CESifo WORKING PAPERS

10011
2025

Original Version: October 2024
This Version: June 2025

Ranking for Engagement: How Social Media Algorithms Fuel Misinformation and Polarization

Fabrizio Germano, Vicenç Gómez, Francesco Sobbrio

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.ifo.de/en/cesifo/publications/cesifo-working-papers
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.ifo.de/en/cesifo/publications/cesifo-working-papers

Ranking for Engagement: How Social Media Algorithms Fuel Misinformation and Polarization*

Fabrizio Germano† Vicenç Gómez‡ Francesco Sobbrio§

June 2025

Abstract

Social media are at the center of countless debates on polarization, misinformation, and even the state of democracy in various parts of the world. An essential feature of social media is their recommendation algorithm that determines the ranking of content presented to the users. This paper investigates the dynamic feedback loop between recommendation algorithm and user behavior, and develops a theoretical framework to assess the impact of popularity-based parameters on platform engagement, misinformation, and polarization. The model uncovers a fundamental trade-off: assigning greater weight to online social interactions—such as likes and shares—increases user engagement but also increases misinformation (*crowding-out the truth*) and polarization. Building on this insight, the analysis considers how a simple "engagement tax" on social interactions can mitigate these negative externalities by altering platform incentives in the design of profit-maximizing algorithms. The framework is extended to include personalized rankings, demonstrating that personalization further amplifies polarization. Finally, empirical evidence from survey data in Italy and the United States indicates that Facebook's 2018 "Meaningful Social Interactions" update—which increased the emphasis on certain engagement metrics—contributed to increased ideological extremism and affective polarization.

Keywords: Social media, recommendation algorithm, ranking algorithm, feedback loop, engagement, misinformation, polarization, popularity ranking, algorithmic gatekeeper.

*A previous version of the paper circulated with title "Crowding out the truth? A simple model of misinformation, polarization, and meaningful social interactions" (October 2022). We thank Roland Bénabou, Filipe Campante, Alexander Frug, Gaël Le Mens, and audiences at the 2022 AI+Economics Workshop (ETH Zurich), the 8th International Conference on Computational Social Science in Chicago (IC^{2}S^2 2022) and at the 5th Economics of Media Bias Workshop (Berlin 2022) for helpful comments. Fabrizio Germano acknowledges financial support from Grant PID2020-115044GB-I00/MICIU/AEI/10.13039/501100011033 and from the Spanish Agencia Estatal de Investigación (AEI), through the Severo Ochoa Programme for Centres of Excellence in R&D (Barcelona School of Economics CEX2019-000915-S). Francesco Sobbrio is grateful to Nando Pagnoncelli and IPSOS for allowing access to the data of the Polimetro.

†Department of Economics and Business, Universitat Pompeu Fabra, and BSE, fabrizio.germano@upf.edu

‡Department of Information and Communications Technologies, Universitat Pompeu Fabra, Barcelona, vicen.gomez@upf.edu

§Department of Economics and Finance, Tor Vergata University of Rome and CESifo, francesco.sobbrio@uniroma2.it

"Many of the familiar pathologies of social media are, in my view, relatively direct consequences of engagement optimization" (*Understanding Social Media Recommendation Algorithms*, Narayanan 2023, page 35)

1 Introduction

Digital platforms play an increasingly central role in the consumption of political news (Aridor et al., 2024). A defining feature of these platforms is the use of recommendation algorithms that rank content based on measures of user engagement—most commonly clicks, likes, shares, or retweets. Put simply, these algorithms decide what information to show to users and, importantly, also in what order to show it. While such algorithms are effective at boosting traffic and user activity, there is growing public concern (CNN, 2021; Amnesty International, 2022; EU DisinfoLab, 2022; UNESCO, 2023; United Nations, 2023)—and related empirical evidence (Levy, 2021; Braghieri et al., 2025)—they may also distort the information environment, contributing to the spread of misinformation and the amplification of political polarization.[1]

This paper develops a formal model of a digital platform to study how algorithmic ranking, based on popularity measures derived from user behavior, shapes individual information acquisition and aggregate outcomes. In our framework, individuals seek information about a continuous, unknown state of the world (e.g., the net benefit of a vaccine, the credibility of a political candidate, or the severity of climate change). They choose among news items presented by a platform, each of which contains a signal about the state. The platform displays (i.e., ranks) these items based on their popularity, defined as a weighted average of clicks and highlights, where a highlight may correspond to a user "liking", sharing, or otherwise endorsing a piece of content. A central parameter in the model, denoted η, governs the weight the platform assigns to highlights relative to clicks.

The model incorporates three key behavioral elements. First, individuals are more likely to click on content whose perceived stance aligns with their prior beliefs (Gentzkow and Shapiro, 2010; Yom-Tov et al., 2013; White and Horvitz, 2015; Flaxman et al., 2016; Braghieri et al., 2025). Second, users highlight items only when the content is sufficiently close to their prior, with a higher propensity to highlight when their beliefs are more extreme. This microfoundation captures robust empirical regularities about online behavior and political sharing (An et al., 2014; Bakshy et al., 2015; Grinberg et al., 2019; Pew, 2019; Pogorelskiy and Shum, 2019; Garz et al., 2020; Hopp et al., 2020; Konovalova et al., 2023; Braghieri et al., 2025; Fraxanet et al., 2025). Third, individuals are more likely to click on content that appears higher in the ranking (Pan et al., 2007; Novarese and Wilson, 2013; Glick et al., 2014; Epstein and Robertson,

[1] For evidence on the overall impact of social media on polarization and misinformation see also Bursztyn et al. (2019); Di Tella et al. (2021); Müller and Schwarz (2021, 2023); Fujiwara et al. (2024).

2015), creating an important externality and feedback element between user engagement and future content visibility.

We show that the dynamic feedback between recommendation algorithm and individual behavior gives rise to a tradeoff in the platform's design problem. Increasing the weight on highlights (i.e., raising η) boosts engagement by promoting content that is more likely to be liked and shared. However, this same mechanism amplifies the visibility of extreme content, as individuals with more extreme beliefs are more likely to highlight content closely aligned with their own views. As a result, the distribution of content users see and ultimately click on becomes increasingly bimodal and polarized. This amplification of extremes also lowers the average quality of news items consumed (as measured by the closeness of the news signals to the true state). We refer to this mechanism as *crowding-out the truth*. Our results suggest that ranking algorithms may contribute to generating the "missing middle" (Bartels, 2016) and affective polarization (Iyengar et al., 2019) in the political realm. We also point out the presence of a related insight. Besides increasing actual polarization, a boost in the highlighting weight may also increase *perceived* polarization, since it makes it more likely that an individual will see more extreme content–both like-minded and not like-minded–in higher-ranked positions. This is in line with Bail (2021), who argues that social media act as a "prism" that conveys a distorted image of others and ends up muting moderates, while fueling actual and perceived polarization (see also Yang et al. 2016).

To sharpen the comparative statics analysis, we formally characterize limiting clicking and highlighting behavior as a function of the platform's ranking weights and derive closed-form expressions for key outcome measures: engagement, misinformation, and polarization. We also define an index that reflects both the platform's interest in maximizing engagement and the societal interest in minimizing informational distortions. Our results reveal that the platform-optimal highlighting weight (η) is generally high and detrimental in terms of misinformation and polarization. That is, the very same ranking parameters that increase user activity also degrade the informativeness and balance of the content consumed. We then discuss a simple policy instrument—a per-unit tax on highlights ("engagement tax")—and show that it can mitigate these negative externalities by altering platform incentives in the design of profit-maximizing algorithms. The tax enters the platform's objective as a marginal cost on highlighted content. This reduces the platform's optimal weight on highlights, thereby shifting the algorithm toward rankings that perform better in terms of misinformation and polarization. The result parallels standard Pigouvian logic: a tax can correct a misalignment between private and social marginal benefits without relying on heavy policy regulations (e.g., banning highlights, mandating algorithmic structure or directly measuring the misinformation contained in social media platforms). We also discuss implementation strategies, such as differentiating taxes by content domain (e.g., politics vs. entertainment).

We extend the model to incorporate personalized rankings, where the content shown to an individual is shaped more heavily by the behavior of like-minded users. We show that personalization further increases polarization by reducing cross-cutting exposure and reinforcing echo chambers. Unlike the highlighting weight, however, personalization has limited effects on misinformation, yet continues to raise engagement.

To assess the empirical relevance of the model's mechanisms, we examine Facebook's "Meaningful Social Interactions" (MSI) update, implemented in early 2018. This algorithmic change significantly increased the platform's emphasis on user shares and reactions in determining content rankings, effectively raising η in our framework.[2] Using representative survey data from Italy and the United States and estimating a Differences-in-Differences empirical model, we document that, after the MSI update, individuals relying on the internet (and in particular Facebook) for political information became significantly more likely to report extreme ideological positions and exhibited greater affective polarization relative to non-users. These findings are consistent with the model's prediction that increasing η amplifies political extremism and inter-group hostility.[3]

Taken together, our findings underscore the potentially divergent objectives of platforms and society. While engagement-based ranking systems can successfully capture attention, they may do so at the cost of truth and social cohesion. Our model provides a tractable framework to analyze these tradeoffs and to inform the ongoing debate about the regulation and design of algorithmic information environments.

1.1 Related Literature

To the best of our knowledge, this is the first paper providing a theoretical framework—and related empirical evidence— to assess how the dynamic feedback between recommendation algorithms and individuals' behavior may affect platform engagement, misinformation and polarization. The model generalizes and extends Germano et al. (2019); Germano and Sobbrio (2020) to the context of a social media platform, giving individuals the option to highlight content, thereby further affecting the ranking, besides allowing for a broader set of individual clicking behavior and also a larger signal space. Moreover, we model individuals as positioned on a one-dimensional line (e.g., representing left to right) and abstract from modeling a fully-fledged user network. In this sense, our model is complementary to the one of Acemoglu et al. (2024) who study endogenous social networks and fact-checking.[4] Acemoglu et al. (2024) show that platforms

[2] The MSI update also increased the extent of personalization of rankings as it assigned a multiplier between 0.3 and 0.5 to contents/reactions from indirect links (e.g., non-friends). As discussed in Section 5.2, our theoretical framework suggests that an increase in personalized rankings further contributes to political polarization. For details on the MSI algorithmic update see Horwitz (2023) as well as `https://www.documentcloud.org/documents/21093256/pages/1/?embed=1`; `www.edition.cnn.com/2021/10/27/tech/facebook-papers-meaningful-social-interaction-news-feed-math` and `https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215?mod=article`.

[3] It is worth noticing that there is a growing body of literature documenting the spillover effects of social media on legacy media (see Hatte et al. (2021); Cagé et al. (2022)). In the presence of spillover effects of the MSI update on the control group, our results should be interpreted as a lower bound for the actual effect of the MSI update on polarization.

[4] See also Hsu et al. (2024) who study misinformation spread as a dynamic game on an exogenous social network. Törnberg et al. (2023) study simulations of an agent-based model in conjunction with a large language model to evaluate the effect of

have an incentive to increase personalization (in the sense of preferring a more homophilic sharing network) as this increases platform engagement. In their setting, this is detrimental in terms of social welfare as it increases the level of misinformation. In a related setting, Acemoglu et al. (2025) further study the impact of social media on polarization. While Acemoglu et al. (2024, 2025) model the platform as designing the sharing network of users, we study a platform that decides on the popularity weights of the ranking algorithm. Importantly, by explicitly modeling the endogenous dynamic ranking that emerges from the platform's algorithm, we are able to provide insights on the incentives—and possible perverse effects on misinformation and polarization—of such platforms to boost the relative weight given to the highlighting of content in their ranking algorithm.[5] Our framework also connects to broader questions about platform design, incentives, and the societal consequences of attention-maximizing algorithms (e.g., Allcott et al. 2022; Matias and Wright 2022; Matias 2023; Narayanan 2023)

Besides the direct empirical evidence provided in the paper, the predictions of our model are also consistent with those of several recent papers studying social media. In particular, our theoretical framework emphasizing the role of ranking algorithms (designed to maximize engagement) on polarization speaks directly to Braghieri et al. (2025), who show that the exposure to Facebook's feed—by itself—accounts for 82% of the degree of polarization in news consumption on the platform. Drawing from a dataset of around 208 million US-based adult active Facebook users, González-Bailón et al. (2023) shows that ideological segregation is amplified by Facebook's ranking algorithm. The empirical predictions of our paper also relate to Guess et al. (2023a), who present a unique randomized experiment looking at the impact of algorithmic vs. chronological feed exposure on Facebook. The study shows that the algorithmic feed leads to large changes in online behavior, such as more platform engagement (both in terms of time spent on the platform and likes), more exposure to like-minded sources, and less exposure to moderate/mixed sources. In a parallel randomized experiment, Guess et al. (2023b) look at the impact of reshares on Facebook. The analysis documents that shutting down reshares leads to less platform engagement (less time spent on Facebook and fewer clicks), more exposure to moderate news sources, and fewer partisan news clicks. The findings of Guess et al. (2023a,b) document a clear trade-off between platform engagement and exposure to moderate sources, which is very much in line with the key insights of our theoretical model. Yet, Guess et al. (2023a,b) also show that the significant changes in online behavior and news exposure induced by

three different types of ranking algorithms on polarization and toxicity of speech; Chavalarias et al. (2023) study calibrated simulations of different recommendation algorithms optimized for engagement on individuals posting and sharing content on an endogenous network.

[5] Our work is also complementary to papers that—while abstracting from the feedback loop between user behavior and recommendation algorithms—provide alternative mechanisms linking social media, misinformation and polarization. Azzimonti and Fernandes (2022) present a model of diffusion of misinformation on social media via internet bots. van Gils et al. (2024) provide insights on the impact of microtargeting on political outcomes. Beknazar-Yuzbashev et al. (2024) show that advertising-driven platforms can find it profitable to display content that harms users when it is complementary to their time spent on the platform.

algorithmic ranking and reshares, respectively, do not translate into significant changes in political attitudes (e.g., affective polarization). As suggested by Guess et al. (2023a,b), this discrepancy, which is also in contrast with our results in terms of misinformation and polarization, may be due—among other factors—to the short time window of their study (three months before the 2020 US election) and to the fact that such randomized intervention on a small (and self-selected) fraction of platform users may miss potentially relevant general equilibrium effects of algorithmic changes. In this respect, we can reconcile our empirical findings with the ones of Guess et al. (2023a,b). Indeed, we document a significant impact of an algorithmic change boosting the weight given to reshares and likes in the ranking on ideological extremism and affective polarization, as potentially arising from the general equilibrium effects of a population-wide change in the algorithm. Importantly, our event-study estimates suggest that the impact of the algorithm change on ideological extremism may not materialize on impact, thus further suggesting that it may take time to translate changes in online behavior into changes in political attitudes and knowledge. Similarly, Kalra (2021) provides evidence from India showing that recommendation algorithms tend to increase user engagement while fostering toxic content. The trade-off between social media engagement and negative externalities predicted by our model also speaks directly to the empirical evidence by Beknazar-Yuzbashev et al. (2022), who shows that curbing toxic content on social media platforms leads to a reduction in user engagement.[6]

The remainder of the paper is structured as follows. Section 2 introduces the theoretical model, describing the behavior of individuals and the platform's ranking algorithm. Section 3 characterizes the equilibrium clicking and highlighting distributions and analyzes the effects of ranking parameters on engagement, misinformation, and polarization. Section 4 presents the implications of our model and discusses how an *engagement tax* may reduce misinformation and polarization by affecting platform incentives. Section 5 discusses extensions of the model allowing for: (a) alternative distributions of highlighting behavior; (b) incorporating personalized rankings. Section 6 provides empirical evidence based on the 2018 Facebook "Meaningful Social Interactions" algorithmic update using survey data from Italy and the United States. Section 7 concludes.

2 The Model

At the center of the model is a digital platform characterized by its recommendation algorithm, which ranks and directs individuals to different news items (e.g., websites, Facebook posts, tweets), based on the

[6] The theoretical predictions of the model pointing out the role of social media algorithms in fostering misinformation, are also consistent with Vosoughi et al. (2018), who provide evidence that false stories spread faster than true ones on Twitter. Similarly, Mosleh et al. (2020) points out the presence of a negative correlation between the veracity of a news item and its probability of being shared on Twitter.

popularity of individuals' choices. Such news items may be used by individuals to obtain information on an unknown, cardinal *state of the world* $\theta \in \mathbb{R}$ (e.g., net benefit of a vaccine, consequences of inaction on global warming, adequacy of a presidential candidate, etc.).

The ranking of each news item is inversely related to its popularity, where the popularity is determined by the number of clicks and the number of "highlights" received by the given item (e.g., likes received by a Facebook post/number of shares, likes, retweets of a tweet, etc.). Each click has a weight of one, and each "highlight" has a weight of $\eta \geq 0$, so that a subject clicking and highlighting an item increases its popularity by $1 + \eta$. We briefly illustrate the working of the model before entering into some more details regarding the ranking algorithm.

2.1 Individual Clicking and Highlighting Choices

There are N *individuals*, each of whom receives a private informative signal on the state of the world, $x_n \in \mathbb{R}$, which is drawn randomly and independently from $N(\theta, \sigma_x^2)$ (we use $f(x; \theta, \sigma_x^2)$ to denote the corresponding density function). There are also $M > 2$ *news items* that individuals can click on, each of which carries an informative signal on the state of the world $y_m \in \mathbb{R}$, also drawn randomly and independently from $N(\theta, \sigma_y^2)$ (we use $f(y; \theta, \sigma_y^2)$ to denote the corresponding density function).

2.1.1 Individual clicking choices absent ranking

To model individuals' clicking choices, we assume there is a benchmark $\widehat{\theta} \in \mathbb{R}$—non-informative with respect to θ—which allows individuals to sort news items into "like-minded" or not. Specifically, we assume that, besides the order of the news items provided by the ranking algorithm, individuals are able to see whether a news item is reporting a "like-minded" information or not. Yet they need to click on the news item in order to see the actual signal y_m. This assumption is meant to capture a rather typical situation, where individuals observe the "coarse" information provided in the landing page by the platform (e.g., infer the basic stance of a news item, whether Left or Right, pro or anti something, from the website title, Facebook post intro, first tweet in a thread, etc.), yet, in order to learn the actual content of the news (i.e., the cardinal signal y_m) and update her beliefs, the individual has to click on the news item.

We formally translate this setting into assuming that an individual is able to observe whether her own signal x_n and the news items' signals y_m are above or below $\widehat{\theta}$. Accordingly, for each individual, the signal x_n has an associated binary signal indicating whether x_n is above or below $\widehat{\theta}$: $\text{sgn}(x_n) \in \{-1, 1\}$, where $\text{sgn}(x_n) = -1$ if $x_n < \widehat{\theta}$ and $\text{sgn}(x_n) = 1$ if $x_n \geq \widehat{\theta}$. Similarly, for each news item, the signal y_m has an associated binary signal $\text{sgn}(y_m) \in \{-1, 1\}$, where $\text{sgn}(y_m) = -1$ if $y_m < \widehat{\theta}$ and $\text{sgn}(y_m) = 1$ if $y_m \geq \widehat{\theta}$. Let M_- and M_+ denote the sets of news items with binary signal respectively -1 and 1 (by slight abuse of notation, we use $\#M_-$ and $\#M_+$ or sometimes directly M_- and M_+ to denote the number of news items in M_- and M_+ respectively). Thus, given the individual's signal x_n, the benchmark $\widehat{\theta}$ allows the individual to sort news items into "like-minded" or not, before actually clicking or reading. For most of the paper, we focus on the case where the benchmark separates signals in roughly symmetric groups: $\widehat{\theta} \approx \theta$.[7]

At the same time, as mentioned, individuals have to click on news item m in order to learn its cardinal signal y_m. Following Germano et al. (2019); Germano and Sobbrio (2020), we model clicking behavior by means of a stochastic choice rule accounting for the fact that two websites m and m' with the same sign are perfect substitutes (since two items with the same sign $(y_m, y_{m'}$ with $\mathrm{sgn}(y_m) = \mathrm{sgn}(y'_m))$ appear identical to an individual before clicking).[8] Assuming each individual derives value $\gamma_n \in (0,1)$ from clicking on a like-minded news item and $1-\gamma_n$ for clicking on a non-like-minded item, this readily implies that the Luce choice rule propensity is simply γ_n for clicking on a like-minded item and $1-\gamma_n$ for clicking on a non-like-minded item. For simplicity, we assume individuals can be of three clicking types $k \in \{C, E, I\}$:

- *confirmatory type* $(k = C)$: $\gamma_n = \gamma_C > 1/2$;
- *exploratory type* $(k = E)$: $\gamma_n = \gamma_E < 1/2$;
- *indifferent (purely ranking-driven) type* $(k = I)$: $\gamma_n = \gamma_I = 1/2$.

These three types occur with probabilities, respectively, $p_C, p_E, p_I \geq 0$, such that $p_C > 1/2$ and $p_C + p_E + p_I = 1$.[9] While we allow for all three types, all the key results of the model hold even if we were to focus on only one or two of the types (always including the confirmatory types).[10] Dividing by the number of items in the relevant class (whether like-minded or not) to keep track of ex ante identical items, we can write individual n's *ranking-free choice function for clicking* on item m when her type is k as:

$$\varphi_{n,m}(k) = \begin{cases} \gamma_k/[m] & \text{if } \mathrm{sgn}(x_n) = \mathrm{sgn}(y_m) \\ (1-\gamma_k)/[m] & \text{if } \mathrm{sgn}(x_n) \neq \mathrm{sgn}(y_m), \end{cases} \tag{1}$$

where $k \in \{C, E, I\}$ and $[m] = \#M_-$ if $\mathrm{sgn}(y_m) = -1$ and $[m] = \#M_+$ if $\mathrm{sgn}(y_m) = 1$.

[7] Say, $|\widehat{\theta} - \theta| < \min\left\{\frac{\sigma_x}{4}, \frac{\sigma_y}{4}\right\}$. In Appendix B.2 we discuss the case where individuals have heterogeneous benchmarks $\widehat{\theta}_n$; Appendix B.3 discusses the case, where $\widehat{\theta}$ and θ are far apart.

[8] See Luce (1959) and Brock and Marshack (1960) for early papers on stochastic choice rules (Luce rules) and Gül et al. (2014) for rules allowing for identical alternatives.

[9] Similar to the literature on political economy that parametrizes the fraction of different types of voters (Krasa and Polborn, 2009; Krishna and Morgan, 2011; Galasso and Nannicini, 2011), the model does not micro-found the individuals' clicking choices. At the same time, it is easy to see that the confirmatory type might be driven by a preference for like-minded news (Mullainathan and Shleifer, 2005; Bernhardt et al., 2008; Gentzkow and Shapiro, 2010; Sobbrio, 2014; Gentzkow et al., 2015). Yom-Tov et al. (2013); Flaxman et al. (2016); White and Horvitz (2015); Braghieri et al. (2025) provide empirical evidence on confirmation bias by users of digital platforms. Similarly, the exploratory type might be the by-product of incentives to cross-check different information sources (Rudiger, 2013; Athey et al., 2018). Finally, the indifferent type allows us to consider the role of individuals with a high attention bias or search cost (Pan et al., 2007; Glick et al., 2014; Novarese and Wilson, 2013).

[10] Note that we assume clicking types to be independent of individuals' priors x_n, but when formalizing the "highlighting" behavior, we allow highlighting types to be correlated with the priors.

2.1.2 Individual clicking choices with ranking

So far, the ranking of items did not enter the clicking choice. We now allow the individual choice function to take into account that individuals see the news items ordered by the ranking $r_n = (r_{n,m})_{m \in M}$, where $r_{n,m}$ is the rank of news item m as seen by individual n. We assume individuals have an *attention bias* calibrated by the parameter $1 < \beta < 2$, with the interpretation that, a news item of equal sign but placed one position higher in the ranking, has a likelihood β times larger to be clicked on than the lower ranked one (Pan et al., 2007; Glick et al., 2014; Novarese and Wilson, 2013). Together with the propensity to click, absent ranking, these jointly determine the probability with which individuals click on news items. Define the *probability of individual n of clicking type k to click on news item m* as:

$$\rho_{n,m}(k) = \frac{\beta^{(M-r_{n,m})}\varphi_{n,m}(k)}{\sum_{m' \in M} \beta^{(M-r_{n,m'})}\varphi_{n,m'}(k)}. \tag{2}$$

2.1.3 Individual highlighting choices

After clicking on a given news item m, the individual sees the actual signal $y_m \in \mathbb{R}$ and then decides whether or not to *highlight* that item (e.g, like, share, retweet, etc.). This depends on the individual's *highlighting type* $h \in \{A, P\}$. We consider two highlighting types:

- *passive type* ($h = P$): never highlights a news item regardless of her signal;
- *active type* ($h = A$): highlights a news item if and only if the news item's signal is sufficiently close to her own signal, $y_m \in H(x_n)$,

where $H(x_n) \equiv [x_n - \sigma_x/2, x_n + \sigma_x/2]$ and σ_x is the standard deviation of the individual's signal x_n. We assume the active and passive highlighting types occur with probabilities $p_A, p_P \geq 0$, respectively, where $p_A + p_P = 1$. Importantly, the probability of an individual's highlighting type is a function of the signal, where for the *probability of being an active type* p_A, we assume:

$$p_A(x_n) = 1 - e^{-\frac{1}{2\alpha}\left(\frac{x_n - \widehat{\theta}}{\sigma_x}\right)^{2\alpha}}, \quad \alpha \geq 1\,. \tag{3}$$

Thus, we assume: (a) that an individual highlights only if the news item reports a signal sufficiently close to her prior ($y_m \in H(x_n)$) (An et al., 2014; Pogorelskiy and Shum, 2019; Garz et al., 2020; Konovalova et al., 2023; Braghieri et al., 2025);[11] (b) the highlighting probability is increasing in the absolute value of the individual's signal. Specifically, p_A increases with the (square of the) deviation of x_n from the benchmark $\widehat{\theta}$, normalized by σ_x. And, while the specific functional form assumed in Eq. (3) is not crucial

[11] For example, Braghieri et al. (2025), page 30, show that "liberals share left-leaning content around three times more often than moderate content and almost never share right-leaning articles. Conservatives display the mirror image of this behavior: they primarily share right-leaning content, and they almost never share left-leaning articles."

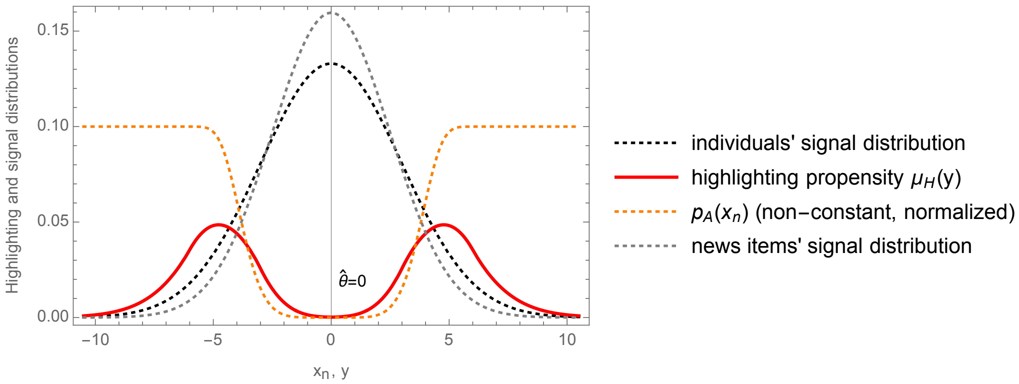

Figure 1: Individuals' signal distribution and highlighting propensity and items' signal distribution for $\widehat{\theta} = \theta = 0$ and $\sigma_y^2 < \sigma_x^2$.

for our results, what does matter is that individuals with more extreme priors are more likely to be active and hence to highlight a given news item y_m.[12]

A function that plays an important role in our analysis is the *highlighting propensity* for an item with signal y:

$$\mu_H(y) = \int_{x \in H^{-1}(y)} p_A(x) f(x; \sigma_x^2) dx, \tag{4}$$

where $H^{-1}(y) = \{x \in \mathbb{R} \,|\, y \in H(x)\}$. It gives the mass of highlights to be expected for an item of signal y (absent ranking). Similarly, the *clicking propensity* for an item is the mass of expected clicks. It is constant and does not depend on the item's signal, apart from its sign (again absent ranking). Given the assumed symmetry of behavior for subjects with positive and negative priors, the clicking propensity is the same constant whether the signal is positive or negative, and we normalize it to 1.

Figure 1 shows the assumed signal distributions of individuals (black dashed line) and of news items (gray dashed line). It also shows the probability of being an active type (orange dashed line) and the resulting highlighting propensity (red solid line).

As with the clicking choice, individuals' highlighting choice is not derived from a maximization problem. Nevertheless, the positive correlation between extreme beliefs and propensity to highlight is reminiscent of the link between overconfidence and ideological extremism modeled by Ortoleva and Snowberg (2015). Importantly, the assumed highlighting behavior is rooted in observed empirical regularities on how individuals with more extreme ideological beliefs tend to be more actively engaged and also more likely to highlight political news items on social media platforms (Bakshy et al., 2015; Grinberg et al., 2019; Pew, 2019; Hopp et al., 2020; Fraxanet et al., 2025). In particular, by using data on over 10 million Facebook users in the US, Bakshy et al. (2015) provide evidence on the ideological distribution of shared political news.[13] The data clearly show a bimodal distribution with large mass on the tails of the distribution, i.e.,

[12] The results of Section 3.3 are also robust to replacing θ for $\widehat{\theta}$ in Eq. (3).

[13] More specifically, they measure the ideological alignment of content shared on Facebook as the average affiliation of sharers

individuals with more extreme preferences account for a larger proportion of the overall shared content on Facebook. Remarkably, such a bimodal distribution is present only when looking at the distribution of shares related to contents defined by Bakshy et al. (2015) as “hard” information (e.g., national news, politics, world affairs). By contrast, no such bimodality is present when looking at “soft” information (e.g., sport, entertainment, travel). This suggests that the bimodal distribution observed in the shares of “hard” information is unlikely to be driven by large tails in the ideological distribution of Facebook’s users (i.e., a bimodal distribution of Facebook users’ ideology) or by a larger density of the network in such tails. Rather, such a bimodal distribution of shares is likely to be driven by users with more extreme ideological preferences that have a higher propensity to share political content. Furthermore, the assumed correlation between extreme beliefs and probability to highlight a news content is also consistent with the evidence provided by Braghieri et al. (2025) showing that extreme articles are approximately 4 times more likely to be heavily shared on Facebook (i.e., shared at least 100 times) and that liberal and conservatives account for 79% of total shares on Facebook. Remarkably, the study points out that the distribution of slant of the articles that are heavily shared on Facebook has much thicker tails than the distribution of slant on the production side.[14]

2.2 Platform and Ranking Algorithm

We consider *popularity-based rankings* that evolve as a function of the clicking and highlighting behavior of individuals.

2.2.1 Popularity ranking

After each individual makes her choices, the algorithm updates the popularity of each news item such that a click has a weight of 1 and a highlight has a weight of $\eta \in \mathbb{R}_+$. That is, starting from $\kappa_{0,m} \in \mathbb{R}_+$, the *popularity* of each item m, $\kappa_{n,m}$, for $n \geq 1$, is updated according to:

$$\kappa_{n,m} = \kappa_{n-1,m} + \begin{cases} 0 & \text{if } m \text{ is not clicked on by } n \\ 1 & \text{if } m \text{ is clicked on and not highlighted by } n \\ 1+\eta & \text{if } m \text{ is clicked on and highlighted by } n. \end{cases} \tag{5}$$

weighted by the total number of shares. Similar evidence is presented by the authors when weighting by the total number of distinct URL shared.

[14] See also Grinberg et al. (2019); Pew (2019); Hopp et al. (2020) for additional empirical evidence on the positive correlation between extremist political preferences and the propensity to share political news on social media. Fraxanet et al. (2025) further documents U-shaped patterns of being active for various types of engagement on Facebook.

The *ranking* of the news that individual n sees $(r_{n,m})_{m\in M}$ is inversely related to the popularity before she clicks:

$$r_{n,m} < r_{n,m'} \iff \kappa_{n-1,m} > \kappa_{n-1,m'}. \tag{6}$$

We also keep track of the traffic a news item receives without counting the highlights. Hence, starting from $\widehat{\kappa}_{0,m} = \kappa_{0,m} \in \mathbb{R}_+$, the *number of clicks* on news item m, $\widehat{\kappa}_{n,m}$, for $n \geq 1$, is updated according to:

$$\widehat{\kappa}_{n,m} = \widehat{\kappa}_{n-1,m} + \begin{cases} 0 & \text{if } m \text{ is not clicked on by } n \\ 1 & \text{if } m \text{ is clicked on by } n. \end{cases} \tag{7}$$

Similarly, we track keep of the number of highlights a news item receives, without counting the clicks. Thus again, defining $\widetilde{\kappa}_{0,m} = \kappa_{0,m} \in \mathbb{R}_+$, the *number of highlights* of news item m, $\widetilde{\kappa}_{n,m}$, for $n \geq 1$, are updated accordingly.

2.3 Evaluation Indices

To evaluate the effects of the highlighting parameter of the ranking algorithm, we formally define a few indices. Let $y(n) \in M$ denote the signal of the news item clicked on by individual n, and let L (R) denote the individuals with signals x_n with $\text{sign}(x_n) = -1$ $(= +1)$. Then we can define the following indices:

- *Engagement* on item m by group g: $ENG^g_m = \widehat{\kappa}^g_{N,m} + \widetilde{\kappa}^g_{N,m}$ (clicking and highlighting by group g)
- *Total Engagement*: $ENG = \sum_{m\in M} \left(ENG^L_m + ENG^R_m\right)$ (total clicking and highlighting)
- *Misinformation*: $MIS = \frac{1}{N}\sum_{n\in N} |y(n) - \theta|$
- *Polarization*: $POL = \frac{1}{N}\left|\sum_{n\in R} y(n) - \sum_{n'\in L} y(n')\right|$.

The first two indices aim to capture a key dimension of what digital platforms care about: user engagement, or the expected amount of activity generated by the individuals.[15] The third index is a straightforward measure of misinformation capturing the average distance between the information carried by the news items chosen by individuals and the true state of the world. The fourth index measures polarization as the average distance between the information provided by the news items chosen by individuals in group R with the respect to the ones chosen by individuals in group L.[16]

[15] In particular, the willingness to increase engagement was behind the update (boost in η in our model) implemented in 2018 by Facebook with the stated objective of increasing *meaningful social interactions* (see Section 6 for a related discussion).

[16] Notice that we abstract from the specific belief updating of each individual. Our focus is on the comparative static effect of changes in the algorithm parameter (η) on misinformation and polarization. Accordingly, the proposed misinformation and polarization indices will be informative on such effects as long as individuals update their beliefs in the direction of the signal carried by the news item they click on, $y(n)$.

3 Engagement, Misinformation and Polarization

In this section, we study the dynamic interplay between individuals' clicking and highlighting behavior and the platform's popularity ranking algorithm. To simplify the discussion, throughout this section we assume that $\widehat{\theta} = \theta$. That is, we focus on the symmetric case where signals are symmetrically distributed around the benchmark.[17] We first present a preliminary discussion of the mechanism linking η and the dynamics of clicking and highlighting. We then provide analytical results characterizing limit clicking and highlighting distributions and the impact of η on the key indices introduced in Section 2.3.

3.1 Boosting Meaningful Social Interactions: Crowding out the truth

A key objective of our model is to understand the effect of the weight of a highlight (η) on engagement, misinformation, and polarization.

It turns out that increasing the weight on highlighting (η) can have desirable properties for the platform, namely, higher engagement, but not necessarily for users since it can result in higher misinformation and higher polarization. To see this, notice first that the combination of normally distributed priors and a propensity to highlight that increases in the extremeness of the prior leads to highlighting behavior that is bimodal (see Section 2.1.3, Figure 1). An important intermediary result shows that, as η becomes large and highlighting becomes relatively more important, the clicking distribution inherits the basic shape of the highlighting distribution. The reason is that, due to the ranking algorithm, when η is large, items that have a high propensity of being highlighted go higher up in the ranking, and highlighting behavior becomes more important as a driver of what individuals see as being ranked prominently and ultimately end up clicking on.

As a result, as η increases, the clicking distribution goes from being roughly unimodal and centered around the true state θ when η is small to being increasingly bimodal as η gets large. Since the clicking distribution reflects what people read, this shows that a higher η increases both misinformation and polarization. This also leads to higher engagement (measured as the sum of clicks and highlights). This is illustrated in Figure 2, where Panels A and B show the clicking distribution for the cases of respectively small and large η. We refer to this phenomenon, whereby individuals are less likely to click on items close to the true state and more likely to click on ones further away, due to a higher parameter η, as *crowding out the truth.* As mentioned in the introduction, the case of large η seems to capture what Bail (2021) refers to as the social media "prism", which he argues besides fueling extremism and polarization, mutes moderates

[17] Allowing for heterogeneous benchmarks across individuals $\widehat{\theta}_n$ does not qualitatively affect the results; see Appendix B.2. Similarly, allowing for small asymmetries in the distribution of signals with respect to the benchmark $\widehat{\theta}$ also does not affect the results qualitatively ($|\widehat{\theta} - \theta| < \min\left\{\frac{\sigma_x}{4}, \frac{\sigma_y}{4}\right\}$) . However, allowing for large asymmetries may change the results; Appendix B.3 discusses the case where $\widehat{\theta}$ and θ are far apart.

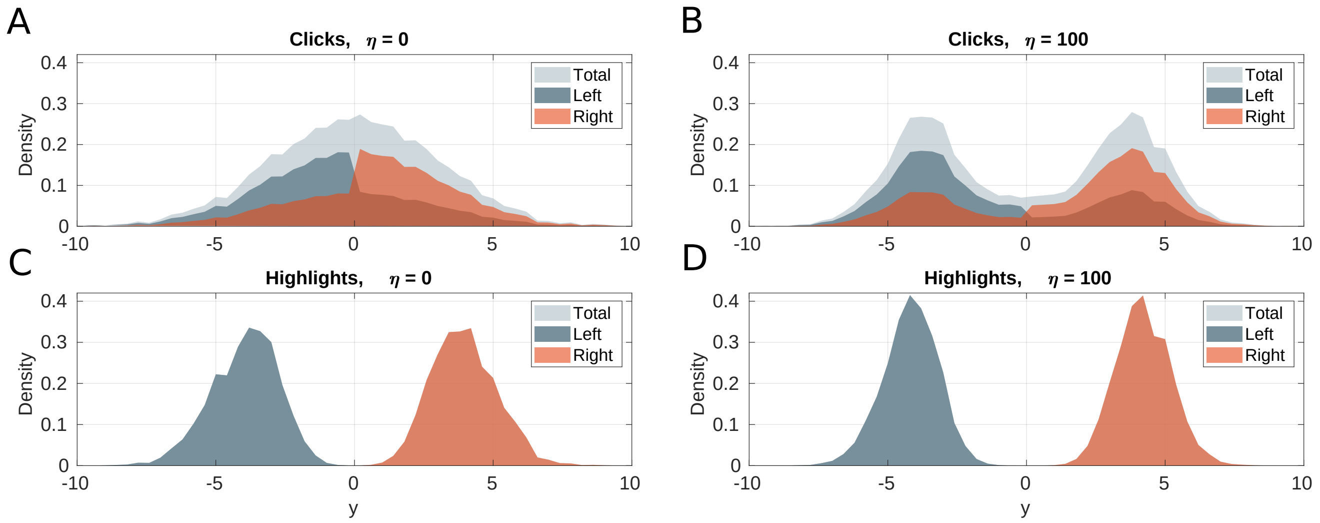

Figure 2: Polarization and misinformation *increase* for increasing values of η. Panels (A) and (B) show users' clicking behavior for $\eta = 0$ and $\eta = 100$, respectively. Panels (C) and (D) show users' highlighting behavior for $\eta = 0$ and $\eta = 100$, respectively. Polarization increases from an average value of 1.8 (SD 0.5) to 2.8 (SD 0.4), and misinformation increases from an average value of 2.4 (SD 0.6) to 3.6 (SD 0.4).

and gives a distorted image of others. To see this, consider Panel B, where we see that individuals on the left, for example, are more likely to click on extreme items from the left when clicking on the left (blue), but are also more likely to click on extreme items from the right when clicking on the right (shaded red on the right). This suggests that they will get a more extreme impression of individuals both on the left and on the right and, given that priors are centered around θ, this is also consistent with higher *perceived* polarization on social media (see also Yang et al. (2016)). In other words, the highlighting parameter η can be seen as directly contributing to the "prism" effect of Bail (2021). Panels C and D, finally, show the highlighting behavior in the case of small and large η respectively, and where it can be seen that there is more highlighting in D than in C, suggesting that engagement increases with η.

In the subsections that follow, we look at the above phenomena more formally and also connect them to metrics of platform's engagement and users' information and polarization.

3.2 Limit Clicking and Highlighting Behavior

In order to evaluate the impact of the parameters of the algorithm (η), it is useful to obtain the actual clicking and highlighting behavior of the individuals, while keeping track of the dynamic feedback between clicking, highlighting and ranking. We do this by characterizing the limit clicking and highlighting distributions, since these ultimately determine what to expect in terms of engagement and changes in posterior beliefs. These are computed for an arbitrarily large number of repetitions of the process described in Sections 2.1 and 2.2. Characterizing the feedback loop of such a ranking process while keeping track of the stochastic behavior of the subjects is not a straightforward task, yet it is fundamental to compute

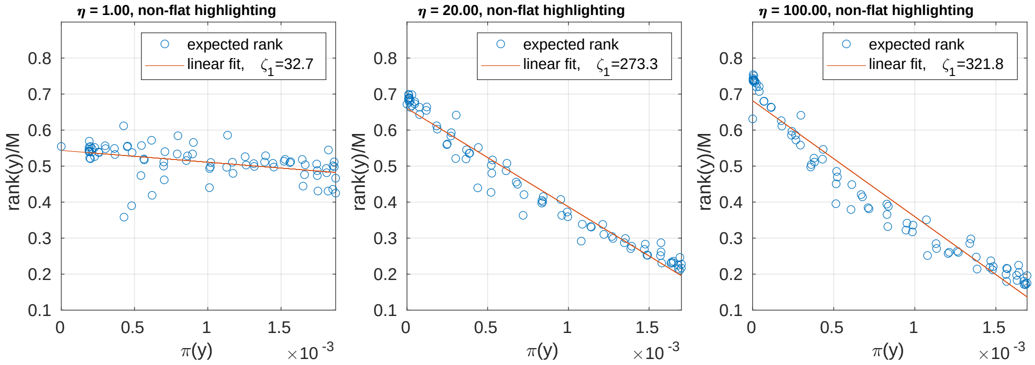

Figure 3: Linear dependence between the expected rank (blue circles, obtained from simulations) and the expected popularity $\pi(y)$ as assumed in Eq. (9). Red line denotes the best linear fit. To compute each blue dot, we binned the item's signals into 81 bins (bin-size = 0.2) and compute the mean popularity and the corresponding mean rank from $T = 10^3$ experiments, each of them with different $M = 20$ item's signals and different $N = 5 \cdot 10^3$ individual's signals.

comparative statics and understand the role of algorithms when evaluating potential effects on opinion formation (Matias and Wright, 2022; Narayanan, 2023).

A central feature of the ranking algorithm is its dependence on the popularity of different items. Define the *expected popularity of an item with signal y, absent ranking*, as the sum of the expected clicking and highlighting propensities, absent ranking, where highlighting is weighted by η:

$$\pi(y) = \frac{1 + \eta \cdot \mu_H(y)}{M \cdot (1 + \eta \cdot \bar{\mu}_H) + \eta \cdot (\mu_H(y) - \bar{\mu}_H)}, \tag{8}$$

where recall $\mu_H(y) = \int_{x \in H^{-1}(y)} p_A(x) f(x; \sigma_x^2) dx$ and $\bar{\mu}_H = \int \mu_H(y) f(y; \sigma_y^2) dy$.[18]

An assumption that we maintain in this and the following section is that the expected rank (over arbitrarily many repetitions) of a given item with signal $y_m = y \in \mathbb{R}$ is approximated by a linear decreasing function of the expected popularity of that item, absent ranking:

$$r(y) \approx \zeta_0 - \zeta_1 \cdot \pi(y), \tag{9}$$

where $\zeta_0, \zeta_1 > 0$ are constants. Simulations of the model with large numbers of individuals (N) and of repetitions (T) seem to consistently support the assumption, see Figure 3.[19]

This assumption significantly simplifies the analysis. It allows us to directly derive the effects of the popularity parameter and to characterize both the limit clicking and highlighting distributions, which

[18] Eq. (8) is derived from:

$$\pi(y_m) = \frac{\int \sum_{k \in \{C,E,I\}} p_k \cdot \varphi_{n,m} \cdot \left(1 + \eta \cdot p_A(x_n) \cdot 1_{\{x_n \in H(y_m)\}}\right) f(x; \sigma_x^2) dx}{\sum_{m' \in M} \int \sum_{k \in \{C,E,I\}} p_k \cdot \varphi_{n,m'} \cdot \left(1 + \eta \cdot p_A(x_n) \cdot 1_{\{x_n \in H(y_{m'})\}}\right) f(x; \sigma_x^2) dx},$$

taking averages over T repetitions for $T \to \infty$.

[19] Further simulations are available upon request from the authors.

represent expectations of T repetitions (for $T \to \infty$) of the process described in Sections 2.1 and 2.2.[20]

Proposition 1 (Limit Clicking and Highlighting Distributions)**.** *Assume Eq. (9), then the limit clicking distribution can be approximated as:*

$$LCD(y) \approx \Lambda_\beta(\pi(y)) \cdot f(y; \sigma_y^2), \tag{10}$$

where, Λ_β is a linear function with $\Lambda'_\beta > 0$ for $\beta > 1$, π is defined in Eq. (8). Accordingly, the limit highlighting distribution can be approximated as $LHD(y) \approx \mu_H(y) \cdot LCD(y)$.

The first result shows a basic feature of our ranking-based dynamics, namely, that the expected traffic on a given item is driven by its expected popularity, absent ranking. Attention bias enters through Λ_β as it's a strictly increasing function of $\pi(y)$ provided $\beta > 1$. A higher $\pi(y)$ increases the rank of an item with signal y, thereby increasing its traffic, the more so, the greater β is, for $\beta > 1$. With Eq.(8) we can write:

$$LCD(y) \propto (1 + \eta \cdot \mu_H(y)) \cdot f(y; \sigma_y^2), \tag{11}$$

meaning that, for any y, $LCD(y)$ is proportional to the expression on the right and directly shows how η affects limit clicking behavior. Both results follow from the linearity assumption in Eq. (9) combined with the functional form of the clicking probabilities assumed in Eq. (2). Notice, however, that the limit highlighting distribution is not normalized to integrate to 1, thereby allowing it to also capture the intensity of highlighting. Importantly, the limit clicking and highlighting distributions are used to compute the effects of algorithmic parameters on metrics of engagement, misinformation, and polarization, as well as to assess the effect of an engagement tax.

3.3 Engagement, Misinformation and Polarization: Analytical Results

We here focus on the comparative statics of the popularity parameter for highlighting (η).[21] As is clear from Section 3.2, as η increases, the expected popularity of an item and hence its expected traffic is increasingly driven by the highlighting propensity. This observation is a central message of the paper and has important consequences for how the parameter η affects engagement, misinformation and polarization.

Proposition 2 (Effect of η on Engagement, Misinformation and Polarization)**.** *Assume Eq. (9) and $\sigma_y \leq \sigma_x$, then increasing the weight on highlighting (higher η) increases user engagement, misinformation and*

[20] The analysis would otherwise require computing distribution over limit rankings, which is an open problem already without highlighting and with $M > 2$ items (Analytis et al., 2022). However, it can be verified that in the case of $M = 2$, using a result from Analytis et al. (2022), there is an exact linear relationship between the expected rank and the clicking propensity as assumed in Eq. (9).

[21] The willingness to increase engagement was behind the boost in η implemented in 2018 by Facebook with the stated objective of increasing *meaningful social interactions*; see Section 6 for a related discussion.

polarization.

Thus, increasing η increases engagement, but also has the adverse effect of increasing misinformation and polarization. It is possible to interpret these results in light of the evidence provided by Bakshy et al. (2015). As discussed above, Bakshy et al. (2015) point out that in the case of "hard" news (e.g, national, political), the propensity to highlight content is indeed higher for individuals with a more extreme prior, whereas the same does not apply to "soft" news (e.g., entertainment). The results suggest that social media platforms have an incentive to choose a high level of η as this results in a high level of engagement across all types of news content. Yet, while this might not be so much a concern for the "quality" of information consumed by the user in the case of "soft" news, we show it might have detrimental effects on misinformation and polarization when it comes to political news content. Accordingly, the results of Proposition 2 formalize the intuition of Narayanan (2023), suggesting that algorithms may be detrimental when it comes to non-entertainment domains.[22] Taken together, these results underscore how seemingly benign algorithmic design choices can have persistent and socially consequential effects. In particular, Proposition 2 reveals that even when a platform's ranking algorithm is not explicitly designed to promote ideological extremes, its interaction with heterogeneous user behavior can generate precisely that outcome. The dynamic feedback between engagement-driven ranking and user highlights implies that algorithmic neutrality cannot be assessed solely based on design intentions. This challenges claims made by industry representatives regarding the ideological impartiality of recommendation systems (Clegg, 2021). Rather, our results show that certain algorithmic features—such as greater weight on user highlights—can implicitly amplify extremism and misinformation, even under symmetric signal structures. These insights highlight the importance of analyzing algorithmic effects not in isolation, but as part of an evolving equilibrium shaped by user behavior and platform incentives.

Notice that, while our theoretical model is deliberately structured to be analytically tractable, the above comparative statics results are not immediate or mechanical consequences of the stated assumptions. The ranking function presented in Section 2.2 evolves endogenously, following the popularity of each item, where popularity is defined as a weighted sum of clicks and highlights. The weights of this function dynamically shape the visibility of the content, which in turn affects future user behavior, both clicking and highlighting. This creates a non-trivial, path-dependent mapping from algorithmic parameters to equilibrium outcomes. Proposition 1 is instrumental in allowing one to study this mapping.

The less trivial effects of increasing η are those on polarization and misinformation. They emerge

[22] See Narayanan (2023), page 37: "Each institution has a set of values that make it what it is, such as fairness in journalism, accuracy in science, and aesthetic values in art. Markets have notions of quality, such as culinary excellence in restaurants and professional skill in a labor market. Over decades or centuries, they have built up internal processes that rank and sort what is produced, such as peer review. But social media algorithms are oblivious to these values and these signals of quality. They reward unrelated factors, based on a logic that makes sense for entertainment but not for any other domain."

from a subtle dynamic interaction between user behavior and algorithmic amplification, and rely in an important way on the U-shaped form of the probability of being active p_A (reflected in Eq. (3)).[23] As η increases, content highlighted by users with more extreme priors—who also highlight more frequently—gains disproportionate visibility. The corresponding items tend to reflect signals that are both more polarized and further away from the truth. Over time, the algorithm with a higher η steers attention away from moderate, truth-adjacent content, not because the content itself disappears, but because it generates less engagement, less highlights and hence also less popularity. The equilibrium outcome is a limit clicking distribution that is increasingly driven by the propensity to highlight μ_H, and is thus increasingly bimodal and with more mass away from the truth. The distortions arise even though neither users nor the platform are actively seeking to misinform or polarize, and even though the content pool itself is symmetric and unbiased. It is the feedback loop between heterogeneous highlighting behavior and engagement-based ranking that gradually pushes the ranking toward crowding-out of the truth.

As for the effect of η on total engagement, it is not obvious a priori that giving more weight to highlights increases engagement, since highlights occur only after a click, and only if the content is sufficiently aligned with the user's prior. Increasing η could, in principle, raise the visibility of niche content, thereby discouraging engagement from the broader (and more moderate) user base. That this does not occur—and that engagement increases in η—is a result of the amplification mechanism. Users who highlight more frequently indirectly drive up visibility for the content they favor, which in turn further attracts increased engagement, without discouraging clicking behavior from the less active user types.

4 The Social Impact of the Recommendation Algorithm

We now assess the impact of highlighting weights on different metrics linked to the platform's profit and to the information consumed by its users, and discuss how a tax may induce platforms to internalize the effect of such weights on polarization and misinformation.

4.1 Highlighting Weights

We consider two key dimensions when assessing the impact of the recommendation algorithm, namely, one based on what the platform cares about: generating high levels of engagement (ENG); and another linked to metrics that affect its users: misinformation (MIS) and polarization (POL). For convenience, we capture all these aspects in a single measure of the form:

$$W_\psi(\eta) = \psi \cdot ENG(\eta) - (1-\psi) \cdot MIS(\eta) \cdot POL(\eta), \tag{12}$$

[23] In Section 5.1, we study the case where p_A is not U-shaped, but rather what we call flat, and show that increasing η actually decreases both polarization and misinformation, while the positive effect on engagement continues to hold.

where $0 \leq \psi \leq 1$ is a weight for the relative importance of the platform's likely goal (high ENG) relative to the metrics reflecting the information consumed by its users (low MIS and POL). From the analysis of the previous sections, we can show the following:[24]

Proposition 3. *Assume Eq. (9) and $\sigma_y \leq \sigma_x$. For small values of ψ, ($\psi \approx 0$), (W_ψ) is maximized at $\eta \approx 0$, while for large values of ψ, ($\psi \approx 1$), it is maximized at arbitrarily large values of η.*

This presents a key result of the paper. It shows that a clear dichotomy arises between the perspective of the platform and that of the users with respect to the desirable weight to be assigned to the highlights. This resonates with the reports leaked by Facebook's whistle-blowers, who underscored the conflicting welfare effects created by the platform's 2018 "Meaningful social interactions" update, which boosted the weights given to content sharing and highlighting in the ranking algorithms. Indeed, while this change increased users' overall engagement on Facebook, it appears to also have led to an increase in the misinformation and polarization components as predicted by Propositions 2and 5 (for the case of non-flat p_A).[25] Section 6 presents direct empirical evidence in this regard.

The following section shows how a simple policy instrument, such as an engagement tax on highlights, might be beneficial in terms of reducing polarization and misinformation.

4.2 Correcting Platform Incentives: An Engagement Tax

As discussed in the previous section, the ranking weight η on highlights generates a divergence between the platform's objective (high engagement) and that of society at large (low misinformation and polarization). In this section, we study how a simple policy instrument—a per-unit "engagement tax" on highlights—may help reduce the misalignment between platform and the information consumed by its users. Suppose the platform is taxed at the rate $\tau \geq 0$ for each highlight. As before, we assume the platform receives value from total engagement—i.e., the sum of clicks and highlights—but now internalizes the monetary cost of highlights via the tax. The platform's objective function becomes:

$$\Pi(\eta; \tau) = ENG(\eta) - \tau \cdot H(\eta), \tag{13}$$

where $ENG(\eta)$ is total engagement as defined in Section 2.3, and $H(\eta) = \sum_{m \in M} \widetilde{\kappa}_{N,m}$ is the total number of highlights across all users and items under ranking weight η.

[24] Notice that the condition $\eta \leq 2M$ in Proposition 2 is not needed for the effect of increasing η on engagement, which is relevant when $\psi \approx 1$. This is clear from the proof of Proposition 2 in the Appendix.

[25] An aspect that may play a role in moderating the platform-optimal level of η is how it impacts advertising through its effect on content, besides the effect on engagement. For example, if advertisers were to dislike a too high level of misinformation/polarization, this could be reflected in W_ψ and, consequently, could affect the platform-optimal level of η. Nevertheless, for $\psi > 0$, i.e., as long as the platform cares about engagement, the level of η chosen by the platform will be higher than the one minimizing misinformation and/or polarization. For theoretical models on the role of advertising in affecting content moderation in social media, see Madio and Quinn (2024); Liu et al. (2021); Jiménez Durán (2022).

Note that $H(\eta)$ is increasing in η by Proposition 2. We assume that the tax revenue is rebated lump-sum to users or redistributed outside the platform, and hence does not directly enter user welfare considerations. Let $\eta^*(\tau)$ denote the platform's optimal highlight weight given tax rate τ. It is immediate to see that as τ increases, the marginal cost of η increases. Hence, $\eta^*(\tau)$ is decreasing in τ, as the first-order condition for the platform's maximization problem is:

$$\frac{dENG(\eta)}{d\eta} - \tau \cdot \frac{dH(\eta)}{d\eta} = 0.$$

Intuitively, the engagement tax reduces the marginal benefit of increasing η for the platform, thereby pushing it to adopt a lower ranking weight on highlights. Therefore, it follows immediately from Proposition 2 that a higher τ is associated with a lower level of misinformation and polarization.

This intuitive result shows that a tax on highlights can reduce the harm created by the platform when maximizing engagement. The appealing feature of this tax is that it can be implemented without banning highlights, dictating algorithmic structure, or directly measuring the misinformation contained in social media platforms; instead, the platform is left to internalize the costs it imposes on users. This result parallels standard Pigouvian logic: when the platform does not internalize the externalities its engagement-maximizing algorithm imposes (via increased polarization and misinformation), a tax can align private and social incentives. A few remarks are in order. In practice, such a tax could be implemented indirectly, e.g., by imposing regulatory costs proportional to engagement metrics (likes, shares), or by limiting monetization (e.g., ad impressions) on content with high virality. A more refined policy could consider differentiated taxation, e.g., taxing highlights only for certain content categories (e.g., politics, health) or imposing higher tax rates on types of engagement more heavily associated with polarization and misinformation (since different types of engagement can exhibit different degrees of U-shape, see e.g., Fraxanet et al. (2025)). If platforms respond by altering personalization (e.g., shifting λ to offset lower η), joint regulation of η and λ may be needed (see Proposition 5). It is also important to point out that Facebook's MSI update can be interpreted as a shift from a low-η to a high-η regime. The empirical evidence that we provide in Section 6 suggests that this move increased ideological extremism and affective polarization, consistent with the model. The engagement tax serves as a conceptual counterfactual: the negative externalities might have been curtailed if platforms had faced a cost for indirectly promoting extreme, high-virality content.

5 Extensions

5.1 Alternative Distributions of Highlighting Behavior

It is important to remark that in Section 2.1.3 we embed a specific (and empirically-driven) assumption concerning the distribution of highlighting behavior into our model (see Eq. (3)). Yet, our theoretical framework is more general in that it allows considering and studying alternative distributions, for example, by looking at how the specific distributions of likes, shares, or comments would translate into the algorithmically driven dynamics of polarization and misinformation.

As an illustration, we now discuss how the implications concerning limit clicking, polarization, and misinformation would change if one were to assume a non-bimodal distribution of highlights, like for example the one suggested by Bakshy et al. (2015) for soft-news (non-political information). Specifically, suppose now that the *probability of being an active type* p_A is *flat*, that is, instead of taking the form assumed in Eq. (3) in Section 2.1.3, assume it is constant for all values of x_n. Then increasing the weight on highlights (η) can be shown to have some nice properties for both the platform and the information consumed by its users, namely, it increases engagement, while also reducing both polarization and misinformation. To see this, suppose then that a constant share ($p_A \in (0,1)$) of individuals who read a given article are also willing to highlight it, provided the news item's signal is sufficiently close to the individual's signal ($y_m \in H(x_n)$). Then as η increases, articles that get highlighted increase in total popularity and hence go up higher in the ranking, meaning that they are in turn also more likely to get clicked on. Since both the news items' and the individuals' signals are normally distributed, there is a relatively higher mass of individuals with signals around the truth (θ) and so, such individuals are more likely to read and highlight articles closer to the truth. This pushes them further up in the ranking. Hence, higher values of η will tend to concentrate clicking around the truth. This decreases polarization and misinformation, and, because there are relatively more individuals with signals around the truth, due to their normal distribution, it also increases engagement. Thus increasing η in the flat case directly increases what Facebook calls *meaningful social interactions*, and at the same time concentrates clicking around the truth, thereby decreasing misinformation and polarization. This is illustrated in Figure B.1 in the Appendix and is to be contrasted with the non-flat case depicted in Figure 2 in Section 3.1. Hence, in the case of soft/non-political information, there is no negative externality generated by an engagement maximizing weight on highlights and accordingly also no need for an engagement tax.

5.2 Popularity Ranking with Personalization

Suppose now that the ranking, while still being based on popularity, can differ from one individual to another and, moreover, weighs clicks (and highlights) differently based on which individuals they are made by. Suppose the algorithm can deduce for each individual whether $x_n \leq \hat{\theta}$ of $x_n > \hat{\theta}$ (e.g., using data from browsing history), then individuals can be assigned by the algorithm into one of two groups L or R depending on whether $x_n \leq \hat{\theta}$ of $x_n > \hat{\theta}$. Choices to click and highlights are determined as before, but the difference is that now there are two rankings, $r^L_{n,m}$ and $r^R_{n,m}$, whereby individuals in L see $r^L_{n,m}$ when doing their search, while individuals in R see $r^R_{n,m}$. Moreover, the ranking of group $g \in \{L, R\}$ depends only in part on the clicks and highlights of individuals from the opposite group. Specifically, starting from $\kappa^g_{0,m} \in \mathbb{R}_+$, the *popularity for group g* of each news item m, $\kappa^g_{n,m}$, for $n \geq 1$, is updated according to:

$$\kappa^g_{n,m} = \kappa^g_{n-1,m} + \begin{cases} 0 & \text{if } m \text{ is not clicked on by } n \\ \lambda(n) & \text{if } m \text{ is clicked on and not highlighted by } n, \\ \lambda(n) \cdot (1+\eta) & \text{if } m \text{ is clicked on and highlighted by } n, \end{cases} \tag{14}$$

where $\lambda(n) = 1$ if $n \in g$ and $\lambda(n) = \lambda$ if $n \notin g$, and where $\lambda \in [0, 1]$ is a parameter of the personalized ranking algorithm and determines how much clicks and highlights from the opposite group $g' \neq g$ count for the ranking seen by group g. When $\lambda = 0$ each group sees a fully personalized ranking, independent of the clicks and highlights of the other group. When $\lambda = 1$ clicks from both groups count the same, so that the two rankings are identical, and we get back the case of a single ranking.

As before, the *ranking* of news item m that individual $n \in g$ sees $(r^g_{n,m})_{m \in M}$ is inversely related to the popularity of m before n clicks:

$$r^g_{n,m} < r^g_{n,m'} \iff \kappa^g_{n-1,m} > \kappa^g_{n-1,m'}, \quad g \in \{L, R\}. \tag{15}$$

We also keep track of traffic and engagement of news items separately for each group. Starting again from $\widehat{\kappa}^g_{0,m} = \kappa_{0,m} \in \mathbb{R}_+$, the *number of clicks by group* g on website m, $\widehat{\kappa}^g_{n,m}$, for $n \geq 1$ and $g \in \{L, R\}$, is updated according to:

$$\widehat{\kappa}^g_{n,m} = \widehat{\kappa}^g_{n-1,m} + \begin{cases} 0 & \text{if } m \text{ is not clicked on by } n \\ 1 & \text{if } m \text{ is clicked on by } n, n \in g \\ 0 & \text{if } m \text{ is clicked on by } n, n \notin g, \end{cases} \tag{16}$$

The same can be done by counting the highlights of group g without counting the clicks, $\widetilde{\kappa}^g_{n,m}$, for $g \in \{L, R\}$.

Considering this more general model with this simple form of personalization, it is not difficult to see that Propositions 1 on the limit distribution and Proposition 2 on the effect of η continue to hold for any degree of personalization ($\lambda \in [0,1]$).[26] The next result concerns the effect of the parameter λ.

Proposition 4 (Effect of λ on Engagement and Polarization)**.** *Assume Eq. (9) and fix $\eta \geq 0$ arbitrarily. Then increasing personalization (lower λ) increases user engagement and polarization, both when individuals' highlighting behavior is flat and when it is non-flat (p_A as in Eq. (3)).*

The fact that more personalization increases polarization is straightforward. Decreasing λ makes the rankings of the two groups increasingly less correlated, which in turn makes users in each group more likely to click on items carrying a signal of the same sign as their own. This directly increases the polarization measure POL. To see the effect on engagement, note first that users that are active types only share items that are close enough to their own signal ($y_m \in H(x_n)$). As λ decreases and the rankings become less correlated, users are more likely to see items that have signals closer to their own more prominently ranked, and are in turn also more likely to click on them. But since items that are more prominently ranked are more likely to be in the set $H(x_n)$, they are also more likely to be highlighted. Overall, whether highlighting behavior is flat or non-flat, a lower λ (more personalization) contributes to an increase in ENG.

One effect that the personalization parameter λ does not have in our model, differently from the highlighting parameter η, is that it does not significantly impact misinformation. This is due to the fact that it mainly contributes towards interchanging clicks made from one group on items with signals of the opposite sign with clicks made by individuals from the other group, who have signals of the same sign. While this contributes to increasing polarization it does not really affect misinformation.

We can now embed the personalization parameter λ in the W_ψ defined in Section 2.3 as follows:

$$W_\psi(\eta,\lambda) = \psi \cdot ENG(\eta,\lambda) - (1-\psi) \cdot MIS(\eta,\lambda) \cdot POL(\eta,\lambda), \tag{17}$$

From the analysis of the previous sections, we can show:

Proposition 5 (Socially Efficient Rankings)**.** *Assume Eq. (9) and $\sigma_y \leq \sigma_x$. If individuals' highlighting behavior is non-flat (p_A as in Eq. (3)), then, for small values of ψ, ($\psi \approx 0$), W_ψ is maximized at $(\eta,\lambda) \approx (0,1)$, while for large values of ψ, ($\psi \approx 1$), it is maximized at $(\eta,\lambda) \approx (\infty,0)$.*

If instead individuals' highlighting behavior is flat (p_A constant), then, for small values of ψ, ($\psi \approx 0$), W_ψ is maximized at $(\eta,\lambda) \approx (\infty,1)$, while for large values of ψ, ($\psi \approx 1$), it is maximized at $(\eta,\lambda) \approx (\infty,0)$.

The above proposition generalizes the key result of the paper in the presence of personalization. Namely, in the empirically relevant case of non-flat propensity to highlight, a clear dichotomy arises between the

[26] This is shown in the Appendix. Moreover, simulations of the model for arbitrary values of λ continue to support the linearity assumption of Eq. (9) in the model with personalization.

perspective of the platform and that of the users with respect to the desirable weight to be assigned to the highlights. This resonates with the reports leaked by Facebook's whistleblowers, who underscored the conflicting welfare effects created by the platform's 2018 "Meaningful social interactions" update, which boosted the weights given to content sharing and highlighting in the ranking algorithms. Indeed, while this change increased users' overall engagement on Facebook, it appears also to have led to an increase in misinformation and polarization, as predicted by Proposition 5 for the non-flat case.[27]

6 Empirical Evidence: Meaningful Social Interactions and Political Polarization

The theoretical predictions of our model discussed in Section 3.3 suggest that an increase in the weight given by the ranking algorithm to the "highlights" (an increase in η) will result in individuals being more exposed to extremist political content and, in turn, in a higher level of political polarization. To connect this prediction to observational data, we exploit Facebook's "Meaningful Social Interaction" (MSI) update implemented in January 2018, which—with the goal of increasing platform engagement—heavily boosted the weight given to shares in Facebook's ranking algorithm.[28] In particular, our theoretical framework suggests that we should observe an increase in extremism and political polarization following such a change in the algorithm. In what follows, we provide empirical evidence in support of such predictions by leveraging different survey datasets from Italy and the US around the time of the change in Facebook's algorithm.

6.1 Evidence from Italy

We first provide evidence from Italy by exploiting a dataset coming from the *Polimetro* (i.e., Political meter) surveys run by the leading Italian public opinion polling company *Ipsos*. The *Polimetro* contains weekly/monthly interviews on a representative sample of the Italian voting population (i.e., aged 18 or above).

In particular, for the purpose of our analysis, the survey asks questions on the main sources of information used by an individual to form a political opinion (i.e., newspapers, radio news, tv news, friends,

[27] An aspect that may play a role in moderating the platform-optimal level of η is how it impacts advertising through its effect on content, besides the effect on engagement. For example, if advertisers were to dislike a too high level of misinformation/polarization, this could be reflected in W_ψ and, consequently, could affect the platform-optimal level of η. Nevertheless, for $\psi > 0$, i.e., as long as the platform cares about engagement, the level of η chosen by the platform will be higher than the one minimizing misinformation and/or polarization. For theoretical models on the role of advertising in affecting content moderation in social media, see Madio and Quinn (2024); Liu et al. (2021); Jiménez Durán (2022).

[28] The MSI update also increased the extent of personalization of rankings as it assigned a multiplier between 0.3 and 0.5 to contents/reactions from indirect links (e.g., non-friends). As discussed in Section 5.2, our theoretical framework suggests that an increase in personalized rankings further contributes to political polarization. For details on the MSI algorithmic update see Horwitz (2023) as well as https://www.documentcloud.org/documents/21093256/pages/1/?embed=1; www.edition.cnn.com/2021/10/27/tech/facebook-papers-meaningful-social-interaction-news-feed-math and https://www.wsj.com/articles/facebook-algorithm-change-zuckerberg-11631654215?mod=article.

internet, etc). It is important to notice that around the time of its MSI update, Facebook was by far the first social network in Italy with 34 million active users per month and a 60% penetration rate in the overall population (corresponding to a penetration rate of almost 80% with respect to the population of Italian internet users), compared with the 33% and 23% penetration rates of Instagram and Twitter, respectively (We Are Social, 2018). Hence, while the Ipsos survey does not directly ask questions about Facebook use, it is possible to proxy the exposure to Facebook content with the use of internet to form a political opinion.

Furthermore, besides providing information on the socio-demographic characteristics of the respondents, the *Polimetro* asks questions regarding the ideological position of the respondent on the left-right scale and on the probability of voting for each party. Accordingly, we make use of these questions to construct two main outcome variables. The first one is a dummy variable taking value zero if a respondent self-identifies with a moderate political position (center, center-left or center-right) and one if she instead identifies with a more extremist position (left, right, extreme-left, extreme-right). This variable is thus meant to capture a simple measure of political extremism. The second one is a measure of affective polarization capturing "the extent to which citizens feel sympathy towards partisan in-groups and antagonism towards partisan out-groups" (Wagner 2021, page 1), see Appendix B.4.1 for further details.

6.1.1 Empirical strategy

We implement a Differences-in-Differences empirical model to assess whether the change in the Facebook algorithm implemented in January 2018 via the introduction of the "Meaningful Social Interaction" weights had an impact on self-declared ideological extremism and affective polarization. Specifically, we look at such outcomes for people who use internet to form a political opinion and who were interviewed after (January-June 2018) vs. before (i.e., June-December 2017) the MSI algorithm was introduced, and then compare it with the ones of people that were not using internet as one of the main sources to form an opinion interviewed after vs. before such change in the algorithm. Accordingly, we estimate the following econometric specification:

$$\begin{aligned}\mathtt{Y_{i,m,t}} &= \alpha + \beta_1 (\mathtt{Opinion\ via\ internet}_{i,m,t} \times \mathtt{Post\ MSI}) \\ &\quad + \beta_2\ \mathtt{Opinion\ via\ internet}_{i,m,t} + \beta_3\ \mathtt{Post\ MSI} + \alpha_m + \mathtt{X_{i,t}} + \varepsilon_{i,m,t} \end{aligned} \tag{18}$$

where $Y_{i,m,t}$ represents the outcome of interest relative to individual i, leaving in municipality m interviewed in the survey wave t (i.e., probability of declaring a non-moderate political ideology or affective polarization). β_1 is the parameter of interest. α_m captures municipality fixed effects (hence, our effects estimate the impact of the algorithmic change holding constant time-invariant geographical characteristics of the place of residence). $X_{i,t}$ represents a vector of socio-demographic control variables, including the respondent's age (and age squared), gender, number of resident family members, level of education, type of occupation, religiosity, and interview format (telephone/mobile assisted or computer-assisted). We also include the interaction terms between the socio-demographic control variables and the post-MSI dummy to account for possible differential trends of individual characteristics correlated with the time of the algorithmic change. Observations are weighted according to the sampling weights provided by Ipsos, and thus, the results are representative of the Italian voting-age population. In more demanding specifications, we also include either survey-wave fixed effects or region-by-survey wave fixed effects (which account for any unobservable shock at the region-survey wave level).

6.1.2 Results

Table 1 shows our baseline results on the effect of the introduction of Facebook's MSI update on the probability that an individual using internet to form an opinion holds a non-moderate political position. Column 1 provides estimates when including municipality-fixed effects only besides individual-level controls. Column 2 provides estimates when interacting individual-level controls with a Post MSI dummy, accounting for possible differential trends of individual characteristics correlated with the time of the algorithmic change. Column 3 also includes survey-wave fixed effects accounting for possible overall time-varying patterns in ideological positions. Column 4 includes fixed effects at the region-survey wave level, thus accounting for any region-time variation in political preferences. The results suggest that in the period after the MSI implementation, individuals using internet to form a political opinion had a higher probability of holding a non-moderate ideology. The effect—in the most demanding specification—accounts for a 1/11 standard deviation increase in such probability.

Figure 4 presents an event study specification in support of the underlying parallel trend assumption. That is, it reports estimates from a specification where we augment the model in equation (19) by including up to five lags and six leads.[29] This event study specification, besides the inclusion of leads and lags, reflects the most conservative one presented in Column (4) of Table 1. That is, it includes municipality and region-by-survey wave fixed effects, and the full set of controls interacted with the Post MSI dummy. Importantly, Figure 4 shows that before Facebook's MSI update, using internet as the main source of information to form a political opinion did not have any significant impact on the probability of an individual self-identifying with a non-moderate ideological position. Instead, after the MSI update the point estimates become positive and statistically significant starting from the third month after the update. Figure 4 thus shows both (a) the absence of pre-trends (hence lending empirical support to the parallel trend assumption beyond the diff-in-diff model estimated in Equation (19); (b) that the impact of the algorithmic change on ideological extremism does not materialize on impact but rather arises over the course of the following

[29] Notice that the lag minus five is not present as the survey was not run in August 2017.

Table 1: MSI and non-moderate ideological position

	(1) Non-moderate Ideology	(2) Non-moderate Ideology	(3) Non-moderate Ideology	(4) Non-moderate Ideology
Opinion via internet × Post MSI	0.058***	0.050***	0.048***	0.046***
	(0.017)	(0.018)	(0.018)	(0.017)
Opinion via internet	-0.006	-0.001	0.001	-0.001
	(0.014)	(0.014)	(0.014)	(0.013)
Post MSI	-0.009	0.021		
	(0.013)	(0.208)		
Observations	26,558	26,558	26,558	26,558
Mean pre-MSI	0.36	0.36	0.36	0.36
SD pre-MSI	0.48	0.48	0.48	0.48
Municipality FE	YES	YES	YES	YES
Individual controls	YES	YES	YES	YES
Individual controls interacted with Post MSI	NO	YES	YES	YES
Survey-wave FE	NO	NO	YES	NO
Region-survey wave FE	NO	NO	NO	YES

Note: Time horizon: June 2017-June 2018. All estimates include the following control variables: age, age squared, gender, number of resident family members, level of education, type of occupation, religiosity of the respondent and interview format (telephone/mobile assisted or computer-assisted). Observations are weighted according to the sampling weights provided by Ipsos and thus the results are representative of the Italian voting age population. Robust Standard Errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

months after the update.

We now turn to the analysis of affective polarization. Table 2 presents our results.[30] The results show a positive, statistically significant, and robust effect in the most demanding specifications. That is, in the period after the MSI algorithmic update, individuals using internet to form a political opinion had a higher level of affective polarization. Also, in this case, the effect is sizeable, accounting for around 0.9 of a standard deviation increase in affective polarization.[31] All in all, Tables 1 and 2 provide evidence in support of one of the key theoretical predictions of our model: increasing the weight attributed by recommender systems to "highlights" leads to an increase in political polarization. Importantly, Appendix Tables B.1 and B.2 show that the results are robust to excluding observations in the pre-electoral period (January-March 2018). Indeed, if anything, when excluding the pre-electoral period the estimates are larger in magnitude.

6.2 Evidence from the US

We now provide evidence from the US by exploiting a dataset coming from the *American Trends Panel* of the Pew Research Center. The Pew surveys are interviews on a representative sample of the US voting

[30] The lower number of observations relative to Table 1 is due to the fact that the questions used as proxies of sympathy score for the different parties are asked less frequently (i.e., in fewer surveys) with respect to the one on the self-declared ideological position.

[31] Figure B.4 in the Appendix presents corresponding event study estimates. Notice, that the measure of affective polarization hinges upon survey questions that were absent from the February and March 2018 survey waves and it was present only in a subset of the December 2017 survey waves. Accordingly, the Figure reflects only the estimates corresponding to the available data in a given month-year.

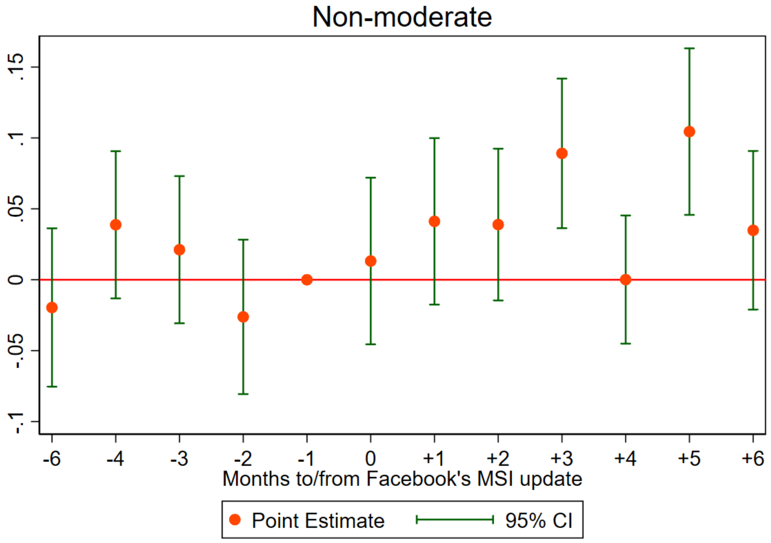

Figure 4: Event study (Italy)

Table 2: MSI and Affective Polarization

	(1) Affective Polarization	(2) Affective Polarization	(3) Affective Polarization	(4) Affective Polarization
Opinion via internet × Post MSI	0.049*	0.066**	0.067**	0.067**
	(0.030)	(0.032)	(0.031)	(0.030)
Opinion via internet	-0.011	-0.018	-0.019	-0.017
	(0.023)	(0.023)	(0.023)	(0.021)
Post MSI	0.126***	-0.070		
	(0.022)	(0.347)		
Observations	14,837	14,837	14,837	14,837
Mean pre-MSI	1.21	1.21	1.21	1.21
SD pre-MSI	0.57	0.57	0.57	0.57
Municipality FE	YES	YES	YES	YES
Individual controls	YES	YES	YES	YES
Individual controls interacted with Post MSI	NO	YES	YES	YES
Survey-wave FE	NO	NO	YES	NO
Region-survey wave FE	NO	NO	NO	YES

Note: Time horizon: June 2017-June 2018. All estimates include the following control variables: age, age squared, gender, number of resident family members, level of education, type of occupation, religiosity of the respondent and interview format (telephone/mobile assisted or computer-assisted). Observations are weighted according to the sampling weights provided by Ipsos and thus the results are representative of the Italian voting age population. Robust Standard Errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

population.[32] In particular, for the purpose of our analysis, there are survey waves where individuals are asked questions on the specific social media they use. Hence, differently from the Italian survey data, it is possible to determine the specific social media used by each individual. This allows us to have more specific treatment (Facebook users) and control (users of other social media) groups. Moreover, similarly to the Italian dataset, the survey contains a question on the ideological position of the respondent. We

[32] See `https://www.pewresearch.org/american-trends-panel-datasets/`

then construct a dummy variable taking value zero if a respondent self-identifies with a relatively moderate political position (liberal, moderate, conservative) and one if she instead identifies with a more extremist position (very liberal, very conservative).[33] However, the frequency of questions regarding social media use is irregular. As such, we focus on survey waves containing information on specific social media use around the time when the Meaningful Social Interaction (MSI) algorithm was introduced (2016-2019).[34] We then provide a robustness analysis when restricting to a time period closer to the change in the algorithm.

6.2.1 Empirical strategy

We implement a Differences-in-Differences empirical model to assess whether the change in the Facebook algorithm implemented in January 2018 via the introduction of the "Meaningful Social Interaction" weights had a causal impact on self-declared ideological extremism. Specifically, We compare the responses of Facebook users before the Meaningful Social Interaction (MSI) algorithm was introduced and then compare it with the ones of Facebook users before the MSI update and, at the same time, compare with the responses of users of other social media platforms before vs. after the MSI update. Accordingly, we estimate the following econometric specification:

$$\begin{aligned} \mathtt{Y_{i,t}} &= \alpha + \beta_1 (\mathtt{Facebook\ User}_{i,m,t} \times \mathtt{Post\ MSI}) \\ &\quad + \beta_2\ \mathtt{Facebook\ User}_{i,m,t} + \beta_3\ \mathtt{Post\ MSI} + \mathtt{X_{i,t}} + \varepsilon_{i,t} \end{aligned} \tag{19}$$

where $Y_{i,t}$ represents the outcome of interest relative to individual i, interviewed in the survey wave t (i.e., probability of declaring a non-moderate political ideology). β_1 is the parameter of interest. $X_{i,t}$ represents a vector of socio-demographic control variables available in the Pew surveys, namely marital status, income segment, age category, gender, education level, ethnicity, religion, and attendance to religious services. We also include the interaction terms between the socio-demographic control variables and the post-MSI dummy to account for possible differential trends of individual characteristics correlated with the time of the algorithmic change. In more demanding specifications, we also include either survey-wave fixed effects (i.e., time FE) or region-survey wave fixed effects (which account for any unobservable shock at the region-time level). Observations are weighted according to the sampling weights provided by the Pew Research Center, and thus, the results are representative of the US voting population.

[33] The Pew survey waves investigating the use of specific social media (around the time of the MSI update) do not contain information on the probability of voting for each party across different surveys. Hence, it is not possible to construct a measure of affective polarization. It is also worth noticing that the Pew dataset contains much less granular geographical information—with respect to the Italian *Polimetro* Dataset—as it reports only the region where the respondent lives (Northeast, Midwest, South. West).

[34] The surveys containing information on specific social media use before the MSI update encompass Wave 14 in January 2016, Wave 19 in July 2016 and Wave 28 in August 2017. The surveys after the MSI update containing information on specific social media use are Wave 35 in May 2018, Wave 37 in July 2018, and Wave 51 in July 2019.

6.2.2 Results

The estimates point out a positive impact of the Facebook MSI update on the probability of Facebook users self-identifying as non-moderate, in the order of around 1/6 of a standard deviation. As the PEW surveys containing information on social media use have an irregular frequency, providing evidence on the parallel trend assumption behind the diff-in-diff specification is not straightforward. Nevertheless, Figure B.5 in the Appendix presents an event study suggesting that no significant pre-trends were present prior to the 2018 MSI update. Moreover, as a robustness, Appendix Table B.3 presents estimates when restricting the time period in an interval close to the change in Facebook algorithm. Specifically, we focus on two survey waves before and two survey waves after January 2018 (over the period July 2016-July 2018).[35] In this case, the magnitude of the effect is even larger (around 1/5 of a standard deviation) with respect to our baseline specification.

Table 3: MSI and non-moderate ideological position - Evidence from the US

	(1) Non-moderate Ideology	(2) Non-moderate Ideology	(3) Non-moderate Ideology	(4) Non-moderate Ideology
Facebook User × Post MSI	0.0511**	0.0436+	0.0729**	0.0724**
	(0.026)	(0.027)	(0.034)	(0.034)
Facebook User	-0.0222+	-0.0203	-0.0507**	-0.0503**
	(0.015)	(0.015)	(0.026)	(0.026)
Post MSI	0.0007	-0.0625		
	(0.015)	(0.065)		
Observations	11,234	11,234	11,234	11,234
Mean pre-MSI	0.19	0.19	0.19	0.19
SD pre-MSI	0.39	0.39	0.39	0.39
Individual controls	YES	YES	YES	YES
Individual controls interacted with Post MSI	NO	YES	YES	YES
Survey wave FE	NO	NO	YES	NO
Region-survey Wave FE	NO	NO	NO	YES

Note: Time horizon: January 2016, July 2016, August 2017, May 2018, July 2018, July 2019. All estimates include the following control variables: marital status, income segment, age category, gender, education level, ethnicity, religion, attendance to religious services, and language of the interview. Observations are weighted according to the sampling weights provided by the Pew American Trend Panel, and thus, the results are representative of the US voting-age population. Robust Standard Errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1, + p<0.15

7 Conclusion

This paper develops a model that describes how social media recommendation algorithms impact individuals' choices about news content. In the theoretical framework, individuals seek information about a continuous state of the world and choose among news items recommended by the platform. The algorithm uses different types of engagement measures (clicking and highlighting in the model) to determine the

[35] The surveys containing information on specific social media use before the MSI update encompass Wave 19 in July 2016 and Wave 28 in August 2017. The surveys after the MSI update containing information on specific social media use are Wave 35 in May 2018 and Wave 37 in July 2018.

ranking and hence visibility of the content. Individuals exhibit empirically grounded behavioral traits, including confirmation bias and attention bias in favor of higher ranked objects (when clicking), and a propensity of engaging that is higher for more extreme types (when highlighting). Our analysis formalizes a central trade-off: Increasing the algorithmic weight on highlights increases user engagement, but it also amplifies political polarization and misinformation, implying a mechanism of *crowding out the truth.* It further shows that a distortion in terms of how individuals perceive each other on the platform may also arise (Bail, 2021; Yang et al., 2016). To show these points, we characterize the dynamic feedback loop between clicking, highlighting, and ranking. We show that the platform-optimal choice of algorithmic weight on highlights is generally too high, as it over-prioritizes engagement at the expense of information quality. We propose a simple policy instrument—a per-unit tax on highlights—that can correct the misalignment. By introducing a marginal cost on those types of highlights that reinforce polarization and misinformation, the engagement tax induces platforms to internalize the social costs of their algorithmic parameter choices. We also extend the model to include personalization in rankings and show that, while it increases engagement, it further exacerbates polarization with limited effect on misinformation.

Although the model is intentionally stylized to allow for analytical tractability and clear-cut comparative statics, the core theoretical insights are not mechanical consequences of the assumptions. The result that behavioral traits such as confirmation bias and asymmetric highlighting propensities lead to systematic crowding out of the truth and amplification of polarization arises from the dynamic feedback between user engagement and algorithmic ranking. Crucially, this effect emerges even though the algorithm does not explicitly target ideological content or personalize the ranking of content (in the baseline model). The distortions stem from the interaction between behavioral heterogeneity—in both prior beliefs and engagement behavior—and the platform's algorithm design. As such, the results do not follow directly from any single assumption, but rather from their joint effect in the limit under endogenous ranking. Taken together, the findings illustrate how relatively innocuous and empirically grounded micro-level behaviors, when combined with engagement-driven algorithms, can lead to wide-spread, persistent and socially consequential aggregate distortions.

In terms of empirical evidence, we exploit survey data from Italy and the United States to show that Facebook's 2018 "Meaningful Social Interactions" algorithmic update—effectively an increase in algorithmic weight on highlights—was associated with significant increases in ideological extremism and affective polarization. These findings lend support to the model's predictions and underscore the broader societal implications of engagement-driven design. Importantly, the insights of the model are very much consistent with the recent empirical literature looking at the link between social media, misinformation and polarization (Levy, 2021; González-Bailón et al., 2023; Guess et al., 2023a,b; Braghieri et al., 2025).

Overall, the results suggest that digital platform design entails a fundamental tension between maximizing engagement and preserving a healthy information environment. As recommendation systems continue to shape political discourse and civic life, aligning platform incentives with societal welfare remains a first-order policy challenge.

We conclude by acknowledging that the model does not embed important features of social media such as endogenous networks or fact-checking. Complementary research (Acemoglu et al., 2024, 2025) points out that these additional features may further reinforce the trade-off between platform engagement and social welfare outlined here. All in all, the insights from this line of research provide a "theory of harm". As such, it challenges claims by social media representatives on the neutrality of algorithms (Clegg, 2021) and, vice versa, it indirectly endorses the recent attempt by the European Union to regulate digital platforms.[36] At the same time, our results suggest that there is no need for cumbersome public policy interventions such as a ban on highlights, dictating algorithmic structure, or directly measuring the misinformation contained in social media platforms. A simple "engagement tax" on highlights can reduce the harm created by platforms when maximizing engagement by inducing them to internalize the costs it imposes on users.

Future research combining endogenous dynamic algorithmic ranking and endogenous belief and network formation may provide additional insights to guide public regulators and social media platforms in their efforts to reduce the negative impact of ranking dynamics on social media users and on democratic society at large.

[36] See `https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package`.

References

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius (2024) "A model of online misinformation," *Review of Economic Studies*, 91 (6), 3117–50.

——— (2025) "AI and Social Media: A Political Economy Perspective," Working Paper 33892, National Bureau of Economic Research, 10.3386/w33892.

Allcott, Hunt, Matthew Gentzkow, and Lena Song (2022) "Digital Addiction," *American Economic Review*, 112 (7), 2424–63, 10.1257/aer.20210867.

Amnesty International (2022) "The Social Atrocity. Meta and the Right to Remedy for the Rohingya," `www.amnesty.org/en/documents/ASA16/5933/2022/en/`, [Online; accessed 01-October-2022].

An, Jisun, Daniele Quercia, and Jon Crowcroft (2014) "Partisan Sharing: Facebook Evidence and Societal Consequences," in *Proceedings of the Second ACM Conference on Online Social Networks*, COSN '14, 13–24, New York, NY, USA: Association for Computing Machinery, 10.1145/2660460.2660469.

Analytis, Pantelis P., Francesco Cerigioni, Alexandros Gelastopoulos, and Hrvoje Stojic (2022) "Sequential choice and selfreinforcing rankings," Economics Working Papers 1819, Department of Economics and Business, Universitat Pompeu Fabra, `https://ideas.repec.org/p/upf/upfgen/1819.html`.

Aridor, Guy, Rafael Jiménez-Durán, Ro'ee Levy, and Lena Song (2024) "The Economics of Social Media," *Journal of Economic Literature*, 62 (4), 1422–74, 10.1257/jel.20241743.

Athey, Susan, Emilio Calvano, and Joshua S Gans (2018) "The impact of consumer multi-homing on advertising markets and media competition," *Management science*, 64 (4), 1574–1590.

Azzimonti, Marina and Marcos Fernandes (2022) "Social media networks, fake news, and polarization," *European Journal of Political Economy*, 102256.

Bail, Chris (2021) "Breaking the social media prism," in *Breaking the Social Media Prism*: Princeton University Press.

Bakshy, Eytan, Solomon Messing, and Lada A Adamic (2015) "Exposure to ideologically diverse news and opinion on Facebook," *Science*, 348 (6239), 1130–1132.

Bartels, Larry M (2016) "Failure to converge: Presidential candidates, core partisans, and the missing middle in American electoral politics," *The ANNALS of the American Academy of Political and Social Science*, 667 (1), 143–165.

Beknazar-Yuzbashev, George, Rafael Jiménez Durán, Jesse McCrosky, and Mateusz Stalinski (2022) "Toxic Content and User Engagement on Social Media: Evidence from a Field Experiment," *CESifo Working Paper.*

Beknazar-Yuzbashev, George, Rafael Jiménez-Durán, and Mateusz Stalinski (2024) "A Model of Harmful Yet Engaging Content on Social Media," *AEA Papers and Proceedings*, 114, 678–83, 10.1257/pandp.20241004.

Bernhardt, Dan, Stefan Krasa, and Mattias Polborn (2008) "Political polarization and the electoral effects of media bias," *Journal of Public Economics*, 92 (5-6), 1092–1104.

Braghieri, Luca, Sarah Eichmeyer, Ro'ee Levy, Markus M Mobius, Jacob Steinhardt, and Ruiqi Zhong (2025) "Article-Level Slant and Polarization of News Consumption on Social Media," *Working Paper.*

Bursztyn, Leonardo, Georgy Egorov, Ruben Enikolopov, and Maria Petrova (2019) "Social media and xenophobia: evidence from Russia,"Technical report, National Bureau of Economic Research.

Cagé, Julia, Nicolas Hervé, and Béatrice Mazoyer (2022) “Social media and newsroom production decisions.”

Chavalarias, David, Paul Bouchaud, and Maziyar Panahi (2023) “Can Few Lines of Code Change Society? Beyond fack-checking and moderation : how recommender systems toxifies social networking sites.”

Clegg, Nick (2021) “You and the Algorithm: It Takes Two to Tango,” https://nickclegg.medium.com/you-and-the-algorithm-it-takes-two-to-tango-7722b19aa1c2, [Online; accessed 01-April-2024].

CNN (2021) “Likes, anger emojis and RSVPs: the math behind Facebook’s News Feed — and how it backfired,” www.edition.cnn.com/2021/10/27/tech/facebook-papers-meaningful-social-interaction-news-feed-math, [Online; accessed 01-July-2022].

Di Tella, Rafael, Ramiro Gálvez, and Ernesto Schargrodsky (2021) “Does Social Media Cause Polarization? Evidence from Access to Twitter Echo Chambers During the 2019 Argentine Presidential Debate,” *NBER Working Paper* (w29458).

Epstein, Robert and Ronald E Robertson (2015) “The Search Engine Manipulation Effect (SEME) and its Possible Impact on the Outcomes of Elections,” *Proceedings of the National Academy of Sciences*, 112 (33), E4512—-E4521.

EU DisinfoLab (2022) “EU DisinfoLab’s Position on the 2022 Code of Practice on Disinformation,” Policy Statement, June, Available at https://www.disinfo.eu/advocacy/eu-disinfolabs-position-on-the-2022-code-of-practice-on-disinformation/.

Flaxman, Seth, Sharad Goel, and Justin M Rao (2016) “Filter bubbles, echo chambers, and online news consumption,” *Public opinion quarterly*, 80 (S1), 298–320.

Fraxanet, Emma, Andreas Kaltenbrunner, Fabrizio Germano, and Vicenç Gómez (2025) “Analyzing news engagement on Facebook: Tracking ideological segregation and news quality in the Facebook URL dataset,” *arXiv:2409.13461v2*.

Fujiwara, Thomas, Karsten Müller, and Carlo Schwarz (2024) “The effect of social media on elections: Evidence from the united states,” *Journal of the European Economic Association*, 22 (3), 1495–1539.

Galasso, Vincenzo and Tommaso Nannicini (2011) “Competing on good politicians,” *American political science review*, 105 (1), 79–99.

Garz, Marcel, Jil Sörensen, and Daniel F Stone (2020) “Partisan selective engagement: Evidence from Facebook,” *Journal of Economic Behavior & Organization*, 177, 91–108.

Gentzkow, Matthew and Jesse M Shapiro (2010) “What drives media slant? Evidence from US daily newspapers,” *Econometrica*, 78 (1), 35–71.

Gentzkow, Matthew, Jesse M. Shapiro, and Daniel F. Stone (2015) “Chapter 14 - Media Bias in the Marketplace: Theory,” in Anderson, Simon P., Joel Waldfogel, and David Strömberg eds. *Handbook of Media Economics*, 1 of Handbook of Media Economics, 623–645: North-Holland, https://doi.org/10.1016/B978-0-444-63685-0.00014-0.

Germano, Fabrizio, Vicenç Gómez, and Gaël Le Mens (2019) “The few-get-richer: a surprising consequence of popularity-based rankings?” in *The World Wide Web Conference*, 2764–2770.

Germano, Fabrizio and Francesco Sobbrio (2020) “Opinion dynamics via search engines (and other algorithmic gatekeepers),” *Journal of Public Economics*, 187, 104188.

van Gils, Freek, Wieland Müller, and Jens Prüfer (2024) "Microtargeting, voters' unawareness, and democracy," *The Journal of Law, Economics, and Organization*, ewae002, 10.1093/jleo/ewae002.

Glick, Mark, Greg Richards, Margarita Sapozhnikov, and Paul Seabright (2014) "How does ranking affect user choice in online search?" *Review of Industrial Organization*, 45 (2), 99–119.

González-Bailón, Sandra, David Lazer, Pablo Barberá et al. (2023) "Asymmetric ideological segregation in exposure to political news on Facebook," *Science*, 381 (6656), 392–398, 10.1126/science.ade7138.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer (2019) "Fake news on Twitter during the 2016 US presidential election," *Science*, 363 (6425), 374–378.

Guess, Andrew M., Neil Malhotra, Jennifer Pan et al. (2023a) "How do social media feed algorithms affect attitudes and behavior in an election campaign?" *Science*, 381 (6656), 398–404, 10.1126/science.abp9364.

——— (2023b) "Reshares on social media amplify political news but do not detectably affect beliefs or opinions," *Science*, 381 (6656), 404–408, 10.1126/science.add8424.

Hatte, Sophie, Etienne Madinier, and Ekaterina Zhuravskaya (2021) "Reading Twitter in the Newsroom: How Social Media Affects Traditional-Media Reporting of Conflicts."

Hopp, Toby, Patrick Ferrucci, and Chris J Vargo (2020) "Why do people share ideologically extreme, false, and misleading content on social media? A self-report and trace data–based analysis of countermedia content dissemination on Facebook and Twitter," *Human Communication Research*, 46 (4), 357–384.

Horwitz, Jeff (2023) *Broken Code: Inside Facebook and the Fight to Expose its Harmful Secrets*: Doubleday.

Hsu, Chin Chi, Amir Ajorlou, and Ali Jadbabaie (2024) "A game-theoretic model of misinformation spread on social networks," *Available at SSRN 4808800*.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood (2019) "The origins and consequences of affective polarization in the United States," *Annual Review of Political Science*, 22 (1), 129–146.

Jiménez Durán, Rafael (2022) "The economics of content moderation: Theory and experimental evidence from hate speech on Twitter," *Available at SSRN*.

Kalra, Aarushi (2021) "Hate in the time of algorithms: Evidence from a large-scale experiment on online behavior."

Konovalova, Elizaveta, Gaël Le Mens, and Nikolas Schöll (2023) "Social media feedback and extreme opinion expression," *Plos one*, 18 (11), e0293805.

Krasa, Stefan and Mattias K Polborn (2009) "Is mandatory voting better than voluntary voting?" *Games and Economic Behavior*, 66 (1), 275–291.

Krishna, Vijay and John Morgan (2011) "Overcoming ideological bias in elections," *Journal of Political Economy*, 119 (2), 183–211.

Levy, Ro'ee (2021) "Social media, news consumption, and polarization: Evidence from a field experiment," *American economic review*, 111 (3), 831–70.

Liu, Yi, Pinar Yildirim, and Z. John Zhang (2021) "Social Media, Content Moderation, and Technology," 10.48550/ARXIV.2101.04618.

Madio, Leonardo and Martin Quinn (2024) "Content moderation and advertising in social media platforms," *Journal of Economics & Management Strategy.*

Matias, J Nathan (2023) "Humans and algorithms work together—so study them together," *Nature*, 617 (7960), 248–251.

Matias, J Nathan and Lucas Wright (2022) "Impact Assessment of Human-Algorithm Feedback Loops," *Social Science Research Council.*

Mosleh, Mohsen, Gordon Pennycook, and David G Rand (2020) "Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter," *Plos one*, 15 (2), e0228882.

Mullainathan, Sendhil and Andrei Shleifer (2005) "The market for news," *American economic review*, 95 (4), 1031–1053.

Müller, Karsten and Carlo Schwarz (2021) "Fanning the flames of hate: Social media and hate crime," *Journal of the European Economic Association*, 19 (4), 2131–2167.

——— (2023) "From Hashtag to Hate Crime: Twitter and Antiminority Sentiment," *American Economic Journal: Applied Economics*, 15 (3), 270–312, 10.1257/app.20210211.

Narayanan, Arvind (2023) "Understanding Social Media Recommendation Algorithms."

Novarese, Marco and Chris Wilson (2013) "Being in the Right Place: A Natural Field Experiment on List Position and Consumer Choice," *Working Paper.*

Ortoleva, Pietro and Erik Snowberg (2015) "Overconfidence in political behavior," *American Economic Review*, 105 (2), 504–35.

Pan, Bing, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka (2007) "In Google We Trust: Users' Decisions on Rank, Position, and Relevance," *Journal of Computer-Mediated Communication*, 12, 801–823.

Pew (2019) "National Politics on Twitter: Small Share of U.S. Adults Produce Majority of Tweet,"Technical report, Pew Research Center.

Pogorelskiy, Kirill and Matthew Shum (2019) "News we like to share: How news sharing on social networks influences voting outcomes," *Available at SSRN 2972231.*

Rudiger, Jesper (2013) "Cross-Checking the Media," MPRA Paper 51786, University Library of Munich, Germany.

Sobbrio, Francesco (2014) "Citizen-editors' endogenous information acquisition and news accuracy," *Journal of Public Economics*, 113, 43–53.

Törnberg, Petter, Diliara Valeeva, Justus Uitermark, and Christopher Bail (2023) "Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms."

UNESCO (2023) "Social media: UNESCO leads global dialogue to improve the reliability of information," Press Release, February, Available at `https://www.unesco.org/en/articles/social-media-unesco-leads-global-dialogue-improve-reliability-information`.

United Nations (2023) "Policy Brief 8: Information Integrity on Digital Platforms," Our Common Agenda – UN Secretary-General, June, Available at `https://indonesia.un.org/sites/default/files/2023-06/Our%20Common%20Agenda%20Policy%20Brief%20Information%20Integrity%20(EN).pdf`.

Vosoughi, Soroush, Deb Roy, and Sinan Aral (2018) “The spread of true and false news online,” *Science*, 359 (6380), 1146–1151.

We Are Social (2018) “Digital in 2018 Report,” `https://wearesocial.com/it/blog/2018/01/global-digital- report-2018/` , [Online; accessed 01-October-2022].

White, Ryen W and Eric Horvitz (2015) “Belief Dynamics and Biases in Web Search,” *ACM Transactions on Information Systems (TOIS)*, 33 (4), 18.

Yang, JungHwan, Hernando Rojas, Magdalena Wojcieszak et al. (2016) “Why are “others” so polarized? Perceived political polarization and media use in 10 countries,” *Journal of Computer-Mediated Communication*, 21 (5), 349–367.

Yom-Tov, Elad, Susan Dumais, and Qi Guo (2013) “Promoting Civil Discourse Through Search Engine Diversity,” *Social Science Computer Review*, 1–10.

Appendix A. Proofs

Proof of Proposition 1. From Eqs. (2) and (5), we have that clicks on item m get updated according to:

$$\widehat{\kappa}_{n,m} - \widehat{\kappa}_{n-1,m} = \frac{\beta^{M-r_{n,m}} \varphi_{n,m}}{\sum_{m' \in M} \beta^{M-r_{n,m'} \varphi_{n,m'}}}.$$

By symmetry of the clicking types around $\widehat{\theta}$, the expectations of the ranking free propensities to click satisfy $\mathbf{E}[\varphi_{n,m}] \approx 1/(2[m])$, so that for the expected changes in clicks we can write:

$$\begin{aligned}
\mathbf{E}[\widehat{\kappa}_{n,m} - \widehat{\kappa}_{n-1,m}] &= \begin{cases} \frac{\beta^{M-r_{n,m}}/(2M_+)}{\sum_{m' \in M_+} \beta^{M-r_{n,m}}/(2M_+) + \sum_{m' \in M_-} \beta^{M-r_{n,m}}/(2M_-)} & \text{if } m \in M_+ \\[2ex] \frac{\beta^{M-r_{n,m}}/(2M_-)}{\sum_{m' \in M_-} \beta^{M-r_{n,m}}/(2M_+) + \sum_{m' \in M_+} \beta^{M-r_{n,m}}/(2M_-)} & \text{if } m \in M_- \end{cases} \\
&\overset{(1)}{\approx} \frac{\beta^{M-r_{n,m}}}{\sum_{m' \in M_+} \beta^{M-r_{n,m'}} + \sum_{m' \in M_-} \beta^{M-r_{n,m'}}} \qquad \text{for all } m \in M \\
&= \frac{\beta^{M-r_{n,m}}}{\sum_{m' \in M} \beta^{M-m'}} = \frac{\beta^M \beta^{\frac{M-r_{n,m}}{M}}}{\beta^M \sum_{m' \in M} \beta^{\frac{M-m'}{M}}} = \frac{\beta^{\frac{M-r_{n,m}}{M}}}{\sum_{m' \in M} \beta^{\frac{M-m'}{M}}},
\end{aligned}$$

where (1) follows from $\mathbb{P}(\mathrm{sgn}(x_n) = \mathrm{sgn}(y_m)) \approx \mathbb{P}(\mathrm{sgn}(x_n) \neq \mathrm{sgn}(y_m)) \approx 1/2$ and $[m] = M_+$ or $[m] = M_-$, where again by symmetry of the distribution of the items' signals g around $\widehat{\theta}$, $\mathbf{E}[M_+] \approx \mathbf{E}[M_-] \approx M/2$.[1]

From this we can write the probability of an item with signal $y_m = y$ being clicked by an average individual in an average run out of T runs (for $T \to \infty$), where such an item is present among the M items, as:

$$\mathbf{E}[\hat{\kappa}(y)] \;=\; \mathbf{E}\left[\frac{\beta^{\frac{M-r(y)}{M}}}{\sum_{m' \in M} \beta^{\frac{M-m'}{M}}}\right] \overset{(1)}{\approx} \frac{1 + (\beta - 1)\frac{M - \mathbf{E}[r(y)]}{M}}{1 + (\beta - 1)\sum_{m' \in M} \frac{M-m'}{M}} \overset{(2)}{\approx} \frac{1 + (\beta - 1)\frac{M - \zeta_0 + \zeta_1 \cdot \pi(y)}{M}}{1 + (\beta - 1)\sum_{m' \in M} \frac{M-m'}{M}},$$

where (1) follows from taking the binomial approximation, since $\frac{M-r(y)}{M} < 1$ and $1 < \beta < 2$, and (2) follows from applying Eq. (9) to the expected rank $r(y)$. Taking into account the distribution of the items' signals g, this readily implies as the limit clicking distribution:[2]

$$LCD(y) = \Lambda_\beta(\pi(y)) \cdot g(y), \tag{A.1}$$

where, for $z \geq 0$:

$$\Lambda_\beta(z) = \frac{M + (\beta - 1)(M - \zeta_0 + \zeta_1 \cdot z)}{M + (\beta - 1)\sum_{m' \in M}(M - m')} = \frac{M + (\beta - 1)(M - \zeta_0 + \zeta_1 \cdot z)}{M\,(1 + (\beta - 1)(M - 1)/2)}, \tag{A.2}$$

so that $\Lambda'_\beta(z)$ is a constant and $\Lambda'_\beta(z) \equiv \Lambda'_\beta = \frac{\zeta_1 \cdot (\beta - 1)}{M(1 + (\beta - 1)(M-1)/2)} > 0$ since $\beta > 1, \zeta_1 > 0$.

With this we can write the limit highlighting distribution as:

$$LHD(y) = \mu_H(y) \cdot LCD(y), \tag{A.3}$$

[1] To cut on notation, throughout the proofs, we write $f(x)$ for the density of the individuals signals ($f(x; \sigma_x^2)$ in the main text) and $g(y)$ for the density of the items' signals ($f(y; \sigma_y^2)$ in the main text).

[2] More formally, this can be derived by dividing the space around $\widehat{\theta}$ into an arbitrarily large number, say $2K$, of bins of equal size, say $\nu > 0$, K on each side of $\widehat{\theta}$. Then, for $K \to \infty$, $\nu \to 0$, the expected probability of a click on an item in a given bin, assuming there is an item with a value y in the given bin, is approximated by $\mathbf{E}[\hat{\kappa}(y)]$. Similarly, the probability that there is an item with a value y in the bin is approximated by $g(y)$.

which is not normalized and hence does not integrate to 1. □

Proof of Proposition 2. Set $\theta = \widehat{\theta} = 0$ and fix $\beta > 1$. To simplify notation we drop the subscript β from the function Λ_β and write just Λ. Given Eq. (9), we can apply Proposition 1 and write engagement (ENG), polarization (POL) and misinformation (MIS), respectively as:

$$\begin{aligned}
ENG &= \int (LCD(y) + LHD(y))\, dy = \int (1+\mu_H(y))\, LCD(y) dy = \int (1+\mu_H(y))\, \Lambda(\pi(y)) f(y) dy \\
MIS &= \int |y - 0|\, LCD(y) dy = \int |y|\, \Lambda(\pi(y)) f(y) dy \\
POL &= \int \left| yLCD^R(y) - yLCD^L(y) \right| dy = \int \left| y\Lambda^R(\pi(y)) - y\Lambda^L(\pi(y)) \right| f(y) dy,
\end{aligned}$$

where for $g \in \{L, R\}$, LCD^g is the limit clicking distribution of individuals from group g and can be written as:[3]

$$LCD^g(y) = \Lambda^g(\pi(y)) g(y),$$

where $\Lambda^g(\pi(y))$ is now the expected probability an item with signal $y_m = y$ will be clicked on by an individual in group g. Note that, while clicking and highlighting by a given group is heavily dependent on the sign of the signal of the item, (that is, whether $m \in M_+$ or $m \in M_-$), both groups share the same ranking which depends on the total clicking and highlighting propensities of the two groups ($\pi(y)$).

From this we can compute the effect of a change in η on the three variables. It suffices to compute:

$$\begin{aligned}
\frac{\partial ENG}{\partial \eta} &= \frac{\partial}{\partial \eta} \int (1+\mu_H(y))\, \Lambda(\pi(y)) g(y) dy \\
&= \int \frac{\partial (1+\mu_H(y))}{\partial \eta} \Lambda(\pi(y)) g(y) dy + \int (1+\mu_H(y))\, \frac{\partial \Lambda(\pi(y))}{\partial \eta} g(y) dy \\
&\overset{(1)}{=} \int (1+\mu_H(y))\, \Lambda' \frac{\partial \pi(y)}{\partial \eta} g(y) dy \\
&\overset{(2)}{=} \int (1+\mu_H(y))\, \Lambda' \frac{(M-1)(\mu_H(y) - \bar{\mu}_H)}{(M(1+\eta\bar{\mu}_H) + \eta(\mu_H(y) - \bar{\mu}_H))^2} g(y) dy \\
&\overset{(3)}{>} 0,
\end{aligned}$$

where (1) follows because $\frac{\partial \mu_H(y)}{\partial \eta} = 0$ and $\frac{\partial g(y)}{\partial \eta} = 0$, (2) follows from Eq. (8), and (3) follows since $\Lambda'(\pi(y)) = \Lambda' > 0$ and $\mu_H(y) \geq 0$ for all y and using Jensen's inequality, $\int \mu_H(y)(\mu_H(y) - \bar{\mu}_H) g(y) dy \geq 0$. This readily implies $\int \frac{(1+\mu_H(y))(\mu_H(y)-\bar{\mu}_H)}{(M(1+\bar{\mu}_H)+\eta(\mu_H(y)-\bar{\mu}_H))^2} g(y) dy > 0$, $\mu_H(y) > \bar{\mu}_H$ on a strictly positive mass of signals y. Similarly:

$$\begin{aligned}
\frac{\partial MIS}{\partial \eta} &= \frac{\partial}{\partial \eta} \int |y| \Lambda(\pi(y)) g(y) dy = \int |y| \frac{\partial \Lambda(\pi(y))}{\partial \eta} g(y) dy \\
&= \int |y| \Lambda' \frac{(M-1)(\mu_H(y) - \bar{\mu}_H)}{(M(1+\eta\bar{\mu}_H) + \eta(\mu_H(y) - \bar{\mu}_H))^2} g(y) dy \\
&\overset{(1)}{>} 0,
\end{aligned}$$

where (1) follows because $\Lambda' > 0$ and $\mu_H(y) \geq 0$, and $|y|\mu_H(y) > |y|\bar{\mu}_H$ on a large enough mass of signals y, given $\sigma_y^2 \leq \sigma_x^2$, ensuring that $\int |y|(\mu_H(y) - \bar{\mu}_H) g(y) dy > 0$.

Finally, to compute the effect on polarization, we need to keep track of clicking in the two groups. While

[3] The superscript g indicating the group of individuals is not to be confused with $g(y)$, which is the distribution of the items, which in the main text is written as $f(y; \sigma_y^2)$ and in the proofs simply as $g(y)$.

there is a unique ranking (since $\lambda = 1$), individuals in the different groups nonetheless behave differently.

$$\begin{aligned}
\frac{\partial POL}{\partial \eta} &= \frac{\partial}{\partial \eta}\left|\int y\Lambda^R(\pi(y)) - y\Lambda^L(\pi(y))g(y)dy\right| \\
&= \frac{\partial}{\partial \eta}\left|\int y\left(\Lambda^R(\pi(y)) - \Lambda^L(\pi(y))\right)g(y)dy\right| \\
&\overset{(1)}{=} \frac{\partial}{\partial \eta}\left(\int_{y\leq 0}(-y)\left(\Lambda^L(\pi(y)) - \Lambda^R(\pi(y))\right)g(y)dy + \int_{y>0} y(\Lambda^R(\pi(y)) - \Lambda^L(\pi(y)))g(y)dy\right) \\
&\overset{(2)}{=} \frac{\partial}{\partial \eta}\left(2\int_{y>0} y\left(\Lambda^R(\pi(y)) - \Lambda^L(\pi(y))\right)g(y)dy\right) \\
&= 2\int_{y>0} y\left(\frac{\partial \Lambda^R(\pi(y))}{\partial \eta} - \frac{\partial \Lambda^L(\pi(y))}{\partial \eta}\right)g(y)dy \\
&\overset{(3)}{=} 2\int_{y>0} y\left(\Lambda^R_+\pi(y)\frac{\partial \pi(y)}{\partial \eta} - \Lambda^L_+\pi(y)\frac{\partial \pi(y)}{\partial \eta}\right)g(y)dy \\
&= 2\int_{y>0} y\left(\Lambda^R_+ - \Lambda^L_+\right)\pi(y)\frac{\partial \pi(y)}{\partial \eta}g(y)dy \\
&= 2\int_{y>0} y\left(\Lambda^R_+ - \Lambda^L_+\right)(1+\eta\mu_H(y))\frac{(M-1)(\mu_H(y) - \bar{\mu}_H)}{(M(1+\eta\bar{\mu}_H) + \eta(\mu_H(y) - \bar{\mu}_H))^3}g(y)dy \\
&\overset{(4)}{>} 0,
\end{aligned}$$

where (1) follows because on $\mathbb{R}_-$ we have $\Lambda^L(\pi(y)) > \Lambda^R(\pi(y))$, and on $\mathbb{R}_+$ we have $\Lambda^R(\pi(y)) > \Lambda^L(\pi(y))$, (2) follows by symmetry of the limit clicking distribution, (3) follows since Λ^R, Λ^L are linear and hence, for y on $\mathbb{R}_+$, $\Lambda^{R\prime}(y) \equiv \Lambda^R_+$ and $\Lambda^{L\prime}(y) \equiv \Lambda^L_+$ are positive constants with $\Lambda^R_+ > \Lambda^L_+$, and finally (4) follows for the same reasons as with the previous case of ENG using Jensen's inequality and since and $\Lambda^R_+ > \Lambda^L_+$ on $\mathbb{R}_+$, ensuring $\int_{y>0}(1+\eta\mu_H(y))(\mu_H(y) - \bar{\mu}_H)g(y)dy > 0$. Also, note that $M(1+\bar{\mu}_H) + \eta(\mu_H(y) - \bar{\mu}_H)$ $= M + \eta\mu_H(y) + (M-1)\eta\bar{\mu}_H > 0$. □

Proof of Proposition 4. Set again $\theta = \widehat{\theta} = 0$ and fix $\beta > 1$, and write Λ for the function Λ_β, thus dropping the subscript β. Applying Eq. (9) to the personalized algorithm Eq. (14), we obtain for the expected rank:

$$r^g(y) \approx \zeta_0 - \zeta_1 \cdot \frac{\pi^g_g(y_m) + \lambda\pi^{\neg g}_g(y_m)}{1+\lambda}, \; g \in \{L, R\}, \tag{A.4}$$

where $\zeta_0, \zeta_1 > 0$ are constants and $\neg g$ denotes the group in $\{L, R\}$ other than g. Here the expressions $\pi^g_g(y)$ and $\pi^{\neg g}_g(y)$ denote respectively the popularity from individuals in g and in $\neg g$ in the ranking seen by group g:[4]

$$\pi^g_g(y) = \frac{1 + \eta \cdot \mu^g_H(y)}{M(1+\eta\cdot\bar{\mu}^g_H) + \eta(\mu^g_H(y) - \bar{\mu}^g_H) + \lambda\left(M(1+\eta\cdot\bar{\mu}^{\neg g}_H) + \eta(\mu^{\neg g}_H(y) - \bar{\mu}^{\neg g}_H)\right)}$$

and

$$\pi^{\neg g}_g(y) = \frac{1 + \eta \cdot \mu^{\neg g}_H(y)}{M(1+\eta\cdot\bar{\mu}^g_H) + \eta(\mu^g_H(y) - \bar{\mu}^g_H) + \lambda\left(M(1+\eta\cdot\bar{\mu}^{\neg g}_H) + \eta(\mu^{\neg g}_H(y) - \bar{\mu}^{\neg g}_H)\right)},$$

where $\mu^g_H(y)$ is the propensity to highlight by individuals in g:

$$\mu^g_H(y) = \int_{x\in H^{-1}(y)} \mathbb{I}_{\{x\in g\}} p_A(x)f(x)dx.$$

[4] The distinction is necessary because of the normalizations that get applied to the two different rankings and that therefore change the denominators in the two cases.

We can apply Lemma 1 and write engagement as:

$$\begin{aligned} ENG &= \sum_{g=L,R} \int \left(LCD^g(y) + LHD^g(y)\right) dy = \sum_{g=L,R} \int \left(1+\mu_H^g(y)\right) LCD^g(y) dy \\ &= \sum_{g=L,R} \int \left(1+\mu_H^g(y)\right) \Lambda^g \left(\frac{\pi_g^g(y) + \lambda \pi_g^{\neg g}(y)}{1+\lambda}\right) g(y) dy, \end{aligned}$$

where as in the proof of Proposition 2, $\Lambda^g\left(\frac{\pi_g^g(y)+\lambda\pi_g^{\neg g}(y)}{1+\lambda}\right)$ is the probability of being clicked by an individual in g:

$$\Lambda^g\left(\frac{\pi_g^g(y) + \lambda\pi_g^{\neg g}(y)}{1+\lambda}\right) = \frac{M + \log\beta \cdot \left(M - \zeta_0 + \zeta_1 \frac{\pi_g^g(y)+\lambda\pi_g^{\neg g}(y)}{1+\lambda}\right)}{M + \log\beta \cdot \sum_{m' \in M}(M - m')}, \tag{A.5}$$

We can compute

$$\begin{aligned} \frac{\partial ENG}{\partial\lambda} &= \sum_{g=L,R} \frac{\partial}{\partial\lambda} \int \left(1+\mu_H^g(y)\right) \Lambda^g\left(\frac{\pi_g^g(y)+\lambda\pi_g^{\neg g}(y)}{1+\lambda}\right) g(y)dy \\ &= \sum_{g=L,R} \int \left(1+\mu_H^g(y)\right) \frac{\partial \Lambda^g\left(\frac{\pi_g^g(y)+\lambda\pi_g^{\neg g}(y)}{1+\lambda}\right)}{\partial\lambda} g(y)dy \\ &= \sum_{g=L,R} \left(\int_{y\le 0} \left(1+\mu_H^g(y)\right) {\Lambda_-^g}' \frac{\partial\left(\frac{\pi_g^g(y)+\lambda\pi_g^{\neg g}(y)}{1+\lambda}\right)}{\partial\lambda} g(y)dy \right. \\ &\qquad\qquad \left. + \int_{y>0} \left(1+\mu_H^g(y)\right) {\Lambda_+^g}' \frac{\partial\left(\frac{\pi_g^g(y)+\lambda\pi_g^{\neg g}(y)}{1+\lambda}\right)}{\partial\lambda} g(y)dy \right) \\ &= 2\sum_{g=L,R} \int_{y>0} \left(1+\mu_H^g(y)\right) {\Lambda_+^g}' \frac{\partial\left(\frac{\pi_g^g(y)+\lambda\pi_g^{\neg g}(y)}{1+\lambda}\right)}{\partial\lambda} g(y)dy \\ &= 2\sum_{g=L,R} \int_{y>0} \frac{\left(1+\mu_H^g(y)\right) {\Lambda_+^g}'}{1+\lambda} \left(\frac{\partial\pi_g^g(y)}{\partial\lambda} + \lambda\frac{\partial\pi_g^{\neg g}(y)}{\partial\lambda} - \frac{\pi_g^g(y) - \pi_g^{\neg g}(y)}{1+\lambda}\right) g(y)dy \end{aligned}$$

Now,

$$\frac{\partial\pi_g^g(y)}{\partial\lambda} = \frac{-(1+\eta\cdot\mu_H^g(y))\left(M(1+\eta\cdot\bar{\mu}_H^{\neg g}) + \eta(\mu_H^{\neg g}(y) - \bar{\mu}_H^{\neg g})\right)}{\left(M(1+\eta\cdot\bar{\mu}_H^g) + \eta(\mu_H^g(y)-\bar{\mu}_H^g) + \lambda\left(M(1+\eta\cdot\bar{\mu}_H^{\neg g}) + \eta(\mu_H^{\neg g}(y) - \bar{\mu}_H^{\neg g})\right)\right)^2} < 0,$$

and

$$\frac{\partial\pi_g^{\neg g}(y)}{\partial\lambda} = \frac{-(1+\eta\cdot\mu_H^{\neg g}(y))\left(M(1+\eta\cdot\bar{\mu}_H^{g}) + \eta(\mu_H^{g}(y) - \bar{\mu}_H^{g})\right)}{\left(M(1+\eta\cdot\bar{\mu}_H^g) + \eta(\mu_H^g(y)-\bar{\mu}_H^g) + \lambda\left(M(1+\eta\cdot\bar{\mu}_H^{\neg g}) + \eta(\mu_H^{\neg g}(y) - \bar{\mu}_H^{\neg g})\right)\right)^2} < 0,$$

which immediately implies:

$$\sum_{g=L,R} \int_{y>0} \frac{\left(1+\mu_H^g(y)\right) {\Lambda_+^g}'}{1+\lambda} \left(\frac{\partial\pi_g^g(y)}{\partial\lambda} + \lambda\frac{\partial\pi_g^{\neg g}(y)}{\partial\lambda}\right) g(y)dy < 0,$$

since ${\Lambda_+^g}' > 0$ and $\lambda, \mu_H^g(y) \ge 0$ for $g \in \{L, R\}$.

Moreover, applying the definitions of π_g^g and $\pi_g^{\neg g}$, it is easy to see that:

$$\int_{y>0} \frac{\left(1+\mu_H^R(y)\right)\Lambda_+^{R\prime}}{1+\lambda}\frac{(-1)(\pi_R^R(y)-\pi_R^L(y))}{1+\lambda}g(y)dy<0$$

since $\pi_R^R(y)>\pi_R^L(y)$ on $\mathbb{R}_+$. Further, since $\mu_H^R(y)\geq\mu_H^L(y)$ on $\mathbb{R}_+$, it also follows that:

$$\left|\int_{y>0} \frac{\left(1+\mu_H^R(y)\right)\Lambda_+^{R\prime}}{1+\lambda}\frac{(-1)(\pi_R^R(y)-\pi_R^L(y))}{1+\lambda}g(y)dy\right|>\left|\int_{y>0} \frac{\left(1+\mu_H^L(y)\right)\Lambda_+^{L\prime}}{1+\lambda}\frac{(-1)(\pi_L^L(y)-\pi_L^R(y))}{1+\lambda}g(y)dy\right|$$

which in turn implies:

$$\sum_{g=L,R}\int_{y>0} \frac{\left(1+\mu_H^g(y)\right)\Lambda_+^{g\prime}}{1+\lambda}\frac{(-1)(\pi_g^g(y)-\pi_g^{\neg g}(y))}{1+\lambda}g(y)dy<0.$$

This shows that $\frac{\partial ENG}{\partial\lambda}<0$ so that less personalization (larger λ) decreases engagement both with flat and non-flat highlighting.

Finally,

$$POL \quad = \quad \left|\int yLCD^R(y)dy-\int yLCD^L(y)dy\right|,$$

so that, using the same reasoning as in the proof of Proposition 2, we can write:

$$\begin{aligned}\frac{\partial POL}{\partial\lambda} \quad &= \quad \frac{\partial}{\partial\lambda}\left(2\int_{y>0} y\left(\Lambda^R\left(\frac{\pi_R^R(y)+\lambda\pi_R^L(y)}{1+\lambda}\right)-\Lambda^L\left(\frac{\pi_L^L(y)+\lambda\pi_L^R(y)}{1+\lambda}\right)\right)g(y)dy\right)\\ &= \quad 2\int_{y>0} y\Lambda_+^{R\prime}\left(\frac{\partial\pi_R^R(y)}{\partial\lambda}+\lambda\frac{\partial\pi_R^L(y)}{\partial\lambda}-\frac{\pi_R^R(y)-\pi_R^L(y)}{1+\lambda}\right)g(y)dy\\ &\qquad -2\int_{y>0} y\Lambda_+^{L\prime}\left(\frac{\partial\pi_L^L(y)}{\partial\lambda}+\lambda\frac{\partial\pi_L^R(y)}{\partial\lambda}-\frac{\pi_L^L(y)-\pi_L^R(y)}{1+\lambda}\right)g(y)dy.\end{aligned}$$

Similar calculations as for ENG show that the first integral is negative and dominates in absolute value the second one, showing overall that $\frac{\partial POL}{\partial\lambda}<0$. Hence, less personalization (larger λ) decreases polarization both with flat and non-flat highlighting. □

Proof of Proposition 5. Recall from Eq. (12):

$$W_\psi(\eta,\lambda)=\psi\cdot ENG(\eta,\lambda)-(1-\psi)\cdot MIS(\eta,\lambda)\cdot POL(\eta,\lambda).$$

Hence, for $\psi=0$, we have $W_0=MIS\cdot POL$, while, for $\psi=1$, we have $W_1=ENG$. The results then follow directly from Propositions 2 and 4.

Consider the non-flat case. It follows immediately that W_0 is maximized at a smallest possible value of η, since $-MIS\cdot POL$ is maximized when $MIS\cdot POL$ is minimized and $\frac{MIS}{\partial\eta}>0$, $\frac{POL}{\partial\eta}>0$. Also, W_0 is maximized at a largest possible value of λ again since $\frac{POL}{\partial\lambda}<0$ (less personalization decreases POL) while $\frac{MIS}{\partial\lambda}\approx 0$. The contrary is true for $\psi=1$.

By contrast, by analogous argument, in the flat case, W_0 is maximized at a largest possible value of η and at a largest possible value of λ, while for $\psi=1$, W_1 is maximized at a largest possible value of η and a smallest possible value of λ. □

Appendix B Additional Figures and Results

We here present and discuss some additional figures and results that were not discussed or only very briefly mentioned in the main text of the paper.

Appendix B.1 Clicking and highlighting distribution in the case of constant p_A

As discussed in Section 5.1, Figure B.1 illustrates how in the presence of a constant probability of being an active type, polarization and misinformation both decrease when the highlight weight is higher.

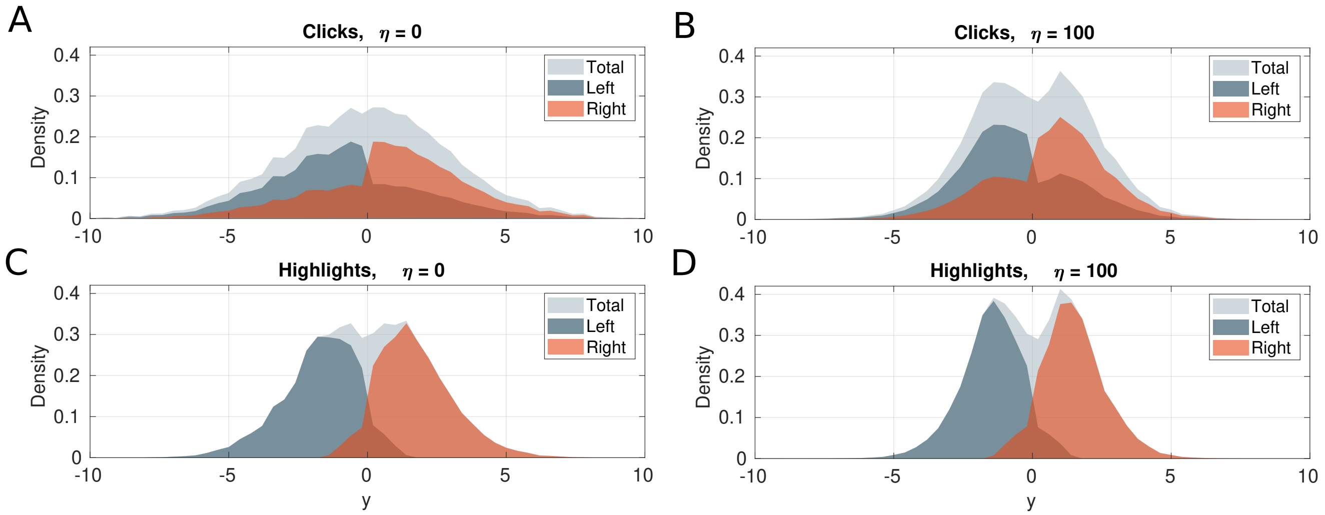

Figure B.1: Polarization and misinformation *decrease* for increasing values of η. Panels (A) and (B) show users' clicking behavior for $\eta = 0$ and $\eta = 100$, respectively. Panels (C) and (D) show users' highlighting behavior for $\eta = 0$ and $\eta = 100$, respectively. Polarization decreases from an average value of 1.8 (SD 0.5) to 1.3 (SD 0.3), and the misinformation also decreases from an average value of 2.4 (SD 0.6) to 1.7 (SD 0.3).

Appendix B.2 Heterogeneous benchmark

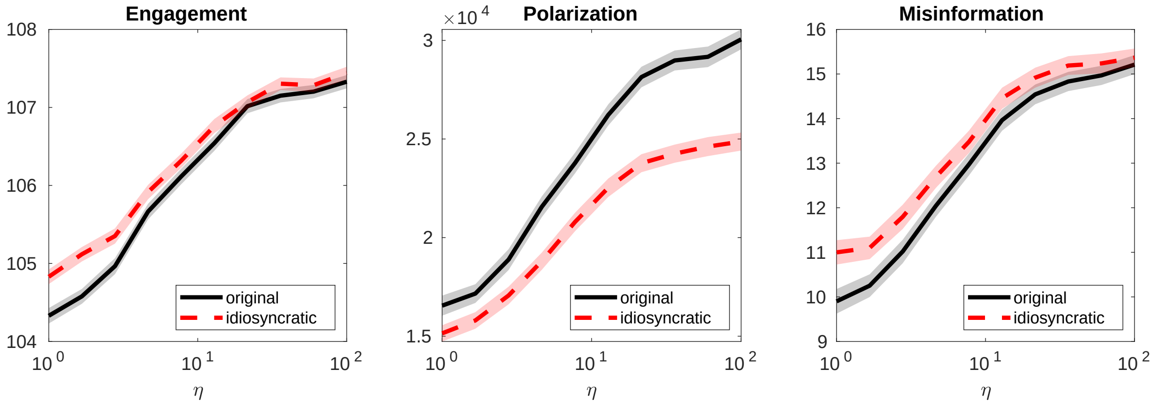

Figure B.2: Engagement, polarization, and misinformation as a function of the highlighting parameter η with a common benchmark $\widehat{\theta}$ (solid line) and heterogeneous benchmarks $\widehat{\theta}_n$ (dotted line). The shaded areas represent the 95% confidence intervals.

Consider the case where individuals can have idiosyncratic benchmarks $\widehat{\theta}_n$. The results do not change qualitatively. In fact, the simulations suggest that, for $\widehat{\theta}_n$'s centered around θ and not too dispersed ($\widehat{\theta}_n \sim N(\theta, \sigma^2_{\widehat{\theta}})$ with $\sigma_{\widehat{\theta}} \leq \min\{\frac{\sigma_x}{4}, \frac{\sigma_y}{4}\}$), our main results on engagement, popularity and misinformation are rather close to the cases where individuals have a common benchmark, $\widehat{\theta}_n = \widehat{\theta}$ for all n. This is illustrated in Figure B.2 that shows the effect of η on the variables ENG, POL, MIS.

Appendix B.3 Non-centered benchmark

While it is natural to assume that the benchmark $\widehat{\theta}$ splits the signals roughly in half in symmetric environments, so that $\widehat{\theta} \approx \theta$, it may occur occasionally that the two are far apart. In such a situation, individuals' and news items' signals are shifted away from the benchmark $\widehat{\theta}$. This means that a potentially large mass of individuals have a prior belief far from $\widehat{\theta}$ and are hence likely to highlight news items far from it but potentially close to θ. Accordingly, an increase in η can contribute to both higher engagement and at the same time lower misinformation. To see this consider Figure B.3 that illustrates a situation where clearly $\widehat{\theta} \neq \theta$. Here $x^* \approx \theta$ so that a large mass of individuals with a signal close to the truth has a large highlighting propensity. An increase in η leads to a more prominent ranking for items around $x^* \approx \theta$, which in turn, through the effect on the clicking distribution, leads to a lower level of misinformation as measured by MIS. Increasing the weight on highlights here actually accelerates individuals clicking on news items carrying truthful signals.

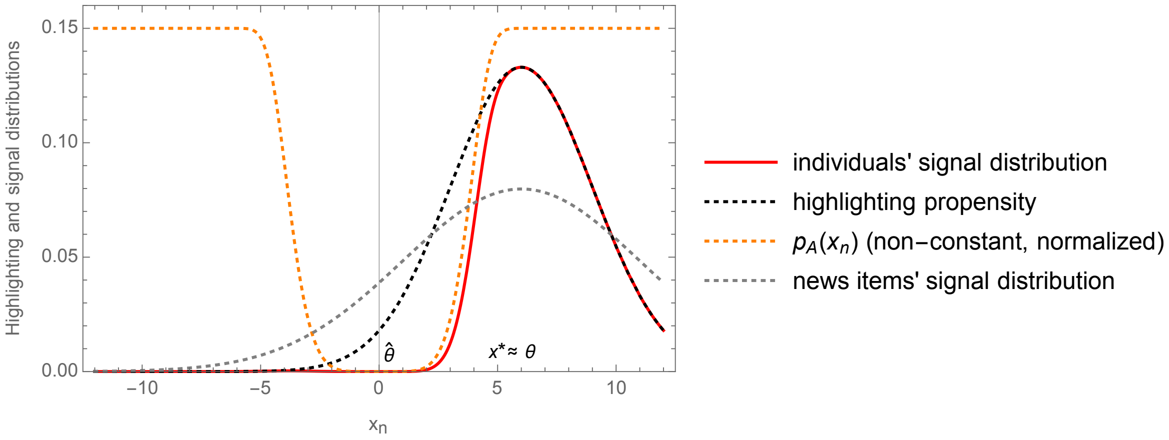

Figure B.3: Individuals' signal distribution and highlighting propensity with non-centered $\widehat{\theta}$, (with $\theta = 6$ and $\widehat{\theta} = 0$); x^* denotes the value of x_n where the highlighting propensity is locally maximal.

Appendix B.4 Empirical Evidence on Meaningful Social Interactions and Political Polarization: Additional Information

Appendix B.4.1 Affective Polarization in Italy

Since Italy is a multi-party political system, we follow Alvarez and Nagler (2004) and Wagner (2021) and define a measure of Weighted Affective Polarization (WAP) for individual i as:

$$WAP = \sqrt{\sum_{p=1}^{P} v_p * \mid symp_{ip} - \overline{symp_i} \mid}, \tag{B.1}$$

where v_p is the vote share of party p (measured as a proportion ranging from 0 to 1), $symp_{ip}$ is measured with the probability attached by individual i to voting for party p (ranging from 0 to 10), and $\overline{symp_i}$ is individual i's weighted average party sympathy score. That is:

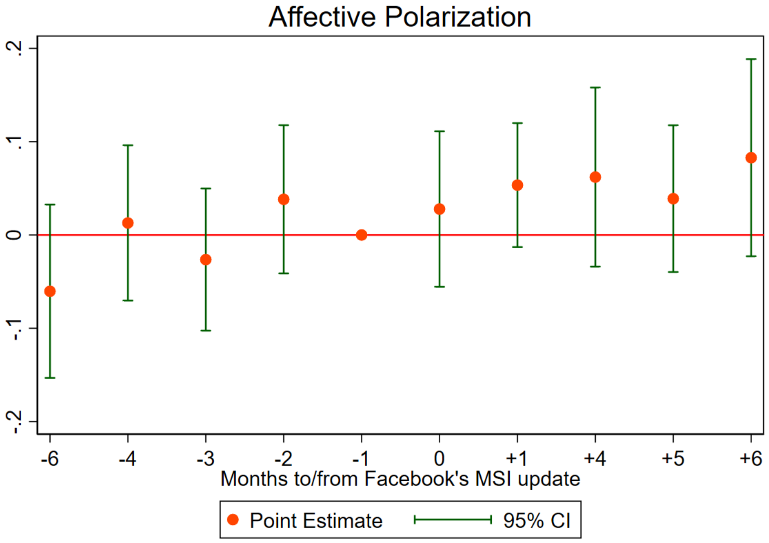

Figure B.4: Event study: Affective Polarization (Italy)

$$\overline{symp_i} = \sum_{p=1}^{P} v_p * symp_{ip} \,. \tag{B.2}$$

Appendix B.4.2 Evidence from Italy: Robustness

Figure B.4 presents event-study estimates on affective polarization providing evidence on the absence of pre-trends. Notice, that the measure of affective polarization hinges upon survey questions that were absent from the February and March 2018 survey waves and it was present only in a subset of the December 2017 survey waves. Accordingly, the Figure reflects only the estimates corresponding to the available data in a given month-year. One possible concern regarding the causal interpretation of our results linking Facebook's MSI update and political polarization in Italy is due to the concurring general elections in Italy in March 2018. With respect to this issue we notice that, by including survey-wave fixed-effect (or region-survey wave fixed effects), our empirical strategy takes into account and controls for any general trend in political polarization over time and across regions. At the same time, one might argue that the presence of elections might have led to a differential trend in political polarization between individuals who used internet to form an opinion and the ones who did not which was not due to the MSI algorithm *per se* (e.g., increase in online fake news before elections). In response to this argument, we first point out that the MSI algorithm might have further amplified the diffusion of fake-news as predicted by our model. Second, we provide below evidence suggesting a polarization effect even when dropping the months immediately after the MSI update and before the elections (i.e., January-March 2018). Specifically, Tables B.1 and B.2 present results when comparing the period June-December 2017 (pre-MSI) with April-December 2018 (post-MSI and post-elections).

Appendix B.4.3 Evidence from US: Event Study and Robustness

Figure B.5 presents the results of an event study specification with one lags and two leads (according to the most demanding specification of Table 3).

Table B.3 provides robustness estimates of the impact of the MSI update on non-moderate ideological position in the US, when restricting the time frame around the timing of the MSI update (July 2016-July 2018).

Table B.1: MSI and non-moderate ideological position: Robustness

	(1) Non-moderate Ideology	(2) Non-moderate Ideology	(3) Non-moderate Ideology	(4) Non-moderate Ideology
Opinion via internet × Post MSI	0.069***	0.054**	0.052**	0.050**
	(0.022)	(0.022)	(0.022)	(0.021)
Opinion via internet	-0.004	0.001	0.003	0.003
	(0.014)	(0.015)	(0.014)	(0.014)
Post MSI	-0.003	0.184		
	(0.016)	(0.211)		
Observations	19,820	19,820	19,820	19,820
Mean pre-MSI	0.36	0.36	0.36	0.36
SD pre-MSI	0.48	0.48	0.48	0.48
Municipality FE	YES	YES	YES	YES
Individual controls	YES	YES	YES	YES
Individual controls interacted with Post MSI	NO	YES	YES	YES
Survey-wave FE	NO	NO	YES	NO
Region-survey wave FE	NO	NO	NO	YES

Note: Time horizon: June 2017-December 2017 and April-December 2018. All estimates include the following control variables: age, age squared, gender, number of resident family members, level of education, type of occupation, religiosity of the respondent and interview format (telephone/mobile assisted or computer-assisted). Observations are weighted according to the sampling weights provided by Ipsos and thus the results are representative of the Italian voting age population. Robust Standard Errors in parenthesis.
*** p<0.01, ** p<0.05, * p<0.1

Table B.2: MSI and Affective Polarization: Robustness

	(1) Affective Polarization	(2) Affective Polarization	(3) Affective Polarization	(4) Affective Polarization
Opinion via internet × Post MSI	0.077*	0.100**	0.095**	0.080**
	(0.041)	(0.042)	(0.041)	(0.040)
Opinion via internet	-0.004	-0.010	-0.012	-0.008
	(0.024)	(0.024)	(0.024)	(0.023)
Post MSI	0.264***	0.303		
	(0.029)	(0.418)		
Observations	10,802	10,802	10,802	10,802
Mean pre-MSI	1.21	1.21	1.21	1.21
SD pre-MSI	0.56	0.56	0.56	0.56
Municipality FE	YES	YES	YES	YES
Individual controls	YES	YES	YES	YES
Individual controls interacted with Post MSI	NO	YES	YES	YES
Survey-wave FE	NO	NO	YES	NO
Region-survey wave FE	NO	NO	NO	YES

Note: Time horizon: June 2017-December 2017 and April-December 2018. All estimates include the following control variables: age, age squared, gender, number of resident family members, level of education, type of occupation, religiosity of the respondent and interview type (telephone or computer). Observations are weighted according to the sampling weights provided by Ipsos and thus the results are representative of the Italian voting age population. Robust Standard Errors in parenthesis.
*** p<0.01, ** p<0.05, * p<0.1

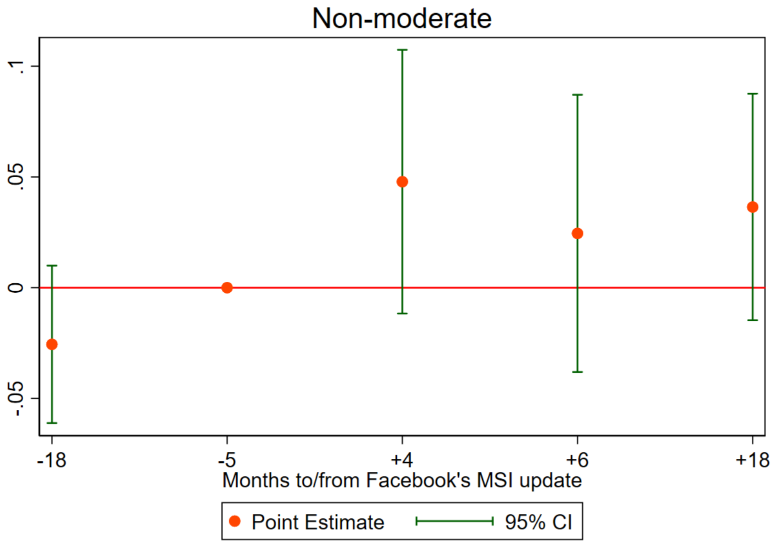

Figure B.5: Event study (US)

Table B.3: MSI and non-moderate ideological position - Restricted time frame around MSI update

	(1) Non-moderate Ideology	(2) Non-moderate Ideology	(3) Non-moderate Ideology	(4) Non-moderate Ideology
Facebook User × Post MSI	0.0890***	0.0848**	0.0842**	0.0839**
	(0.034)	(0.036)	(0.038)	(0.038)
Facebook User	-0.0596**	-0.0561**	-0.0554**	-0.0547**
	(0.023)	(0.024)	(0.026)	(0.026)
Post MSI	-0.0322	-0.1464*		
	(0.024)	(0.085)		
Observations	6,738	6,738	6,738	6,738
Mean pre-MSI	0.19	0.19	0.19	0.19
SD pre-MSI	0.39	0.39	0.39	0.39
Individual controls	YES	YES	YES	YES
Individual controls interacted with Post MSI	NO	YES	YES	YES
Survey wave FE	NO	NO	YES	NO
Region-survey Wave FE	NO	NO	NO	YES

Note: Time horizon: July 2016, August 2017, May 2018, July 2018. All estimates include the following control variables: marital status, income segment, age category, gender, education level, ethnicity, religion, attendance to religious services, and language of the interview. Observations are weighted according to the sampling weights provided by the Pew American Trend Panel, and thus, the results are representative of the US voting-age population. Robust Standard Errors in parenthesis. *** p<0.01, ** p<0.05, * p<0.1, + p<0.15