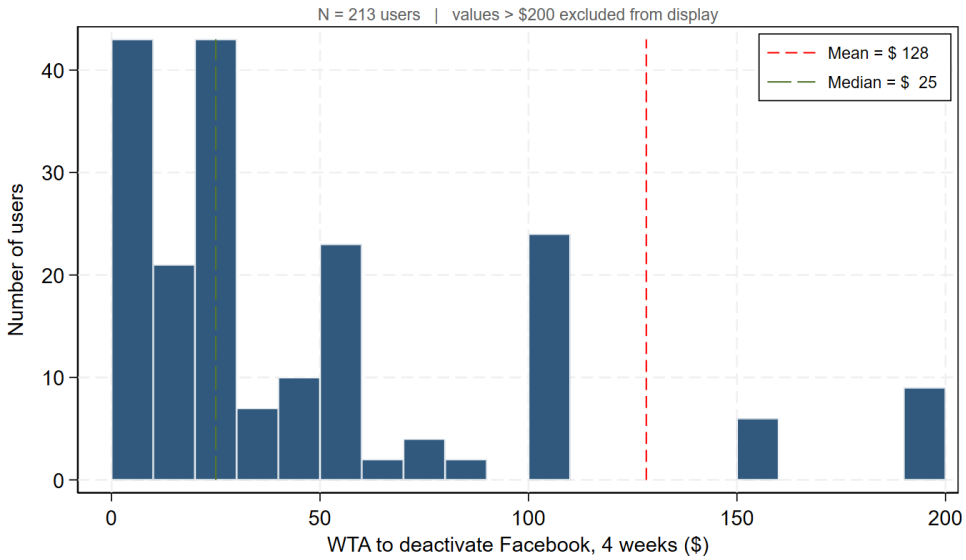


FIGURE A31: WILLINGNESS TO ACCEPT TO DEACTIVATE FACEBOOK FOR FOUR WEEKS

Note: This figure reports the distribution of participants' willingness to accept—the minimum dollar amount they would accept to deactivate their Facebook account for four weeks. Bars are $10 wide; the horizontal axis is truncated at $200 for display, while the mean and median (the dashed vertical lines) are computed on all values. The figure is based on the sample recruited during the supplementary data collection (see Section 4.4).

E.3 LLM Annotation of the Mobile Screen Recordings

For the mobile app browsing session, we reconstruct a counterpart to the desktop extension's event dataset directly from each participant's screen recording using OpenAI's GPT-5.4-mini. For every feed element (a post or an ad) we recover its type, its descriptive attributes, and the timestamps at which its edges enter and leave the screen, together with the number of feed comments visible inline beneath it.

Because the model cannot ingest video, we sample frames from each recording with OpenCV and pass them in temporal order, each labeled with its elapsed time in seconds. Sampling is motion-adaptive: we keep a floor of one frame per second and add frames during scrolling—whenever the feed advances by about half a screen since the last kept frame—so that the edge crossings that determine on-screen exposure are sampled densely while static stretches are not. Frames are downscaled to 512 pixels wide and sent in chunks of at most forty-five seconds of video; the model returns structured JSON that we assemble into a per-participant event file.

The model codes the Facebook feed only; any other screen—another app or Facebook tab; a full-screen photo, video, story, or reel; or an opened comment thread—is logged as a context event (feed left or returned) rather than a feed element, so time away from the feed is not counted as attention. Within the feed it records one element per card, classified as an ad only if it carries an explicit marker (e.g., a "Sponsored" label) and as a post otherwise. For ads it transcribes the advertiser name, ad text, call-to-action, and media type (image, video, or carousel). All on-screen text is transcribed verbatim, and the model is instructed to leave a field blank rather than guess. For each post and ad it logs its viewport edge-crossings—each being the card's top or bottom edge entering or leaving the screen, four types in all—with timestamps, on both downward and upward scrolls, and a running counter of cumulative on-screen time that mimics the desktop extension's counter-based time-on-screen metric.

To reduce noise, each recording is coded three times independently, with elements matched across codings by identity—author name for posts, advertiser name or ad text for ads. How we combine the three codings depends on the quantity. Element counts (posts, ads, and comments) are discrete detections that a single coding can occasionally miss or double-count, so we take the median across codings, which is insensitive to such errors. The timestamp-based measure of time on screen is reconstructed from the union of the edge-crossings detected across codings—a crossing is used if any coding records it, timed by the median of the

codings that do—and the visible intervals are then summed. The counter-based measure of time on screen is instead a continuous reading whose coding-to-coding variation reflects noise rather than a missed detection, so we average it across the three codings to use all three readings.

Some features of our measurements warrant caution when interpreting the mobile results. First, an element's ad status is inferred by the model from on-screen markers, so it is subject to classification error. Second, the model may fail to record some feed elements; because ad load is the share of elements that are ads, if ordinary posts are missed more often than the more salient, marked ads, the mobile ad load would be biased upward. Third, reconstructing events from sampled frames yields coarser timing than the extension—accurate to a few tenths of a second at scroll boundaries and about a second elsewhere. For these reasons we treat the mobile data as a complementary, secondary measurement and read desktop–mobile comparisons with caution; fully validating the model's classifications would require manual annotation of the recordings.