The Economics of Social Media

Guy Aridor, Rafael Jiménez-Durán, Ro'ee Levy, and Lena Song*

We provide a guide to the burgeoning literature on the economics of social media. We first define social media platforms and highlight their unique features. We then synthesize the main lessons from the empirical economics literature and organize them around the three stages of the life cycle of content: (1) production, (2) distribution, and (3) consumption. Under production, we discuss how incentives affect content produced on and off social media and how harmful content is moderated. Under distribution, we discuss the social network structure, algorithms, and targeted advertisements. Under consumption, we discuss how social media affects individuals who consume its content and society at large, and explore consumer substitution patterns across platforms. Throughout the guide, we examine case studies on the deterrence of misinformation, segregation, political advertisements, and the effects of social media on political outcomes. We conclude with a brief discussion of the future of social media.

JEL Classification: L82, L96, P00, D6

* Aridor: Northwestern Kellogg. Email: guy.aridor@kellogg.northwestern.edu. Jiménez-Durán: Bocconi University, IGIER, and Chicago Booth Stigler Center. Email: rafael.jimenez@unibocconi.it. Levy: Tel Aviv University and CEPR. Email: roeelevy@tauex.tau.ac.il. Song: University of Illinois Urbana-Champaign. Email: lenasong@illinois.edu. We thank the editor David Romer, five anonymous referees, Hunt Allcott, Luis Armona, Luca Braghieri, Leonardo Bursztyn, Yeon-Koo Che, Alex Coppock, Dean Eckles, Sarah Eichmeyer, Ruben Enikolopov, Daniel Ershov, Matthew Gentzkow, Brett Gordon, Emeric Henry, Anna Kerkhof, Alexey Makarin, Solomon Messing, Brendan Nyhan, David Rand, Chris Roth, Carlo Schwarz, Ananya Sen, Catherine Tucker, Nils Wernerfelt, Pinar Yildirim, and Ekaterina Zhuravskaya for helpful comments and suggestions. We also thank Dotan Miller and Michael Reeve for their excellent research assistance. We gratefully acknowledge funding support from the William and Flora Hewlett Foundation and thank Anna Harvey who was instrumental in starting this project.

1. Introduction

Social media platforms play an essential role in the modern economy. While these platforms began as niche websites for interacting with friends, they have become ubiquitous and transformed how people interact and communicate. In 2023, there were 4.76 billion social media users worldwide, comprising 60% of the world population and over 90% of internet users (Kemp, 2023). Internet users spend almost 2.5 hours daily on social media platforms, more than any leisure or media activity besides television (Kemp, 2023). The mass adoption of these applications has resulted in a speed and range of information flow that is unprecedented in history. Businesses, organizations, and politicians use social media to directly connect with individuals, target users with ads, and offer algorithmically curated content to the most relevant consumers. Meanwhile, many individuals receive a large consumer surplus from using these services, become better informed about the world (Allcott et al., 2020), and maintain connections that are helpful in the labor market (Armona, 2019).

While social media platforms provide various benefits, they also bring several new challenges to society. First, the ease of diffusing misinformation (Allcott and Gentzkow, 2017) and hate speech (Müller and Schwarz, 2021) have purportedly affected important political beliefs and behavior (Zhuravskaya, Petrova and Enikolopov, 2020; Guriev et al., 2023). Second, individuals' beliefs and behavior could be influenced by the algorithms used to distribute content on these platforms (Levy, 2021), but there is limited oversight concerning these algorithms. Third, the sheer amount of time spent on these platforms has sparked debates about social media overuse (Allcott, Gentzkow and Song, 2022) and whether growing negative trends in mental health, especially amongst children and young adults, are tied to their rise (Braghieri, Levy and Makarin, 2022).

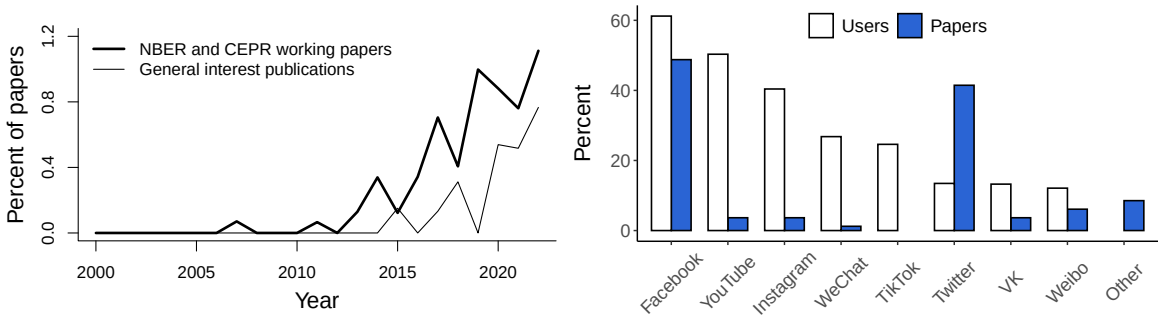
These challenges have led to substantial interest from policymakers, industry players, and academics in understanding the incentives users and platforms face, their societal implications, and how to regulate them. For instance, governments have begun to discuss and craft regulations to increase the accountability of platforms, from Germany's 2017 Network Enforcement Act, to Europe's 2022 Digital Services Act (DSA), to the debate over Section 230 in the United States.

Driven in part by this surge in public interest, academic research studying social media has grown exponentially in recent years across disciplines. Figure 1 shows that the percentage of economics papers published in general-interest journals that study social media has increased four-fold between 2015 and 2022, with much of this work focusing on Facebook and Twitter. The growing supply of social media research and the policy relevance of these topics generate demand for a synthesis and a framework to organize the literature.

This guide covers primarily empirical papers in economics that study social media, and discusses related work in political science, communication, marketing, and computer science when relevant. Our first task is to define what we consider social media in order to determine the scope of the guide, clarify which platforms we cover, and characterize the key economic features of social media that differentiate it from traditional media and other online platforms.

Defining Social Media. We deconstruct the term "social media platforms" into its three components, noting their core features. The "social" component refers to most content being generated by users and involving interactions among them. The "media" component draws on a similarity to traditional

Figure 1. : Social Media Research in Economics



(a) Number of Papers (b) Platform Representation

Notes: Panel (a) shows the share of economic papers that study social media. Social media papers are those whose title or abstract contain "social media," "online social network," "douyin," "facebook," "instagram," "kuaishou," "reddit," "snapchat," "telegram," "tiktok," "twitter," "vk," "wechat," "weibo," or "youtube." The thick line shows the share of social media papers among NBER working papers and CEPR discussion papers. NBER and CEPR papers with the same authors, uploaded within one year, and whose titles have a Levenshtein distance lower than five are counted as a single paper. The thin line illustrates papers published in the following general-interest journals from the EconLit database: AEJ: Applied Economics, AEJ: Microeconomics, AEJ: Policy, American Economic Review, Econometrica, Economic Journal, Journal of the European Economic Association, Journal of Political Economy, Review of Economic Studies, Review of Economics and Statistics, and Quarterly Journal of Economics. In Panel (b), the white bars show the share of users in each platform among all global social media users (Kemp, 2023) and the blue bars show the share of 2000-2022 papers studying each platform. The papers are limited to only empirical papers that analyze data from at least one platform. The figure displays the five most popular platforms and any platform mentioned in more than one paper. TikTok also includes Douyin. The Other column includes the total number of papers that analyzed any other platform divided by the number of papers. We do not include the number of users for these platforms. The bars do not sum to one because users can have accounts on multiple or none of the platforms in the figure and because papers can analyze multiple platforms.

media—that it is typically a two-sided market (Rochet and Tirole, 2003) with users on one side and advertisers on the other. Finally, "platforms" refers to online Internet-based applications that use algorithms to deliver content. Based on these three components, we define social media as two-sided platforms that primarily host user-generated content distributed via algorithms, while allowing for interactions among users.

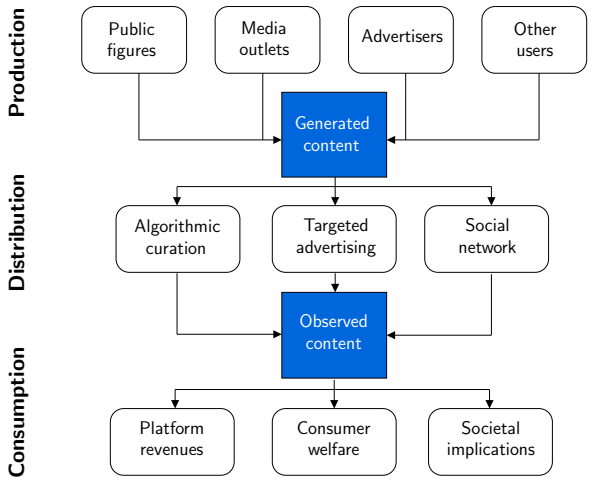
Facebook, Twitter, TikTok, Instagram, and to some extent, YouTube are examples of social media platforms based on our definition. This definition excludes related technologies that lack key components. For example, streaming services typically lack the social component, crowd-sourced discussion forums lack the media component, and email services lack the platform component.

Our definition also helps distinguish social media platforms from traditional media. Perhaps the biggest difference is that any user on the platforms can produce content, which means that the amount of content available is vastly greater than in traditional media. As one example, Meta estimates that its users share one billion "stories" (disappearing posts) every day.[1] As a result of this large scale, social media platforms largely play the role of aggregators—unlike traditional media production, which follows an editorial process. This role introduces a new set of challenges not present in traditional media, such as content moderation (determining allowable content) and algorithmic curation (choosing which content to show users). Facebook, for example, ranks about 1,000 posts per user per

[1]About Stories, Meta: `https://www.facebook.com/business/help/329494947852688?id=2331035843782460`

day for its 2 billion users.[2] Furthermore, social media platforms enable rich social interactions as users are exposed to content and how others react to it (such as through "likes" or comments). The platform choices for these various components influence the type of content that gets both produced and consumed.

Figure 2. : Flow of content



The Structure of the Guide. As depicted in Figure 2, we organize the literature around one key component of social media: the flow of content, starting from its production, to its distribution, to its eventual consumption. While this division is not always clear-cut (for example, content moderation happens at all stages), it clarifies the economic agents involved and their interactions at each of these stages.

We posit a stylized economic framework to elucidate the key economic forces that our guide focuses on at each of these stages. The purpose of this framework is *not* to fully capture the complex set of economic interactions but to point out the high-level incentives at each stage that are the primary focus of our guide. The building block of the framework is a post $x \in \mathbb{R}^K$, represented as an abstract vector of characteristics, which can include, for example, sentiment expressed or indicators of whether posts are ads or contain misinformation.

We begin by discussing the *production* of content in Section 2. The main economic agent at this stage is the producer, j, who solves the following problem:

$$\max_{\mathbf{x}_j^p} \mathbb{E}[u_j^p(\mathbf{x}_j^p)] - c_j(\mathbf{x}_j^p). \tag{1}$$

This problem follows a standard utility maximization: producer j chooses a (possibly empty) set of posts $\mathbf{x}_j^p$ to maximize their expected utility minus costs of production. The types and quantity of content depend on producer beliefs $\mathbb{E}[\cdot]$, the monetary and nonmonetary rewards $u_j^p(\mathbf{x}_j^p)$ that producers get from posting content, and the cost $c_j(\mathbf{x}_j^p)$ of producing the content (for example, the opportunity or physical cost involved in creating or sharing content). We thus begin our discussion of production by focusing on the various incentives and factors that shape $u_j^p(\mathbf{x}_j^p)$ and subsequently the quantity and type of content that gets produced. We then explore how platforms can deter the production of harmful content, such as misinformation and hate speech, by making it more costly to produce (increasing $c_j(\mathbf{x}_j^p)$) or shifting the expectations about its probability of distribution (shifting $\mathbb{E}[\cdot]$).

Section 3 then discusses the *distribution* of social media content. The main economic agent at this stage is a platform that solves the following revenue maximization problem:

$$\max_{\{\mathbf{x}_i\}_i \subset \cup_j \mathbf{x}_j^p} \sum_i \alpha(\mathbf{x}_i) t_i(\mathbf{x}_i). \tag{2}$$

This problem is conceptually simple: The platform chooses a targeting rule that picks, for each user i, a personalized subset $\mathbf{x}_i$ from

[2] See https://engineering.fb.com/2021/01/26/core-infra/news-feed-ranking/

the total pool of posts to show the user. Platforms choose the posts that maximize the revenue-weighted (long-run) time spent or user engagement ($t_i(\mathbf{x}_i)$) on the platform. $\alpha(\mathbf{x}_i)$ represents the monetary gains the platform gets per unit of time spent from showing $\mathbf{x}_i$ to i. In an advertising-based business model, this parameter equals the product of the ad load (share of posts that are advertisements) and the average price paid for the ad.

We partition our discussion into nonmonetized (organic posts) and monetized (advertisements) content and discuss other considerations driving these distribution decisions. First, we discuss how platforms target users with specific posts or ads. For organic content, we discuss the role of social networks, while for advertisements, we discuss the role of off-platform data. Second, we discuss empirical work that quantifies the extent to which targeting "works" (i.e., whether t_i increases due to targeting). Third, we discuss which posts, $\mathbf{x}_i$, tend to get chosen at the distribution stage (for example, whether social media algorithms promote low-quality content). Finally, we discuss the implications of targeting for downstream outcome variables. In particular, we devote a large portion of our discussions to its political consequences: whether platform algorithms result in echo chambers and the effects of targeted political advertisements.

In Section 4 we discuss *consumption*, the final stage of the flow of content. End consumers solve the following problem:

$$\max_{t_i, \mathbf{a}_i} \mathbb{E}[u_i^c(t_i, \mathbf{a}_i; \mathbf{x}_i)]. \tag{3}$$

This is fundamentally a time allocation problem: Consumer i chooses how much time to spend on social media t_i and other activities $\mathbf{a}_i$, as a function of content observed on social media $\mathbf{x}_i$, in order to maximize their expected utility ($\mathbb{E}[u_i^c]$).

We begin this section by clarifying the intricacies around understanding consumer choice t_i and the associated individual welfare from social media consumption. Specifically, we discuss what enters into the utility function, highlight the role of consumption spillovers, time inconsistency, and habit formation, and interpret the differences across various welfare measures. Next, we turn to the societal implications of social media consumption, which occur through beliefs and off-platform activities $\mathbf{a}_i$. We summarize the channels through which social media consumption can lead to aggregate impacts and present case studies on how it affects political knowledge, political participation, polarization, and offline violence in democracies. Finally, we describe consumer substitution patterns across different social media platforms and their economic implications.

Throughout, we mention open questions specific to the topic of each respective section. In Section 5, we conclude the guide with a discussion of the future of social media, highlighting areas for future research that are relevant across the stages in the life cycle of content.

2. Content Production

Social media companies rely on user-generated content to attract users. However, as a result of low entry costs and the large scale of these platforms, they typically cannot directly control the content produced. Instead, they use platform design—the features, incentives, and rules of a platform—to indirectly shape content (Luca, 2015). This indirect shaping contrasts starkly with the editorial process in traditional media, which more directly shapes content production. As part of this process, social media companies often trade off increasing content production and engagement with the risks associated

with certain types of content.

In this section, we first describe what incentivizes the production of content and the implications both on and off the platform. We then review case studies of how platforms and community members define boundaries of acceptable content and deter negative content. Throughout this section, we discuss both the production of original content and the resharing of existing content.[3] In both cases, users implicitly inform the algorithm that content is important and should be shown to their friends.

2.1. How Social Media Affects Content Production

How do incentives—including reactions from others, the algorithm, and the revenue structure—affect content production? We first discuss the effects on content generated within social media platforms and then discuss spillovers to content generated outside social media.

2.1.1. User-Generated Content

The nonrivalrous nature of social media content makes it akin to a (possibly excludable) public good. Two distinct types of incentives induce users to produce this good: nonmonetary and monetary incentives.

Nonmonetary Incentives. Theoretical work focused on social media has modeled roughly five types of nonmonetary incentives to share or produce content: 1) receiving attention or attracting eyeballs, 2) improving social image or reputation, 3) receiving peer awards or feedback (including badges, reactions, likes, and comments), 4) persuading others, and 5) intrinsic or altruistic motives, which can also include keeping up with friends (Abreu and Jeon, 2020; Acemoglu, Ozdaglar and Siderius, Forthcoming; Filippas, Horton and Lipnowski, 2021; Bursztyn et al., 2023*b*; Guriev et al., 2023).[4]

Existing empirical work typically studies policies or experimental interventions that vary multiple types of incentives simultaneously. For example, a content producer who receives additional likes could derive direct benefits from them but also update her beliefs about how much attention her posts get, how reputable she is, and how persuasive her content is. We therefore refer collectively to these as nonmonetary incentives, but we note that a gap in this literature is to disentangle the effect of each type of incentive. A recent contribution in this vein is Guriev et al. (2023), who calibrate a structural model of news-sharing decisions using data from an experiment of misinformation interventions (described in more detail in Section 2.2). The authors find that both reputational and partisan motives for sharing political information are important and that the persuasion motive dominates partisan signaling.

A first lesson from the empirical literature is that nonmonetary incentives moderately increase the quantity and frequency of content produced (with short-lived impacts, typically lasting less than a week), across different types of incentives and different platforms. For example, Eckles, Kizilcec and Bakshy (2016) exploit an experiment that led Facebook users to receive more likes and comments and found an elasticity of posts produced of 0.07 (i.e., doubling the number of likes or comments received increases the number of posts produced by 7%). Zeng et al. (2022) find that producers on a Chinese video-sharing social media platform who could randomly see "pokes" (nudges) that other users sent

[3]More than a quarter of posts in Facebook feeds are reshared (Guess et al., 2023*a*).

[4]Nonmonetary rewards also incentivize content in other contexts such as Wikipedia (Zhang and Zhu, 2011) or recommender systems (Chen et al., 2010).

them increased their content production by 13% in the first day after the intervention. Comparable effects have been found with field experiments on Reddit, by randomly giving badges (Burtch et al., 2022) or AI-generated comments on posts (Srinivasan, 2023). Huang and Narayanan (2020) and Mummalaneni, Yoganarasimhan and Pathak (2023) find similar results with platform experiments that increased the prominence of content on an art-sharing social network and on Twitter, respectively. Moreover, in some contexts, even *negative* peer awards (downvotes on Reddit) have been found to increase content creation (Deolankar, Fong and Sriram, 2023).

A second lesson is that nonmonetary incentives given to one content producer may propagate to other producers; that is, the recipient of an incentive becomes more likely to give nonmonetary incentives to other producers (Eckles, Kizilcec and Bakshy, 2016; Huang and Narayanan, 2020; Mummalaneni, Yoganarasimhan and Pathak, 2023). This finding suggests that partial equilibrium estimates can differ from general equilibrium responses that account for this propagation effect. Zeng et al. (2022) use their experimental estimates to calibrate a structural model of network diffusion and find that the general equilibrium effect on content production is 8% higher than the partial equilibrium effect.

A third lesson is that the effect of nonmonetary incentives on content production is increasing in the perceived quality of the incentive. Srinivasan (2023) finds that randomly allocating six AI-generated comments, as opposed to three, on Reddit users' posts has a lower effect on the number of posts produced, which is partly explained by comments in the former treatment arm being perceived as lower quality (more likely to be accused of being bots and downvoted). Zeng et al. (2022) show that the effect of nudges on video producers is higher when the producer also follows the user who sent the nudge.

A fourth lesson is that nonmonetary incentives have a small effect on the quality of content produced, often proxied by the number of likes received (Zeng et al., 2022; Srinivasan, 2023). Given the evidence that subsequent content produced becomes more similar to the content that receives a nonmonetary incentive (Burtch et al., 2022), a follow-up question is whether these incentives allow content producers to better learn the tastes of their audience.

Monetary Incentives. Influencers and major content creators may also receive monetary incentives to generate content, such as participating in revenue-sharing programs, posting sponsored content, or receiving direct payments from other users.[5]

Do monetary incentives increase the amount of user-generated content? The answer is not obvious; higher monetary rewards increase the marginal benefit of producing content, but they might also change nonmonetary incentives—for example, by making users appear less pro-social (Bénabou and Tirole, 2006). Nevertheless, the literature has found a strong positive effect of ad-revenue-sharing programs. Kerkhof (2020) studies a sudden increase in the salience of YouTube's revenue-sharing rules and provides evidence of an increase in the monthly number of uploaded videos. Abou El-Komboz, Kerkhof and Loh (2023) find that creators who lost access to YouTube's ad-revenue-sharing program posted 86% fewer monthly videos relative to the mean of those who did not lose access.

Monetary incentives can also impact the

[5]Career concerns can also motivate user contributions (Lerner and Tirole, 2002), but evidence of these drivers is scarce in the context of social media. An exception is Petrova, Sen and Yildirim (2021), who document that donations to politicians running for U.S. Congress increase after they open a Twitter account.

quality and variety of content supplied, but the evidence in this regard is scarce and mixed, as in the case of nonmonetary incentives. Some studies find evidence consistent with ad-revenue-sharing programs increasing the quality of content produced and its originality or differentiation from existing content (Abou El-Komboz, Kerkhof and Loh, 2023). However, an early study by Sun and Zhu (2013) finds that the introduction of a revenue-sharing program by Sina (a precursor of the Chinese social media platform Weibo) increased quality but decreased differentiation. Kerkhof (2020) found opposite results: Increased advertising opportunities for YouTube content creators increased differentiation but reduced quality. These differences across studies could be driven by differences in the status-quo that they analyze: Removing a program (Abou El-Komboz, Kerkhof and Loh, 2023) could differ from introducing a program (Sun and Zhu, 2013; Kerkhof, 2020), since the former potentially entails losing status as a platform partner. Another explanation for the different findings is the presence of confounders: Some of the studied interventions varied not only producer incentives but also the amount of advertisements shown to consumers (Sun and Zhu, 2013; Kerkhof, 2020), which could lower their willingness to like content.

Lastly, content creation is increasingly viewed as a viable career or income source.[6] A natural next step to the existing evidence on the elasticity of the content supply curve concerns the labor economics of this activity, studying questions such as the effects of unions for content creators or whether monetary incentives crowd out nonmonetary motives. Beyond ad-revenue-sharing programs, other monetary incentives that have been increasingly used by platforms (for example, allowing users to subscribe to producers) remain understudied, perhaps due to missing data. Indeed, Ershov, He and Seiler (2023) estimate that 96% of sponsored content on Twitter is undisclosed. Future research will need to overcome these data challenges to understand the effect of new business models on content production.

To conclude, nonmonetary incentives such as receiving peer awards or feedback tend to have a moderate short-run impact on the quantity of content produced, while monetary incentives such as ad-revenue-sharing programs seem to have a strong positive effect. As opposed to quantity, the quality of content produced is relatively more difficult to influence.

2.1.2. Content Outside Social Media

As social media platforms become more prominent, their effects on content production are no longer confined to content produced on the platform. Social media platforms provide content producers with new data on the engagement of their audience and often serve as a primary gateway for news. They also threaten the business models of traditional news producers.[7] Qualitative evidence documents that online traffic and social media algorithms can affect news production processes, for example, by having editors prioritize social media traffic (for example, Smith, 2023). However, there is limited rigorous evidence for this phenomenon, perhaps due to the challenges in identifying a causal effect.

Cagé, Hervé and Mazoyer (2022) provide direct evidence for the effect of social media

[6] "Social media and gaming" was the fourth most popular career choice for UK kids according to a 2018 survey: `https://www.educationandemployers.org/wp-content/uploads/2018/01/DrawingTheFuture.pdf`.

[7] Angelucci, Cagé and Sinkinson (Forthcoming) show how the entry of television threatened newspapers' revenue and affected the content they produced. An emerging literature, which we do not discuss in detail in this guide, analyzes how search engines and social media platforms affect the profits of news publishers. Holder et al. (2023) estimate that Meta owes $1.9 billion to news publishers in the United States as fair payment for the engagement generated via their content.

on online news production. The authors exploit social media news pressure (a measure of the amount of activity on the platform in the hour before the first "seed" news post) and the centrality of the user who posted the "seed" post to instrument the popularity of the news story on Twitter. They find that social media popularity increases mainstream media coverage.[8]

At least two mechanisms could explain the effect of social media on news production: Social media may provide journalists with a novel source for news, and it may give editors information on consumers' interests. These two mechanisms have been studied separately. In terms of user-generated content, Hatte, Madinier and Zhuravskaya (2023) exploit internet outages to show that social media posts provide new information on the Israeli-Palestinian conflict. These posts increase the emotional coverage of the conflict and the focus on civilians by traditional media. In terms of information on consumer preferences, Leung and Strumpf (2023) find that the New York Times is more likely to change the headline of articles following negative comments on Twitter. Relatedly, Sen and Yildirim (2015) find that editors increase coverage of online news stories receiving more clicks, providing further evidence that information on popularity shapes content production.

More research is needed on how social media algorithms could affect the production of other types of content, beyond news. For example, it has been argued anecdotally that TikTok is driving songwriters to focus on brief danceable 15-second snippets.

[8]Fortunately, not only like-minded content (see section 3.1.1) and emotional content (Brady et al., 2017) are popular on social media. Cagé, Hervé and Viaud (2020) find that original content also receives more views on social media, and therefore outlets still have incentives to invest in newsgathering.

2.2. Deterring the Production of Harmful Content

One major challenge for social media platforms is content moderation: defining rules outlining the types of content that users are allowed to produce and enforcing sanctions against those that violate these rules. The role that social media platforms play in regulating online speech has raised concerns about these companies becoming "arbiters of the truth." For this reason, academics and policymakers have sought to understand the online and offline effects of this self-regulation and the incentives of platforms to engage in it.

Most of the literature studies interventions targeting misinformation and toxic content due to their policy relevance, so we divide this section based on these two types of content. We define these types of content below, but we refer to them as "harmful" because the literature works with the assumption that they impose negative externalities on certain segments of the population. These externalities could harm other social media users; for example, one-third of adult Americans were harassed online (including through social media) in 2022, which could bring them a reduction in utility (u_i^c).[9] There can also be externalities on nonusers; for example, even if misinformation is only 0.15% of Americans' daily media diet (Allen et al., 2020), it might lead to poorly informed voters or other welfare-reducing offline actions ($\mathbf{a}_i$).

All platforms moderate content to some extent, forbidding illegal content and typically a combination of hate speech, harassment, misinformation, spam, and graphical content. They use a mix of algorithms and human supervision (which can include moderators contracted by platforms

[9]See: `https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2023`.

or users themselves) to detect content that violates their rules and impose sanctions. Sanctions include post-level interventions, such as deletions, algorithmic filtering (also called "shadowbanning"), and adding labels or tags, and user-level interventions such as account suspensions or bans (Gillespie, 2018).[10] Theoretical work has assumed that platforms moderate to maximize their profits, by avoiding regulatory penalties, optimizing user engagement, or increasing advertisers' willingness to pay (Liu, Yildirim and Zhang, 2022; Madio and Quinn, 2023; Jiménez Durán, 2022). This work has also shown that the incentives of the platform to moderate content are not necessarily aligned with those of the users. For example, Beknazar-Yuzbashev, Jiménez Durán and Stalinski (Forthcoming) argue that engagement increases on social media need not correspond with increases in user utility because harmful content may decrease utility while being complementary with engagement, similar to ads in Becker and Murphy (1993).

Moderation and related interventions operate at all stages of the content life cycle. This section focuses on the deterrence of the production and sharing of content (which affect the set of potential posts that platforms can display to users), while Section 4.2.2 reviews the effects on content consumption.

2.2.1. Interventions Targeting Misinformation

We use "misinformation" as an umbrella term encompassing many others (for example, "disinformation" and "fake news"), referring to content that is determined to be false by an authoritative third party. This definition roughly captures the definition used by the academic literature, social media companies, and regulators. In practice, while algorithms that detect misinformation exist and are used by platforms at scale, the empirical literature typically measures misinformation using a set of news, outlets, or URLs considered to be "ground truth" (rated as true or false by professional fact-checkers).

The literature has mostly focused on the sharing as opposed to the production of misinformation. One potential reason for this imbalance is the role of resharing in diffusing misinformation. For example, Vosoughi, Roy and Aral (2018) find evidence that false stories diffuse more broadly than true stories. An important gap in this literature is to understand the determinants of the *production* of misinformation, beyond the sharing of existing articles.

Existing work has primarily studied interventions targeting misinformation initiated by the research teams themselves or by third parties such as fact-checkers. The theoretical literature suggests that these interventions affect the sharing of misinformation by altering 1) the cost of sharing content; 2) how users update their beliefs about the veracity of content; and 3) social-image concerns such as the reputation from sharing misinformation (Papanastasiou, 2020; Acemoglu, Ozdaglar and Siderius, Forthcoming; Thaler, 2021).

An initial question is whether these interventions work. We split the question into two parts; this section focuses on the impact on sharing and Section 4.2 reviews the impact on user beliefs. Studies that analyze sharing decisions typically measure success based on sharing *discernment*—the proportion of true news shared or intended to be shared minus the proportion of false news shared or intended to be shared. The use of sharing discernment by these studies as a main outcome responds to a lesson from early work that misinformation interventions typically affect not only the sharing of false information but also the sharing of true in-

[10] See also `https://www.platformgovernancearchive.org` for an archive of social media platforms' content moderation and legal policies.

formation. An important tradeoff when evaluating these interventions from a social welfare perspective is whether they can reduce misinformation sharing while having a non-negative impact on the sharing of truthful information.

In general, meta-analyses and literature reviews (Kozyreva et al., Forthcoming; Pennycook and Rand, 2022; Martel and Rand, 2023*a*; Blair et al., 2023) show that nudging users to think about the accuracy of content or the prevalence of misinformation, journalistic fact-checking, administering digital literacy campaigns (occasionally also known as inoculation or prebunking), and adding friction to the sharing process are effective at reducing the willingness to share misinformation.

Nudges. Pennycook and Rand (2022) meta-analyze over 20 randomized experiments that nudge users to think about the accuracy of content before sharing decisions and find an average effect size of 3.8 percentage points (71.7%) increase in stated sharing discernment, primarily by reducing sharing intentions for false news. Arechar et al. (2023) find that this effect is robust across 16 different countries, but there is substantial variation in the magnitude of the effect.

Fact-Checking. There is evidence that providing journalistic fact-checking information decreases self-reported sharing of misinformation (Kreps and Kriner, 2022). Beyond stated preferences, Henry, Zhuravskaya and Guriev (2022) document that journalistic fact-checking decreases the sharing of false news and increases the sharing of the fact-checking information. Importantly, merely offering users the option of voluntarily accessing fact-checking is as effective as imposing it on them, with a two percentage points (45%) decrease in sharing. This finding could be driven by users being primed to think about the accuracy of news or updating their beliefs about veracity. This result is relevant given that a common argument against interventions is the infringement of freedom of expression, which would suggest that voluntary interventions are more politically feasible to implement.

There is evidence that adding warning labels to posts (for example, indicating that they have been disputed by fact-checkers) can decrease sharing intentions of false news and increase stated sharing discernment (Martel and Rand, 2023*a*,*b*). However, some studies find null effects or even an increase in the sharing of certain false news (Kreps and Kriner, 2022). Moreover, Pennycook et al. (2020) find evidence of an implied truth effect, whereby adding tags can reduce the willingness to share tagged false news but increase the willingness to share *untagged* false news if users interpret the absence of tags as a signal of veracity. One gap in this literature is to disentangle a potential dual role of fact-checking interventions, which affect not only the users' perceived veracity of the content they are about to share but also the perceived likelihood that they will be fact-checked by the platform in the future.

Digital Literacy. Interventions promoting digital literacy typically provide tips to detect misinformation (Guess et al., 2020). These tips can be administered in different ways, including educational videos (Roozenbeek et al., 2022) or text messages (Athey et al., 2023*b*). One commonly mentioned advantage of conducting digital literacy campaigns over fact-checking individual posts is that the skills learned can be transferable across types of content—users can learn to distinguish between true and false information as opposed to learning that a specific piece is false. In terms of effectiveness, campaigns that train users to identify emotional manipulation are particularly effective at in-

creasing stated sharing discernment, with effect sizes of at least 0.2 SD (Roozenbeek et al., 2022). As Athey et al. (2023*b*) show, these effects persist for a few months and are not explained by making the topic of misinformation more salient. Moreover, tips to identify emotional manipulation are relatively more effective than those teaching reasoning-based techniques.

Friction. Increasing the mechanical cost of sharing (c_j in Equation 1)—for example, by requiring additional clicks or requiring users to pause before sharing—can decrease the likelihood of sharing misinformation. Henry, Zhuravskaya and Guriev (2022) find that requiring an additional confirmation decreases the likelihood of sharing both false news and fact-checking information. Guriev et al. (2023) further show that requiring an extra click to share news decreased the sharing of false news by 3.8 percentage points and had an insignificant effect on the sharing of true news. These policies, however, can have unintended consequences depending on the relative elasticity of sharing different types of content to the friction cost. Ershov and Morales (2024) find that when Twitter increased the cost of reposting content, the overall sharing of news decreased, with left-wing news outlets being relatively more affected than right-wing outlets—inducing a likely unintended partisan bias on sharing decisions. This policy, intended to make users pause before sharing content, had to be reversed due to its unintended effects.[11]

Beyond the success of these measures in isolation, other policy-relevant questions are to compare the effectiveness *across* interventions and to disentangle the mechanisms that underlie the estimated effects. Guriev et al. (2023) make important progress on both fronts and document that nudging users to think about the prevalence of fake news is more effective (in terms of changing the balance of shared news toward true content) than 1) adding friction to the sharing process; 2) nudging users to think about the accuracy and partisan slant of content; and 3) offering the option to access fact-checks. Athey et al. (2023*b*) also compare different interventions and find that the effect of digital literacy courses is more than double the effect of accuracy nudges, with no evidence of complementarity between these policies.

In terms of mechanisms, counterfactual simulations from the structural model in Guriev et al. (2023) rule out that these interventions substantially affect users' beliefs about the veracity of the content and instead the effect is driven by 1) how interventions increase the salience of reputational concerns from sharing misinformation and 2) how they increase the friction of the sharing process. The structural model also shows that digital literacy training reduces the circulation of fake news primarily by changing the sender's beliefs that better-informed receivers would not be persuaded and would negatively update their view of the sender's knowledge.

Given the success of these interventions, a natural question is whether they are scalable to the level at which social media platforms operate. These effects may be smaller when scaled due to general equilibrium adjustments or having a broader user sample. For example, Lin et al. (2024) find a muted effectiveness of accuracy nudges displayed via ads in large-scale experiments on Facebook and Twitter, in line with effect sizes observed in digital advertisement experiments (see Section 3.2). A promising measure that platforms have implemented at scale is crowd-sourced fact-checking, which relies on users adding notes and contextual annotations to others' posts. An example

[11] https://blog.twitter.com/en_us/topics/company/2020/2020-election-update

of this tool is Twitter's Community Notes—an algorithm that publishes user-generated notes that are highly rated by users of different viewpoints (Wojcik et al., 2022). One of the main challenges with implementing such an algorithm is to align user incentives to provide truthful fact-checking. Indeed, partisanship better predicts the ratings that users give to fact-checking notes than the content of the notes and the fact-checked posts (Allen, Martel and Rand, 2022). Nevertheless, despite the important role of partisanship, crowd ratings are still strongly correlated with professional fact-checker evaluations (Martel et al., 2024). The effect of these crowd-sourced fact-checks on the production and sharing of misinformation, and the extent to which the algorithm adequately incentivizes the crowd remain to be studied.

Besides crowd-sourcing, platforms conduct other content moderation measures at scale, such as downranking or removing posts, banning groups, and suspending user accounts. Correlational evidence in Lin et al. (2024) suggests that resharing is almost entirely eliminated after Facebook classifies posts as containing misinformation, which may lead to sanctions such as labeling the posts or downranking them. However, more research is needed to understand the causal effects of these "harder" interventions on the production of misinformation and the mechanisms through which they operate, whether they crowd out fact-checking efforts by the users, and the net welfare effect of sanctions.

To summarize, across interventions, digital literacy campaigns and nudging users to think about the prevalence of misinformation seem to be the most effective policies to increase sharing discernment. An important challenge is implementing these solutions at the large scale at which platforms operate.

2.2.2. Interventions Targeting Hate Speech and Toxic Content

There is no single definition of hate speech but almost all platforms forbid it either explicitly or include it in broader categories such as personal attacks. Platforms' guidelines typically borrow from U.S. antidiscrimination law and define hate speech as attacks based on protected categories such as race or gender (Gillespie, 2018). Besides hate speech, platform rules cover related content such as harassment—attacks that do not have to be based on a protected category.

Classifying posts as hate speech or other similar types of content is an inherently subjective task. Even expert content moderators disagree substantially in their judgments (Lucas, Alm and Bailey, 2019), and it is unclear whether this disagreement reflects "vertical" differentiation (in beliefs about the likelihood that content is hateful) or "horizontal" differentiation (in tastes for hateful content). Platforms and researchers alike deal with this challenge by combining approaches that range from manual annotation to algorithmic classification.[12] Platforms' internal algorithms are often trained to predict the probability that content violates their rules (Ribeiro, Cheng and West, 2022; Thomas and Wahedi, 2023) but they also use—in line with most of the academic literature—algorithms that predict other outcomes such as the toxicity of content (Katsaros, Yang and Fratamico, 2022). In many applications, "toxicity" is defined as rude, disrespectful, or unreasonable messages that are likely to make someone leave a discussion.[13] Given that the literature does not study one widely established outcome, in

[12]See, for instance, the rulebook that Facebook gives its content moderators, which was leaked to the press in 2017: `https://www.theguardian.com/news/gallery/2017/may/24/hate-speech-and-anti-migrant-posts-facebooks-rules`.

[13]This definition follows from the one used by Google's Perspective algorithm, which is widely used in the industry and as a benchmark in academic studies. See `https://www.perspectiveapi.com/how-it-works/`.

this guide we use "toxicity" as an umbrella term that captures many commonly studied types of language (for example, racist, xenophobic, or misogynistic language).

Counterspeech. One way to reduce toxicity is through counterspeech—sending messages to challenge users who post toxic language. There are several takeaways from this literature. First, the effectiveness of counterspeech largely depends on the design of the message. Messages that prime users to be more empathetic (Hangartner et al., 2021) or that include moral references (Siegel and Badaan, 2020; Munger, 2021) tend to be successful (with small effect sizes, in the order of a 0.1 SD decrease in posts in one month). In contrast, messages using humor or warning users of the consequences of their posts on others or themselves tend to have insignificant effects (Hangartner et al., 2021). The credibility of the counterspeech message—which can be signaled by the number of followers (Munger, 2017) or by referring to an authority in the message (Yildirim et al., 2021)—also matters. Second, counterspeech interventions can also reduce the production of nontoxic speech (Hangartner et al., 2021), but the mechanism underlying this effect is largely understudied. Third, these interventions can also impact other users who observe the counterspeech (besides the producers of the toxic content): Siegel and Badaan (2020) find that exposing survey respondents to some forms of counterspeech reduces the rating they give to hate speech posts and decreases their willingness to share these posts (although the effect is not precisely estimated).

An open question is what determines the equilibrium provision of counterspeech and how to incentivize users to provide this public good (similarly to fact-checking). One possibility is for platforms to provide counterspeech. In practice, they conduct a similar type of intervention, with the difference that they typically nudge users *before* they post content. In a large-scale experiment conducted by Twitter, asking users to review toxic language before replying to other users moderately decreased their number of toxic replies over six weeks by 6.4% relative to the control group mean (or 0.02 relative to the control group SD), without significantly decreasing the total replies sent (Katsaros, Yang and Fratamico, 2022). While the effect of this intervention was small, it has been implemented at scale by other platforms including Instagram, YouTube, and TikTok, potentially due to its low cost and since it did not decrease engagement.

Content Filtering. Platforms commonly hide or limit the visibility of content whose toxicity score exceeds certain thresholds (Ribeiro, Cheng and West, 2022), in part due to the concern that toxic content is contagious; that is, that higher exposure to it will increase the incentives of users to produce or spread this type of content. Along these lines, Beknazar-Yuzbashev et al. (2022) conduct an experiment using a browser extension that hides content on Facebook, Twitter, and YouTube whose toxicity exceeds a certain threshold. Reducing the exposure of users to toxic content for six weeks reduced the average toxicity of the content they posted (with an elasticity as high as 0.3, or an effect size of around 0.10 SD), in line with prior evidence on the contagion of toxicity (Kim et al., 2021). They also provide suggestive survey evidence that individuals' evaluations of what constitutes toxic content do not change. Therefore, other mechanisms such as reciprocity (for example, responding to toxic content with more toxic content), changing beliefs about the social acceptability of toxicity, or the likelihood of being moderated could be at play.

Ex-Post Moderation. Ex-post moderation consists of removing posts or restricting or suspending user accounts or groups. These actions are typically more visible than content filtering (for example, platforms leave notices that the post has been removed). A challenge with providing evidence about this type of intervention is the intensive data requirements. Researchers need internal data—which is difficult to obtain since this is a sensitive topic for platforms—or to track content or accounts in real time to measure when they get deleted.

Recent work providing causal evidence of these sanctions tends to find insignificant or small deterrence effects, with a negligible impact on the sanctioned users' subsequent engagement with the platform. Jiménez Durán (2022) exploits the reporting tool on Twitter that allows flagging toxic content which is then presumably reviewed by the platform. Randomly reporting posts with hateful slurs increases by 66% (1.4 pp) the likelihood that Twitter deletes them and there is evidence that the platform imposes other sanctions such as locking users' accounts for some time. However, the engagement (an index of posts and likes given) of the reported accounts or the average toxicity of their posts does not change significantly up to five months after treatment. Ribeiro, Cheng and West (2022) use a regression discontinuity design exploiting Facebook's automatic deletion of comments with toxicity above a certain cutoff and find a 0.1 SD decrease in subsequent rule violations over the four weeks after the deletion. The effect on engagement (measured by the number of comments given) is not significant after two weeks.

Similar small deterrence effects are seen when studying the spillover of interventions on other individuals close to the sanctioned users. Thomas and Wahedi (2023) exploit the staggered banning of hundreds of core members across six hateful organizations on Facebook. Among surviving users, they report a precisely estimated null effect on the fraction and amount of hateful content created. The number of views on hateful content drops by 0.06 SD, but the views on hateful content as a fraction of total views do not change, which is explained by a small drop in engagement. In a similar vein, Müller and Schwarz (2022) find that Twitter followers of Donald Trump decreased their monthly number of toxic tweets by 0.037 SD and total number of tweets by 0.05 SD relative to non-followers after Twitter suspended his account on January 8th, 2021. An open question is whether targeting more prominent users decreases engagement on the platform as a whole and whether platform incentives to moderate more visible cases differ from accounts with smaller reach.

Government regulation provides another source of variation for content moderation. For example, Germany's 2017 NetzDG law introduced fines of up to 50 million euros for social media companies that fail to promptly remove hateful content. The passage of this law was associated with a subsequent decrease in the prevalence of toxic content on social media in the order of 0.08 SD (Andres and Slivko, 2021; Jiménez Durán, Müller and Schwarz, 2022). This policy seems effective at reducing the prevalence of toxic content but it remains unclear whether the effect is mechanical (due to the removal of posts and users) or due to the deterrence of hateful behavior. Moreover, this type of regulation introduces the potential for spillovers to more extreme niche platforms, given that it typically applies only to large platforms.[14] If

[14]The NetzDG covers platforms with more than 2 million active German users. The European DSA introduces obligations for Very Large Online Platforms with more than 45 million users in the EU. These obligations include mitigating risks such as the dissemination of illegal content, disinformation, and gender-based violence.

coordination on social media is a mechanism for the link between online hate and offline violence (see Section 4.2.1), then selectively regulating big platforms may be ineffective, as users can find other places to coordinate. Indeed, Beknazar-Yuzbashev et al. (2022) show that lowering users' exposure to toxicity leads them to increase their engagement on other social media websites, but more causal evidence is needed on whether moderation on one platform increases the toxicity produced on others.

More evidence is needed on the connection between content moderation and advertising. Specifically, there is limited research examining how content moderation influences advertisers, and conversely, how advertising dynamics influence content moderation decisions. One notable exception is the study conducted by Ahmad et al. (Forthcoming), which demonstrates that consumers tend to avoid companies whose advertisements are featured on misinformation outlets. However, more evidence is needed to understand the effect of hate speech and other types of content on user interactions with advertisements and whether content moderation policies can alleviate any potential negative effects.

To conclude, interventions reducing the exposure of users to toxic content and some forms of counterspeech can deter the production of toxic content with small effect sizes (under 0.1 SD), while harder sanctions, such as post deletions, have null or small effects.

3. Content Distribution

After content is produced and posted on social media, the platform decides how to distribute it to users. Individuals are typically exposed to two types of content in their feeds: organic content, discussed in the previous section, and advertisements. In the context of our framework, the platform's revenue from post consumption, α, is typically zero in the case of organic content and nonzero for advertisements. Thus, while organic content is the primary reason users log in to social media platforms, the platform only accrues revenue when users consume advertisements. We focus on these two types of content separately: we first discuss the determinants and economic ramifications of how platforms choose the set of organic content a user observes and then discuss the implications of advertiser-generated content.

3.1. Organic Content

Since individuals spend several hours per day on social media and since the posts they are exposed to may affect their well-being, economic outcomes, and society at large (see Section 4), it is important to understand what content individuals observe on these platforms. In the past, one's network was the main source of content on social media platforms.[15] The platforms simply showed individuals content generated by their friends in a reverse-chronological-order (RCO) feed. In Section 3.1.1 we discuss these networks, how they form, and their implications. While social networks are still important, today content is typically curated by algorithms.[16] Initially, these algorithms ranked potential posts from the accounts people follow. The algorithms of newer platforms, such as TikTok, show users *any* content that is likely to generate interest. In Section 3.1.2, we discuss algorithms, how they may benefit users, and their potential dangers. One concern that is common for both content shared by friends and content promoted by algorithms is that it may gener-

[15] In fact, these platforms were described as "social networks," as reflected in the title of the 2010 film *The Social Network* about Facebook's founding.

[16] Technically, showing content from friends in an RCO feed is also an algorithm. However, throughout this section, when we mention algorithms, we refer to ranking systems, which determine which posts to show users based on various signals.

ate segregation in news exposure. We discuss this concern in Section 3.1.3.

3.1.1. Online Social Networks

Until recently, the networks people formed were a critical aspect of social media platforms. Beyond their role as an input to algorithms, these networks are studied because they provide a unique opportunity to observe complex social connections and analyze how they evolve. For example, Chetty et al. (2022) use over 20 billion friendships on Facebook to study social capital.

A fundamental question is whether online social networks are characterized by homophily, the tendency of similar individuals to form ties. Audit studies causally answer this question by creating fictional accounts with randomized characteristics, which then follow actual users on Twitter and test whether these users reciprocate. They find that individuals are more likely to follow accounts with congruent ideological identities (Mosleh et al., 2021; Ajzenman, Ferman and C Sant'Anna, 2023). This method cleanly detects a causal effect, but it cannot characterize the full network of connections. Barberá (2015) and Bakshy, Messing and Adamic (2015) estimate the ideology of active Twitter and Facebook accounts, respectively, and find that users are indeed more likely to follow or befriend others aligned with their ideology. Furthermore, Halberstam and Knight (2016) find that the degree of homophily in Twitter's political network resembles other social networks, such as offline high-school friendships. While homophily exists, Barberá (2015) finds that social media are also characterized by many "weak ties" between individuals who do not necessarily have the same ideology (for example, distant family members).

Online social networks influence both the content users observe and the context in which it appears. In terms of content, descriptive research finds that people are more likely to share like-minded political news (Garz, Sörensen and Stone, 2020). Laboratory and survey experiments confirm the tendency to share like-minded social justice posts (Song, 2023) or news (Pogorelskiy and Shum, 2019). In homophilic networks, such sharing behavior may result in individuals being exposed predominantly to like-minded content on social media. On the other hand, content shared by weak ties may expose people to cross-cutting content that they would not have been exposed to otherwise.

Individuals are not simply exposed to personalized content on social media; they see it in specific contexts, as they observe who shared the content and how popular it is. Messing and Westwood (2014) conduct an experiment where participants observe the outlet where an article appears, the number of people recommending the article, or both pieces of information. Unsurprisingly, participants prefer content from like-minded sources, but interestingly, observing the number of recommendations eliminates this preference. Relatedly, Dvir-Gvirsman (2019) finds that social cues (such as likes or comments) moderately affect the attention given to posts and the likelihood of clicking them. These results suggest that users do not treat all social media content equally.

Finally, homophily on social networks can also affect offline behavior. Enikolopov et al. (2024) exploit a conflict between Facebook and Google that generated quasi-random variation in the connections between counties. They find that an exogenous increase in the share of connections with others from socio-economically and politically similar counties increased Facebook usage, demonstrating a demand for homophily. However, increased homophily also reduced people's interactions offline.

To conclude, based on observational, quasi-experimental, and experimental data,

there is demand for homophilic connections on social media platforms. The homophily of social media networks matters because it can affect the content people are exposed to and how they perceive it. Specifically, homophily may result in exposure to like-minded content, which we examine in more detail in Section 3.1.3. However, one's social network is not the only factor affecting the feed—posts shared by friends are still ranked and filtered by algorithms, which we discuss next.

3.1.2. Social Media Algorithms

Today, every major social media platform relies on algorithms to choose content, and it is hard to imagine these platforms having as much influence without algorithms. These algorithms operate similarly to other recommender systems (RS) in the sense that they rank potential content (posts) and determine in which order to provide the content to the user.[17] Even though algorithms may improve the experience of users, they remain a major source of controversy and users are skeptical of them. For example, only 30% of the respondents in the 2023 Reuters News Report survey agreed that having algorithms select stories based on previous news consumption is a good way to get news (Newman et al., 2023). In this section, we discuss the methodological challenges in studying algorithms, their economics, and concerns related to them. We focus mostly on political content, not because that content is especially prevalent on social media, but rather because political content can have important off-platform consequences and thus receives more attention in the literature.

There are both data and design challenges in studying algorithms. First, it is difficult to obtain data on the posts distributed to users and even more difficult to observe the set of potential posts that the algorithm ranks. Second, it is challenging to find or generate random variation in algorithms that can be exploited to estimate causal effects.

Researchers have used several methods to overcome these limitations. One effective approach is through collaborations with platforms. For example, in the US 2020 Election Project (2020EP), a team of external researchers worked with Meta to study Facebook's and Instagram's impact on attitudes related to the elections. Studies cooperating with platforms are often reliable because they provide access to rich internal data and have high external validity. Still, there is a risk in allowing platforms to study their own algorithms. Even if the platform's incentives do not affect the results of a study in any way, these incentives or constraints can affect the questions being asked (Lazer, 2015). Without access to internal data, studies observe the content that algorithms distribute by analyzing platform data that is publicly available or shared by participants (Hosseinmardi et al., 2021; Levy, 2021; Agan et al., 2023). To estimate causal effects, researchers have exploited variations in algorithms, including publicly announced changes or discontinuities in how posts are ranked (Ershov and Morales, 2024; Moehring, 2023). Other studies randomly expose participants to algorithmically curated content to estimate its effects, compared to counterfactual content (Holtz et al., 2020; Aridor et al., 2022).

The economics of algorithms are seemingly straightforward: Algorithms attempt to maximize the company's profits by increasing engagement. This problem is characterized in our framework as the platform choosing a set of posts to maximize the revenue-weighted time spent on the platform. Of course, social media platforms may

[17] There is a vast literature studying RS more broadly (Adomavicius and Tuzhilin, 2005). Social media RS are distinct relative to other typical RS environments since the set of potential items is evolving at a more rapid pace and consuming content is cheap (for example, a few seconds of time), so a large share of consumption is likely driven by recommendations. Furthermore, consumption externalities are stronger relative to other settings.

have other considerations when designing algorithms. For example, they may care about social welfare, and therefore, downrank hateful content even if it increases engagement. Still, the first-order goal is likely maximizing engagement. Indeed, platforms state that they attempt to find the most valuable content (Facebook and Instagram), increase retention and time spent on the platforms (TikTok), and give each potential post a score based on the probability of engagement (Twitter).[18] Algorithms mostly use signals that predict short-term engagement, such as whether a user would spend time on a post, click it, or share it. While short-term engagement is an easier object to maximize, a revenue-maximizing platform would probably focus on long-run engagement. Indeed, there is some evidence of platforms downranking content that may negatively affect users' long-run engagement, such as clickbaits (misleading or sensational headlines attempting to generate clicks).

There is strong evidence that algorithms substantially increase engagement and time spent on the platform. When a study in the 2020EP randomly switched participants for three months from an algorithmically curated feed to an RCO feed, the time users spent on the platforms decreased by 26% for Facebook and 13% for Instagram (Guess et al., 2023*b*). Participants were not explicitly told that they had been switched to an RCO feed and thus this paper arguably isolates the effect of the content itself from the effect of users perceiving highly-ranked posts as being "recommended" by the algorithm and worthy of their time. One limitation in experiments comparing algorithmically curated and RCO feeds is that the analysis does not take into account general equilibrium effects. Users may have chosen which pages to follow and whom to befriend on Facebook knowing that the algorithm would filter out irrelevant content. Still, even milder interventions changing the content promoted by algorithms decrease time spent, including the removal of toxic content (Beknazar-Yuzbashev et al., 2022) and reshared content (Guess et al., 2023*a*). These findings can explain why platforms oppose some attempts to regulate their distribution of content, as such interventions may have two costs: the direct cost of detecting and reprioritizing specific posts (for example, paying moderators) and the indirect costs due to lower engagement.[19]

If algorithms successfully increase time spent on platforms, the incentives of users and the platform may be partially aligned as a better algorithm results in more relevant content for users and higher revenue for platforms. While revealed preference logic may suggest that algorithms improve the user's experience, the literature has raised several concerns about potential algorithmic harms: Algorithms could promote content causing negative externalities, they may not only reflect consumer preferences but also shape them in dangerous ways, they may be biased toward specific content, and they may provide addictive content that increases engagement but does not increase the user's welfare. We discuss the first three concerns in this section and the last one in Section 4.

Low-Quality and Like-Minded Content. Even if algorithms perfectly maximize consumer's utility, they may generate negative externalities, which typically oc-

[18]For Facebook and Instagram see: `https://about.fb.com/news/2023/06/how-ai-ranks-content-on-facebook-and-instagram`; For TikTok see Smith (2021); for Twitter see: `https://blog.twitter.com/engineering/en_us/topics/open-source/2023/twitter-recommendation-algorithm`.

[19]While the takeaway from the literature is that existing algorithms increase time spent on the platform, we should not conclude that any deviation from the status quo would necessarily decrease engagement. For example, an experiment reducing the amount of content from like-minded sources on Facebook did not substantially decrease time spent (Nyhan et al., 2023).

cur outside the platform. For example, in order to maximize engagement, algorithms may promote low-quality or like-minded content, which could arguably distort beliefs and polarize users (Aral, 2021). How worried should we be about these concerns?

In terms of exposure to *like-minded content*, studies relying on Meta's internal data examine the effect of algorithms on exposure by comparing potential exposure to content (the set of posts users could potentially observe based on their friends, pages followed and groups joined) with actual exposure. They find that algorithmic curation contributes to segregation in news exposure (a result consistent with increased exposure to like-minded content), though the results are still being debated and the magnitude of the effect is probably not dramatic (González-Bailón et al., 2023; Bakshy, Messing and Adamic, 2015; Messing, 2023). As the authors note, one caveat is that the comparison group is defined as the set of potential posts, but the inventory of potential posts could also be affected by algorithms that suggest friends and pages to follow. Levy (2021) generates random variation in the pages people follow and also finds that Facebook's algorithm is much more likely to show content from like-minded pages, compared to cross-cutting pages.

In terms of content *quality*, González-Bailón et al. (2023) do not find differences in misinformation between potential and actual exposure, while Guess et al. (2023*b*) find that Facebook's algorithm almost doubles the amount of uncivil content or content containing slur words in the feed but also decreases content from untrustworthy accounts by approximately 40%. Moehring (2023) uses data from Reddit to train an RS and finds evidence suggesting that Reddit's algorithm has heterogeneous effects depending on users' demand for quality and increases the exposure of some users to low-quality publishers.

Overall, these results are consistent with concerns that the content promoted by algorithms may have negative consequences. However, the results are not dramatic or unequivocal. For example, Facebook seems to down-rank untrustworthy accounts, perhaps due to incentives that are not related to short-run engagement.

Rabbit Holes. A second concern regarding algorithms is that they not only reflect preferences for content within the platform but also shape preferences by gradually showing users more extreme content. In social media, users may go down "rabbit holes," i.e., dive deeper and deeper into particular topics. When these rabbit holes expose users to more extreme content, they may gradually develop more extreme opinions or incorrect beliefs. The concern over rabbit holes is similar to the concern over exposure to like-minded content but with several important distinctions. First, rabbit holes are dynamic, with individuals exposed to more extreme content over time (Brown et al., 2022). Second, typically the concern associated with rabbit holes is the radicalization of a small group of users, while the concern associated with exposure to like-minded content is a broad increase in polarization. Finally, rabbit holes have been mostly studied on YouTube since it has been argued that its RS gradually offers more extreme content and radicalizes users (Tufekci, 2018).

Studies on YouTube have not found strong evidence for extreme rabbit holes. Hosseinmardi et al. (2021) observe the browsing sessions of over 300,000 Americans and do not find that videos become more extreme within sessions, suggesting that platform recommendations do not explain the exposure to extreme content. Chen et al. (2023) find that YouTube rarely recommends extremist videos to people who do not already sub-

scribe to these videos' channels. Finally, in an audit experiment, Brown et al. (2022) had participants watch a random video on YouTube and then click the second recommendation. While the recommendations may slightly shift users toward like-minded partisan content, they do not lead the average user toward extreme "rabbit holes."

Algorithmic Bias. A third concern is that algorithms produce discriminatory outcomes by prioritizing or downranking content associated with certain demographic groups. Feeds may be biased because the algorithm itself is inherently biased or because the training data used by the algorithm is biased (Rambachan et al., 2020). Biased training data may be especially common on social media, where people make quick decisions that are more likely to suffer from implicit or subconscious biases. Agan et al. (2023) find that Facebook's news feed algorithm shows people fewer posts from their outgroup (race in the United States, religion in India) than what they state they would like to see, but do not find such bias in Facebook's friend recommendations. As the authors explain, the news feed algorithm might amplify biases since it is based on rushed decisions, in contrast to the friend suggestion algorithm. Even if the algorithm reflects user bias without amplifying it, decision-makers may have a preference for equity (Rambachan et al., 2020), especially in political content.

Perhaps the most prominent example of concerns over algorithmic bias on social media is the argument that major platforms are biased against conservatives. However, this claim has not received strong empirical support. On Twitter, the mainstream political right enjoys higher algorithmic amplification than the mainstream political left (Huszár et al., 2022), and on YouTube, Brown et al. (2022) find that recommendations slightly nudge users toward more conservative content. Of course, such a bias does not have to be intentional and it might reflect the argument that "[r]ight-wing populism is always more engaging," as a Facebook executive told *Politico* (Thompson, 2020). An exception to these results is the finding that most of the URLs that Meta flags as misinformation (based on a third-party fact-checking program) are favored by a conservative audience (González-Bailón et al., 2023). Again, this does not necessarily reflect deliberate bias, and it may also reflect conservatives being exposed to more misinformation on Facebook and Instagram around the 2020 elections.

The evidence so far shows that algorithms affect what users are exposed to. Does the content promoted by algorithms affect the content users engage with? One could argue that individuals with strong preferences for specific content will find ways to consume it regardless of what the algorithm shows them (for example, by skipping irrelevant posts). Nevertheless, the evidence accumulating from various studies suggests that content consumption is often somewhat passive. When individuals see posts from specific sources more (or less) often due to changes in the algorithms, the sources they follow, or the platforms' user interface, they tend to engage with those sources more (or less) often as well (Levy, 2021; Ershov and Morales, 2024; Nyhan et al., 2023).

These results suggest that people are close to indifferent regarding the type of content they consume, that the search costs for new content are relatively high, or that people passively consume the content shown to them due to a default bias or other biases. Future research could examine what drives demand for social media content and when and to what extent content consumption is driven by algorithms.

To conclude, algorithms are not "neutral,"

they promote content that increases engagement, and in some cases may potentially have negative consequences. However, there is limited evidence for some of the concerns expressed regarding algorithms—social media platforms are not systematically biased against conservatives, YouTube's recommendations do not seem to be radicalizing users by driving them down extreme rabbit holes, and algorithms are not necessarily increasing exposure to misinformation, though they may be increasing exposure to like-minded and toxic content.

Future studies could attempt to unpack the algorithmic black box and uncover the different forces driving algorithmic decisions. For example, do algorithms limit misinformation because users are less likely to click it, because users who share misinformation tend to share content that generates less engagement, or because algorithms downrank misinformation, despite its potential popularity? In addition, future research could investigate alternative models for distributing social media content. Clearly, going back to RCO feeds is not viable, as platforms derive profit and users derive utility from algorithmically curated content. Still, by focusing on short-run engagement, current algorithms often ignore negative externalities and the users' long-term utility. An open question is how social media algorithms can optimally increase social welfare and what government incentives can encourage them to do so.

3.1.3. Case Study: Segregation of News on Social Media

The previous sections have shown that individuals are more likely to have like-minded friends on social media, who share articles they agree with, and that algorithms may moderately promote like-minded content. These findings have led to concerns that social media platforms are characterized by echo chambers, loosely defined as segregated environments where people are mostly exposed to like-minded opinions, and that such echo chambers could undermine democracy (Sunstein, 2017).

The concerns over echo chambers predate social media. Economists, political scientists, and communication researchers have long studied selective exposure, the tendency to prefer like-minded content (Stroud, 2008). In a seminal study, Gentzkow and Shapiro (2011) found that segregation in online news consumption is not dramatic and is not lower than in offline social networks. While social media has become an important source of news consumption since the paper was published, later studies also found that segregation in online news consumption is modest (Flaxman, Goel and Rao, 2016; Guess, 2021), though it may be increasing (Peterson, Goel and Iyengar, 2021). The limited segregation online and similarity to news consumption on traditional media could stem from consumer choice: While some individuals have a preference for like-minded content, many consume a large share of content from moderate mainstream sources, regardless of the medium. Based on these studies, scholars have explained that the concerns over echo chambers are overstated (Guess et al., 2018). However, as we discuss below, the results are nuanced and depend on the setting (all visits to online news, visits to news sites through social media, or exposure to posts on social media), on how segregation is defined, and on the population studied.

Even though segregation of online news consumption is moderate, papers consistently find that visits to news sites through social media are more segregated than visits through other channels (Flaxman, Goel and Rao, 2016; Peterson, Goel and Iyengar, 2021; González-Bailón et al., 2023). For example, Gentzkow and Shapiro (2011) and Levy (2021) calculate the isolation in online news consumption, a standard measure of segrega-

tion, and find an isolation index of 0.08 and 0.17, respectively. This means that the difference between the share of conservatives in news sites visited by conservatives and news sites visited by liberals is 8-17%, similar to the isolation index for national newspapers or face-to-face interactions in the workplace. However, for news consumed through Facebook, the isolation index increases to 0.25 (Levy, 2021), similar to face-to-face interaction with family members.

Social media segregation may not be a big cause for concern if most news is not consumed through social media. Across different browser add-ons and time periods, researchers find that only around 6-10% of visits to news sites come from social media clicks (Flaxman, Goel and Rao, 2016; Allcott and Gentzkow, 2017; Levy, 2021; Peterson, Goel and Iyengar, 2021). However, some people may still be disproportionally exposed to social media. For example, more Americans aged 18 to 29 say they get news through social media compared to any other medium (Shearer, 2021). Furthermore, estimates of social media referrals may be downward biased since researchers often do not observe mobile data and since they almost never observe websites visited within social media platforms' apps. Indeed, self-reported data suggests that social media are an important source for news consumption: In a 2023 survey conducted across 46 markets, more people said the main way they came across news online is through social media (30%), compared to directly accessing a news website or app (22%) (Newman et al., 2023). Future studies could focus on improving the measurement of social media news consumption.

While most studies focus on news sites visited through social media, understanding exposure to posts *within* the platform answers the core question discussed in this section: What content do social media platforms distribute to consumers? Furthermore, it is likely that a large share of the time individuals spend engaging with news is through exposure to posts in their feeds.[20] In a recent 2020EP paper, González-Bailón et al. (2023) analyze the news sites that over 200 million Americans were exposed to on Facebook and find that segregation on the platform is higher than previously thought. However, segregation in *exposure* to content is still lower than segregation in the content individuals *engage* with. This suggests that segregation in news consumption results both from social media features (the algorithm, social network, option to personalize one's feed by following specific accounts) and from users' behavior, conditional on the posts distributed to them.

A second point of debate is the definition of echo chambers. As noted previously, papers often measure segregation in news consumption based on an isolation measure. However, Levy and Razin (2019) explain that echo chambers consist of both 'chambers,' the increased exposure to like-minded individuals, and 'echo', the potential polarization that could occur in these chambers. Concerns over polarization are mostly relevant for certain types of segregation. For example, there is less concern if Republicans and Democrats are isolated because they consume different nonpolitical local news about their area and thus visit different websites.

Other studies on echo chambers focus specifically on exposure to like-minded content. Nyhan et al. (2023) find that content from like-minded sources is prevalent on

[20] In 2021 UK users spent 10 minutes per day on news sites, while they spent 71 minutes per day on social media (Ofcom, 2022). If we assume, based on the share of news content in the feed, that around 7% of time on social media is news-related (Nyhan et al., 2023), then for every two minutes users spend visiting top news sites, they spend one minute being exposed to news within social media platforms. An analysis of 2016-2018 US Comscore and Nielsen data results in almost the same 2:1 ratio of online news consumption and social media news exposure (Allen et al., 2020).

Facebook but is far from dominant. They show that for approximately half of Facebook users, at least 10% of the news content they are exposed to is from cross-cutting sources (one limitation in this literature is that posts are typically defined as like-minded or cross-cutting based on their source and not their content). Other papers have also found that while social media increases segregation, it also increases exposure to opposing perspectives (for example, Flaxman, Goel and Rao, 2016). Several explanations can account for this somewhat unintuitive finding: Social media may facilitate weak ties with people who share different perspectives, while algorithms could still promote like-minded content (Barberá, 2015), algorithm rankings may reflect the fact that individuals prefer like-minded content but do not avoid cross-cutting content (Garrett, 2009), and extreme content from both sides of the aisle may be amplified by algorithms if it increases engagement. Future work should disentangle these mechanisms.

In addition to the type of news consumption studied and the definition of echo chambers, the population studied matters as research finds heterogeneity in online segregation. González-Bailón et al. (2023) find that there are far more news sources to which conservatives are almost exclusively exposed on Facebook, relative to sources to which only liberals are exposed to. Consistently, Eady et al. (2019) find that liberals were less likely to follow conservative sources on Twitter compared to conservatives following liberal news. The binary distinction between conservatives and liberals can be misleading. Echo chambers likely exist to some degree among extreme conservatives who are probably the ones visiting the most extreme websites (Guess et al., 2018), while moderate conservatives may still be exposed to potentially more diverse news than liberals.

After making important progress in measuring segregation, the literature has started unpacking the forces contributing to segregation, about which the evidence is more limited. Existing research focuses on users' behavior (selective exposure), algorithms, and social networks. First, there is clear evidence that users prefer to engage with like-minded content and their behavior plays an important role in increasing segregation. For example, D'Amico and Tabellini (2022) show that Reddit users are more likely to comment on negative news about candidates from the opposing party. Second, as discussed in Section 3.1.2, algorithms may moderately increase exposure to like-minded news, partially supporting the notion of "filter bubbles," i.e., of algorithms filtering cross-cutting content or prioritizing like-minded content (Pariser, 2011). Third, ideological segregation is larger among posts shared by pages people follow compared to posts shared by their friends (González-Bailón et al., 2023; Levy, 2021), suggesting that social networks are not the main force increasing segregation. While there is substantial research on homophily in social networks (as discussed in Section 3.1.1), more research is needed on how users decide which pages to follow (for example, the accounts of media outlets or politicians), since those pages may be driving segregation.

To conclude, the literature so far provides several important insights: 1) overall, among all online news consumption, ideological segregation is not very high; 2) segregation is higher on social media compared to other online channels; 3) social media platforms seem to increase exposure and engagement with like-minded news, but may also provide exposure to diverse perspectives; 4) segregation is not symmetric; 5) segregation is more likely to be driven by pages or elite accounts followed than by friends.

3.2. Advertisements

Unlike organic content, advertisements shown to users are not explicitly selected by the platform but rather are determined through auctions among advertisers. Furthermore, while this content appears similar to organic content in user feeds, it is typically marked as sponsored and the production incentives differ from those described in Section 2.1 as advertisements are usually optimized for off-platform purchases.[21] In this section, we discuss what characterizes advertisements on social media, the value they generate for businesses and politicians, and the privacy concerns that they raise.

While targeting is very coarse in advertising on other media such as newspapers and television, social media platforms enable advertisers to microtarget: directly bid on consumer interests, demographics, or even individuals similar to their customer base (i.e., through "lookalike" audiences). Indeed, on social media platforms consumers explicitly provide their demographic information, contact information, and the types of content they are interested in (for example, pages/accounts they follow). The reliance on *explicit* data paired with behavioral data that users create through natural usage on and off the platform enables even more refined targeting not only compared to traditional advertising but also relative to online display advertising (for example, ads on third-party websites like the New York Times). Furthermore, targeting on these platforms does not require "omniscient" knowledge of whom to target, but rather is facilitated by delivery optimization that enables rapid learning of the right audience for advertisements.

[21] Advertisers can also directly pay other accounts to post seemingly organic content that promotes their brands, known as influencer marketing. We do not review the emerging literature on the effectiveness of this type of marketing and the relevant disclosure obligations (as an example, see Ershov and Mitchell Forthcoming).

3.2.1. Value of Social Media Advertising

> Half the money I spend on advertising is wasted; the trouble is I don't know which half
>
> *John Wanamaker (1838-1922)*

A long-standing empirical question is whether and which advertisements are effective. Interestingly, the answer to this question is not obvious—for instance, Blake, Nosko and Tadelis (2015) demonstrate using a large-scale experiment at eBay that paid search advertising for brand keywords has a negligible causal effect on sales. While advertising effectiveness has been analyzed across various media, the tracking enabled by social media, both within and outside the platform, increases the ability to measure its effectiveness.[22] The empirical problem of measuring ad effectiveness is to estimate the incremental effect: How many additional consumers would purchase a good that they would not have purchased without the ad? Even with better measurement tools, this is typically difficult due to the volatility of purchases, delayed effects, and multiple exposures (Lewis and Rao, 2015; Gordon, Moakler and Zettelmeyer, 2023).

The most comprehensive evidence we have for the broad effectiveness of social media advertising comes from a large-scale experiment conducted internally at Meta (Tadelis et al., 2023). The paper quantifies the returns to advertising for over 200,000 establishments and finds that on average each dollar spent on ads yields $3.31 in revenues. This finding indicates that social media advertising often works, but the paper documents significant heterogeneity in performance based on various measures of advertiser sophistication. Specifically, advertisers with more experience and advanced users

[22] The core advancement in measurement relative to existing online advertisements is the ability to have a more stable consumer identifier across time.

of targeting tools provided by Meta have larger advertising returns. Beyond helping businesses more efficiently match with consumers, better-targeted ads can also promote social causes. For example, Breza et al. (2021) and Athey et al. (2023*a*) show that personalized public health messaging on social media during the COVID-19 pandemic increased vaccination rates. In general, while on average consumers do not like being exposed to advertising, these studies provide some evidence that there can be gains for both consumers and firms.

The next question is *how* social media ads work and what are their equilibrium consequences in the downstream product market. Economic analysis of advertising (Bagwell, 2007) posits that advertising primarily works through the following channels: shifting beliefs through information (for example, product awareness, attribute information) or directly shifting consumer preferences (for example, increasing affinity to the brand).

Bergemann and Bonatti (2011) explore the role that targeting plays through the information channel (i.e., product awareness) by characterizing the equilibrium implications of increased targeting capabilities. Their model implies that increased targeting should lead to more consumer-product matches and the entry of smaller advertisers. This is consistent with social media advertising's purported role in the success of "direct-to-consumer" businesses that primarily acquire customers through targeted online media campaigns and with small businesses being relatively more reliant on social media advertisements (Werner, 2022).

Empirical research suggests several unique aspects of social media advertising where these mechanisms interact with the "social" aspect of social media. For instance, Lee, Hosanagar and Nair (2018) find that the content of advertisements plays a large role in effectiveness not only by revealing information about product attributes, but also by building brand affinity through on-platform engagement with "brand personality" content. Furthermore, Bakshy et al. (2012) and Huang et al. (2020) show that displaying the name of a friend who liked the advertising brand (i.e., a social cue) leads to an additional lift in click-through rates beyond targeting alone. Huang et al. (2020) find that this effect is most pronounced for "status goods," such as clothing and cars, indicating that this works through a mix of both information and brand affinity. However, more work is needed to understand the relative role of each of the different mechanisms for ad effectiveness and how they interact with both the increased targeting abilities and the social aspect of social media advertising.

While the literature has shown that social media ads can be effective and also influence the composition of advertisers, we still have little empirical evidence about the implications of social media advertising for downstream product markets. Research on the evolution of the macroeconomic product market finds that there has been an increase in product variety and consumption of "niche" products over the last 15 years (Neiman and Vavra, 2023). One possible explanation of this increase is that it is now less costly for firms that produce niche goods to find their target consumers due to more targeted advertising. Indeed, Baslandze et al. (2023) show that some of the increase in product variety is linked to the introduction of online display advertising. On the other hand, theoretical work highlights that enhanced targeting abilities do not necessarily always improve consumer welfare. For instance, Prat and Valletti (2022) highlight how an increase in social media platform concentration can lead to reduced entry in the product market, and Bonatti, Bergemann and Wu (2023) argue that the equilibrium effects of targeting can lead to ineffi-

cient allocations and increased prices. Overall, these results point to the need to better understand the broader macroeconomic implications of social media advertising and its welfare effects.

3.2.2. Consumer Data and Privacy Concerns

A critical component of targeting is tracking users across a wide range of websites and mobile phone applications to collect data on their behavior and observe whether they eventually purchase the advertised product. However, consumers possibly value their personal data (McClain et al., 2023) and want to have more control over which data is shared with which advertisers. Privacy concerns ambiguously factor into consumers' utility, u_i^c, as they get positive utility from control over their personal data, but negative utility from a more inefficient matching to products and posts. Consumer data can also influence the price advertisers are willing to pay for consumers' attention by increasing the likelihood that the consumer is a good match for the product, which shifts the value of α in our framework. Thus, the fundamental economic tension for policymakers is balancing platform profits via high-quality targeting and the value consumers place on their data. In this section, we discuss the relevant literature on the economics of privacy in the social media context.[23]

Several papers study the first component of the tradeoff: the value of consumer data for platforms and advertisers. There are broadly two types of data that can be used for targeting: on-platform (for example, product usage) and off-platform (for example, other visited websites) data. To quantify the value of off-platform data, Wernerfelt et al. (2022) run a large-scale experiment at Meta that experimentally restricts off-platform data for a subset of advertiser campaigns and finds that the average cost to acquire an additional consumer increases by 37% without it. Furthermore, they find that smaller advertisers benefit more from access to off-platform data compared to larger advertisers. This implies that privacy regulation targeting off-platform data can have anticompetitive consequences, a theme consistent with extant literature that highlights a further tension between privacy regulation and competition (Peukert et al., 2022; Johnson, Shriver and Goldberg, 2023). The ability to use off-platform data has been impacted by recent privacy regulations. For example, Apple's App Tracking Transparency (ATT) policy on iOS allows consumers to opt out of disclosing their phone's advertising identifiers to third-party firms and restricts the ability of social media platforms to use this data. Using a panel of online advertising performance and sales, Aridor et al. (2024) show that ATT had a large and negative impact on new consumer acquisition for Facebook-dependent advertisers, indicating that advertisers were unable to substitute for the targeting capabilities of social media platforms with other forms of advertising.

Studies have used survey-based, willingness-to-accept (WTA) and willingness-to-pay (WTP) measures to estimate the value of the other component of the tradeoff: consumer welfare gains from privacy. Prince and Wallsten (2022) find that in the United States, the average consumer would need to be paid $2.87 to allow Meta to share their friend network with third-party advertisers on the platform. Lin and Strulov-Shlain (2023) elicit incentive-compatible valuations for users' data. They find that users require more money to provide their friend network and posts, compared to their likes and profile. They also find that the distribution of privacy preferences is heavily right-skewed.

[23] We focus primarily on papers published since 2016 as the broader literature on the economics of privacy was recently summarized in Acquisti, Taylor and Wagman (2016).

Collis et al. (2021) inform consumers about Facebook's monetization of data and find that this treatment reduces the dispersion of consumer valuations. These studies provide some quantification of consumer valuation for their social media data, but highlight that consumers are uncertain about their valuations and the difficulties in measuring them via survey-based methods. Existing theoretical work emphasizes that data externalities (i.e., that a consumer's data teaches the firm something about other consumers) may depress valuations (Choi, Jeon and Kim, 2019; Acemoglu et al., 2022; Bergemann, Bonatti and Gan, 2022) and that there might be a large gap between stated and revealed preferences, known as the privacy paradox (Athey, Catalini and Tucker, 2017). Thus, one important direction for future work is to measure consumer privacy valuations using revealed-preference measures based on real platform behavior and others' sharing decisions.

Overall, most work on the consumer side has quantified the value of on-platform data, while the work on the platform and advertiser side has focused more on the role of off-platform data. Future work should more comprehensively quantify the value of different components and characterize the overall welfare effects of privacy regulation, taking into account advertisers, the platform, and consumers.[24]

3.2.3. Case Study: Political Advertising

One particular type of advertising that has received substantial interest is political advertising on social media. These ads have become an important component of campaigns and in the 2020 U.S. election cycle, 13% of all political spending was on Facebook and Google (Tech For Campaigns, 2021). While the political science literature has historically been interested in measuring the effectiveness of campaign advertising (Jacobson, 2015), it gained broader public interest in the social media context after the Cambridge Analytica scandal. The 2018 scandal focused on the collection of rich data on millions of Facebook users by the firm Cambridge Analytica mostly for political advertising. The scandal raised concerns that elections could be determined by manipulating voters with ads based on their psychological profiles (Wylie, 2019).

At its core, the mechanisms behind the effectiveness of political advertisements are similar to advertising more broadly. Political advertising primarily functions through either "direct" persuasion—shifting voter beliefs and subsequently their vote choice—or "indirect" persuasion—shifting the likelihood that a voter goes to the polls (Ridout and Franz, 2011). The political context has useful empirical aspects—individual turnout data and geographically aggregated vote shares are publicly available. Furthermore, partially driven by Meta's response to the Cambridge Analytica scandal, there is a comprehensive database of political ads that is available to researchers.

Using Facebook's political ads library and a database of television advertisements, Fowler et al. (2021) study how the content and composition of advertisers shift with the introduction of Facebook. Consistent with the earlier discussion of how targeting enables the entry of smaller advertisers, the authors find that "challenger" politicians with less funding and in more local races enter into the market due to the decreased cost of advertising to their smaller target market. Furthermore, the advertising content

[24] Aridor, Che and Salz (2023) highlight that the tension between consumer privacy and the data needed for targeting is not always zero-sum. They show that for an advertising intermediary in the online travel market, the European General Data Protection Regulation (GDPR) enabled consumers who value their personal data to opt out, but the remaining set of consumers was of higher value to advertisers, leading to a reduction in revenues, but not as steep as would be suggested by opt-out rates alone.

of social media campaigns shifts to be more partisan and focused on indirect persuasion of voters likely to vote for the candidate as opposed to direct persuasion. This pattern highlights a broader difference between political and non-political ads, which is that political ads include "attack" ads that try to dissuade voters sympathetic to other parties from going to the polls, and not only promotional ads (Ansolabehere et al., 1994).

A large focus of the literature has been on measuring the effectiveness of political ads at shifting electoral outcomes, using both campaign and individual-level experiments, primarily in national elections. Aggarwal et al. (2023) run a $8.9 million field experiment in the 2020 United States presidential election and find that the effect of their pro-Biden advertisements on turnout is negligible. Coppock, Green and Porter (2022) randomize Facebook and Instagram ads across zip codes in the 2018 midterms and do not find a statistically significant effect. They then combine their results with three other experiments in a Bayesian framework. While each study does not detect an effect, the posterior based on all the accumulated evidence finds a small and statistically significant effect on vote share. One contention is that these studies find small or null effects as they are measuring average treatment effects, but they may be masking some dimension of heterogeneity. Coppock, Hill and Vavreck (2020) run a comprehensive set of 59 survey experiments exploring various dimensions of heterogeneity to test this theory and consistently find small effects. While this evidence suggests that ads may not be very effective, platforms can still impact turnout through other channels, such as with "get out the vote" posts (Bond et al., 2012).

It is important to remember from the earlier discussion that advertising effects are typically small and are notoriously difficult to measure, especially in the political context where there is only a single outcome period (i.e., voting day). Indeed, that such a large amount of money is spent on digital advertising in national election campaigns is puzzling and deserves additional research since either these ads are more effective than current research indicates or researchers are wrongly inferring the objectives that campaigns pursue with these ads (for example, fundraising rather than voter persuasion). Furthermore, it may be that electoral outcomes are generally hard to change in national elections, as Kalla and Broockman (2018) show in a meta-analysis of empirical studies that the impact of campaign contact (including ads and more intensive contacts like canvassing) is negligible.

While digital political ads do not seem to have large effects in general national elections in the United States, there is evidence that they can be effective in other countries. For example, Enríquez et al. (2024) find that non-partisan Facebook ads in Mexico increased the vote share of less corrupt municipal incumbent parties. These ads were so effective that they even affected people who were not directly exposed to them. One direction for future work, even for studying these issues within the United States, is to better understand the impact of these types of ads in local elections, which is precisely where we may expect that social media advertising could have a larger impact.

4. Content Consumption

Content production and distribution determine the set of organic and paid content users are exposed to. As the framework illustrates, given posts served by the platform $\mathbf{x}_i$, consumer i allocates their time between using the platform t_i and other activities $\mathbf{a}_i$. These choices have implications for consumer welfare, as well as direct effects on payoffs to content producers (through views) and the platform (through advertising rev-

enue), and indirect effects on social welfare. In this section, we first discuss consumer choice and its implications for consumer welfare. Given that social media can affect off-platform behavior, consumer welfare might not fully capture social welfare. Therefore, we also review the societal implications of social media, including channels for aggregate impacts and four case studies on its political effects in democracies. Finally, in addition to consumption *on* and *off* platforms, we describe substitution patterns *across* platforms and their economic implications.

4.1. Consumer Choice and Welfare

In this section, we discuss economic forces that influence social media consumption and its implications for consumer welfare.

4.1.1. Consumer Choice

Social media platforms attract a vast and diverse user base, many of whom spend a significant amount of time engaging with these platforms. We highlight three economic forces examined in the literature—consumption spillovers, habit formation, and self-control problems—that differentiate the problem of consuming content on social media from the standard consumer's problem and have important implications for how we interpret welfare.

An important feature of social media is the presence of consumption spillovers. Social media may be characterized by positive network effects: As the network size increases, the marginal value individuals gain from consumption increases, leading to higher consumption. These network effects could be large, as they are driven not only by forces found in traditional media (for example, more consumers attract better content from producers), but also by interactions among consumers that further amplify these effects. Consumption by others increases the marginal utility of consumption both directly (through comments or other interactions on the platform) and indirectly through content distribution (as a larger user base allows the platform to collect data and improve the algorithm to make it more engaging). Eckles, Kizilcec and Bakshy (2016) provide evidence of positive network effects on Facebook. Using a randomized encouragement design, they find that feedback from peers on shared content increases engagement on the platform. Similarly, Mummalaneni, Yoganarasimhan and Pathak (2023) find that increased engagement from Twitter peers leads to more time spent and engagement on the platform.

A second force that influences social media consumption is habit formation, where utility from current consumption depends on past consumption choices.[25] Allcott, Gentzkow and Song (2022) provide empirical evidence that social media use is habit-forming in a large-scale randomized online experiment. When participants were given temporary financial incentives to reduce the use of a set of most commonly used social media apps, they not only reduced their usage during the incentive period, but also in subsequent weeks. This persistence is a hallmark prediction from models of habit formation, and has been observed in other experiments targeting individual apps within the bundle.[26] Allcott et al. (2020) found that participants incentivized to deactivate their Facebook continued to use it less even after the experiment ended. Similarly, Aridor (2023) finds a post-deactivation reduction in Instagram usage as well as suggestive evidence of a post-deactivation reduction in

[25]The magnitude of habit formation for social media consumption could be particularly large as it could arise from learning, network investments (past engagement increasing the strength of connections on the platform), improved content distribution (algorithms improving at curating content), and automaticity (notifications automatically drawing a user back to the app).

[26]The paper also provides evidence that people are well aware of habit formation, but interestingly they consume as if they are inattentive to it, consistent with substantial projection bias.

YouTube usage.

A related third force is preference inconsistency. With features like immediate feedback, infinite scrolling, and frequent notifications, some social media apps may be especially tempting, and users end up consuming differently from what they would ideally consume due to self-control problems. In a randomized encouragement design, Hoong (2021) finds that participants significantly reduced use after adopting a soft commitment device, providing evidence for self-control problems on Facebook (but not Instagram). In the aforementioned Allcott, Gentzkow and Song (2022) experiment, some participants randomly received a digital tool that allowed them to set voluntary, personalized daily time limits for individual apps. Participants reduced their social media usage when given access to this tool and they were willing to pay for commitment devices—two pieces of evidence that consumers have self-control problems, and that they are at least partially aware of them. The paper also quantifies the *magnitude* of habit formation and self-control problems (two central features of addiction) through a structural model. On average, around 50 minutes per day or 31% of social media use can be attributed to self-control problems magnified by habit formation.

Partly due to the personalized experiences platforms offer, consumer heterogeneity may be especially substantial in social media consumption. For instance, while self-control problems affect many people, they have a negligible impact on social media consumption for about a quarter of users (Allcott, Gentzkow and Song, 2022). In an experiment with online workers on a crowdsourcing platform, Marotta and Acquisti (2018) find significant variations across users in their adoption of a tool that blocks social media access. Moreover, there is substantial heterogeneity in how different social media apps are used, as highlighted in a survey in Aridor (2023).

There are several avenues for future research. First, given the complex and fast-evolving nature of social media consumption, descriptive evidence detailing consumption behavior would be valuable. While most existing work has concentrated on the time spent on social media, the content consumed and the nature of the engagement are also important. Time spent on active interaction with others and passive browsing may contribute differently to habit formation. Second, while existing research highlights self-control problems among American adults, it is policy-relevant to quantify the extent of these problems in the younger population. Finally, future work could look within platforms and quantify how different design features influence what and how users consume. For example, certain features (for example, content format or algorithms) may exacerbate self-control problems (Rosenquist, Morton and Weinstein, 2021). Defining the key product characteristics and quantifying their effect on consumer choice is an important step forward in understanding the welfare implications of consumption.

4.1.2. Consumer Welfare

Building on the discussion of consumer choice, we turn to a question with significant policy relevance: how consumption choices impact individual well-being and the effect of social media use on individual outcomes.

Consumer Surplus. Measuring consumer welfare is not straightforward. Standard measures of WTP (for example, the amount of money a user is willing to pay to keep using social media) are likely to underestimate welfare since users are not used to paying for social media (Brynjolfsson, Collis and Eggers, 2019). Another approach for calculating consumer

surplus involves using data on consumer time use and converting the value of time to monetary terms, as demonstrated in Brynjolfsson, Kim and Oh (Forthcoming). In this approach, the value of consuming social media is its opportunity cost—time spent that could have been used on work or other leisure activities, converted into monetary terms via the wage rate. However, the presence of self-control problems outlined in the previous section would imply that these revealed-preference measures overestimate consumer surplus. Moreover, valuation of social media can be influenced by behavioral biases such as anchoring (similar to valuation of privacy as in Section 3.2) and projection bias (Allcott et al., 2020). Even in the absence of behavioral biases, complementarity between time spent online and content that decreases utility could imply that *increases* in engagement do not reflect welfare *increases* (Beknazar-Yuzbashev, Jiménez Durán and Stalinski, Forthcoming).

Nevertheless, the literature has made significant progress. To address the challenge that consumers are unaccustomed to paying for social media use, the prevailing method of measuring consumer surplus involves choice experiments that elicit WTA to *stop* using social media (Brynjolfsson, Collis and Eggers, 2019). Using incentive-compatible procedures such as multiple price lists (MPL) or BDM (Andersen et al., 2008; Becker, DeGroot and Marschak, 1964), experiments elicit how much a participant needs to be paid to stop using social media for a predetermined duration.

These measures reveal that users of social media highly value its access.The median monthly value of Facebook, for example, ranges from around $50 per month in Brynjolfsson, Collis and Eggers (2019) to $100 in Allcott et al. (2020) and $160 in Mosquera et al. (2020) in experiments conducted in the U.S. in 2016-2018. In a large-scale incentivized online choice experiment on representative samples from across 13 countries, Brynjolfsson et al. (2023) provide more recent evidence on the valuation of a set of digital goods. They uncovered an overall median monthly value for Facebook of $31, ranging from $11 in Romania to $57 in Norway. There is substantial heterogeneity in welfare gains from consuming social media: Individuals in countries with lower income obtain disproportionately higher welfare gains from social media relative to their higher-income counterparts.

The various documented benefits users get from using social media could lead to these positive welfare measures. For example, social connectedness may lead to positive effects on labor market outcomes. Armona (2019) finds that access to Facebook for an additional year in college substantially increases average earnings (especially for women) and decreases income inequality within a cohort. This occurs due to strengthened social ties, where the alumni network provides support in the labor market. These findings align with what Rajkumar et al. (2022) find from experiments in the network of over 20 million people on LinkedIn over five years: Social media platforms facilitate employment opportunities through connections with weak ties. Further research is needed on whether and how social media affects schooling and other important economic outcomes, in both the short and the long term. If the long-term benefits of social media are more substantial than the short-term gains (for example, connections may be strengthened in the long term), then WTA measures based on disconnecting for a few weeks or months could underestimate the long-run value of social media.

Subjective Well-Being. An alternative measure of welfare, beyond calculating con-

sumer surplus, is to directly measure subjective well-being (SWB) and life satisfaction. While SWB measures might not fully capture what people aim to maximize in their decisions (Benjamin et al., 2012), they circumvent issues in choice-based methods where choices may be distorted by biased beliefs such as projection bias. A large and growing body of interdisciplinary research explores the relationship between social media and well-being (Valkenburg, 2022). The evidence is mixed, with some studies finding a negative correlation, while others find null results or a positive relationship.

Recent work has concentrated on offering causal estimates, often concluding that social media usage negatively impacts well-being and exacerbates symptoms of mental disorders in the general population, with adverse effects extending beyond those with pre-existing mental health conditions. Braghieri, Levy and Makarin (2022) leverage the staggered introduction of Facebook across university campuses and find that Facebook negatively affected mental health, specifically anxiety and depression-related symptoms. This study is notable for accounting for network effects since it studies the introduction of Facebook within entire communities. Experimental studies on later versions of the platform reach largely similar conclusions. Allcott et al. (2020) find that deactivating Facebook for a month led to significant improvements in SWB measures, including improved self-reported happiness, life satisfaction and reduced symptoms of depression and anxiety, shifting the overall SWB index by 0.09 SD. Mosquera et al. (2020) observed that a weeklong break from Facebook reduces feelings of depression (but not other measures of SWB). Allcott, Gentzkow and Song (2022) find that a reduction in smartphone social media use by around an hour per day improved self-reported concentration, though the effects on SWB are insignificant, potentially due to the smaller reduction in use compared to other studies. Outside of economics, deactivation studies (for example, Asimovic et al., 2021; Arceneaux et al., 2023) and experiments significantly reducing use (for example, Hunt et al., 2018; Brailovskaia et al., 2020) generally show a small negative impact of social media use on well-being, with recent psychological research noting the importance of understanding use beyond total time spent in examining this relationship (Kross et al., 2021).

These negative impacts on mental health and concentration could adversely affect an individual's economic outcomes. Marotta and Acquisti (2018) show that blocking access to Facebook and YouTube increases productivity and earnings because it reduces distractions. Braghieri, Levy and Makarin (2022) find that students were more likely to report impairment in academic performance due to depression-related symptoms after Facebook was introduced in their college.

An area of particular public concern, as emphasized by the U.S. Surgeon General (Surgeon General, 2023), has been the risks of social media for children and adolescents. These younger groups differ developmentally from adults and could benefit more from the connections fostered through social media or suffer more from social comparisons that these platforms facilitate. Studies on the effect of internet use suggest that these concerns are not unwarranted. Adverse effects have been observed even with limited social media use: Donati et al. (2022) leverage the introduction of high-speed internet in Italy and provide quasi-experimental evidence for the internet's effect on increased mental disorder diagnoses for children and teenagers. These effects could be further exacerbated by social media. McDool et al. (2020) find that faster internet in the UK between 2012-

2017 is associated with children feeling worse about their appearance. Both papers find heterogeneity by gender, with worse effects for girls. Given the distinctive features of social media consumption, future research focusing specifically on social media use in recent years is needed to understand its effects on adolescent well-being.

The widespread concern over the impact on well-being has led to a proliferation of tools and interventions aimed at regulating social media use. The 2023 U.S. Surgeon General advisory underscores the urgent need for research that evaluates the effectiveness of such programs, policies, design features, and interventions, particularly those targeted at younger populations. For example, digital citizenship curriculums have been developed and implemented in some schools to teach students how to responsibly use social media and other digital technologies, but their effectiveness remains to be evaluated (Weinstein and James, 2022). Rigorous evaluation of these interventions is crucial for informing policy decisions. Moreover, the mechanisms through which social media use affects outcomes are not well understood. There is suggestive evidence that social comparisons and a fear of missing out are important channels through which social media affect well-being (Braghieri, Levy and Makarin, 2022; Bursztyn et al., 2023*a*). Other hypothesized channels include displacement or distractions from other activities, disruption of sleep, and effects on social connectedness. An avenue for future research is to measure the magnitude of channels through which social media impacts well-being and their relationship to platform design features.

Overall Welfare Impact. What is the overall net welfare impact of social media? Existing evidence presents an apparent paradox. On the one hand, individuals require large payments to stop using social media. On the other hand, there is some evidence that it negatively impacts well-being and mental health.

While these contrasting findings could stem from the inherent measurement challenges mentioned above, another explanation, based on the social component of social media, has been proposed by Bursztyn et al. (2023*a*). The paper argues that when non-users derive negative utility from others' social media usage—driven by, for example, a fear of missing out—the standard measure of WTA to individually deactivate social media overestimates welfare. Additionally, users could be "trapped," finding it individually optimal to use social media, even if they would prefer to coordinate with others to stop using it or to reduce their consumption. Using incentivized online experiments with college students, the authors find individual welfare estimates consistent with the literature. However, after accounting for non-user utility, individual welfare turns negative for 60% of TikTok users and for 46% of Instagram users. This evidence suggests that a large fraction of individuals could be using social media, while still deriving negative welfare from it, because the cost of being individually excluded from it is high.

These results do not imply that consumer welfare is negative at every level of social media consumption. The economic literature measuring WTA and mental health has largely focused on extensive-margin measures that shut down entirely or give access to social media. Unlike interventions for other addictive goods, such as cigarettes, strategies for managing social media often focus on modifying behavior on the intensive rather than extensive margin. This suggests that some level of social media use may be welfare-improving. An open question for future work is estimating the effect of the

intensity of social media usage on welfare.

4.2. Societal Implications

Social media's influence extends beyond individual users and their networks, impacting the broader economy and society as a whole. In this section, we discuss the channels through which social media has aggregate impacts and we present case studies on its political effects in democratic systems.

4.2.1. Channels for Aggregate Impacts

The consumption of social media can significantly influence economic and societal outcomes by shaping individuals' beliefs and preferences as well as the way people interact with each other. There are many channels through which these effects may manifest themselves. We broadly categorize them into social media: 1) providing exposure to persuasive content; 2) facilitating coordination of actions; and 3) shifting individuals' perceptions of others.

Social media can affect beliefs and preferences by exposing users to various forms of persuasive content, including information, misinformation, and noninformational materials, such as entertainment. Some of this content is produced with the intention of persuading consumers. One example of persuasive communication is social media advertising, discussed in Section 3.2. Another example is experts using social media to disseminate information and influence public opinion. Ehrmann and Wabitsch (2022) show that Twitter provides a channel for the European Central Bank to relay information to non-experts, leading to more factual tweets by non-experts with more moderate and homogeneous views.

Given the nature of content production, especially the low barriers to entry, the persuasive effect of social media could be different than traditional media. For example, social media can be used to voice concerns and ultimately enhance accountability. Gans, Goldfarb and Lederman (2021) show that consumers use Twitter to more effectively voice quality concerns to airlines. In an analysis across subnational regions in 116 countries, Guriev, Melnikov and Zhuravskaya (2021) investigate the impact of the worldwide expansion of 3G mobile networks, a key driver for the expansion of social media, on government approvals and find that access to 3G reduced government approval in areas with some level of corruption and uncensored internet. Moreover, the effect is particularly pronounced in areas where traditional media are censored, suggesting that social media can help promote accountability.[27] Similarly, Enikolopov, Petrova and Sonin (2018) show that blog posts (an early form of social media) exposing corruption in large state-controlled Russian firms affected the stock market and were associated with management turnover and long-term improvements in corporate governance.

Some social media content, such as posts that express personal opinions, may be produced for other reasons (see Section 2.1) and not necessarily with the intent to persuade, but can still influence the beliefs and preferences of those who consume it. For example, social media has been shown to affect market expectations. Bianchi et al. (2023) find that tweets from Trump criticizing the Federal Reserve affect expectations about future monetary policy, and consequently financial markets. Similarly focusing on Twitter, Bianchi, Cram and Kung (Forthcoming) provide evidence that tweets by members of

[27]The paper also finds that the expansion of 3G in democracies reduced votes for the incumbent government and benefited both right-wing and left-wing populist opposition parties. Relatedly, across twenty European countries, Tabellini, Manacorda and Tesei (2023) find that increased access to mobile internet was associated with a higher vote share for extreme right-wing and communitarian parties. The authors explain that the results are consistent with social media making individuals more easily persuaded by messages of intolerance of outgroups.

Congress influence stock prices through their influence on expectations about future legislative and economic action. Beyond communication from politicians, information-sharing on social media among investors can improve the accuracy of their expectations about monetary policy, especially under uncertainty (Ehrmann and Hubert, 2023), and can improve short-term forecasts (Dessaint, Foucault and Frésard, 2021).[28] In this context, social media facilitates the creation of common knowledge, as the exchange of information shapes not only users' expectations about future policies, but also their view of the expectations of other market participants.

The remaining two channels through which social media influences outcomes are linked to its inherent *social* nature. Social media allows users to coordinate their actions by reducing costs for groups to form and exchange information on organization and tactics. This has been shown to facilitate protests and social movements. A consequential example is the Arab Spring, where offline protests were associated with coordination on social media (Acemoglu, Hassan and Tahoun, 2018; Steinert-Threlkeld et al., 2015). Separately, Enikolopov, Makarin and Petrova (2020) exploit the diffusion of the Russian social media platform VKontakte (VK) to the cities of origin of students who studied with the founder of VK. The paper finds that social media increases protests substantially. This effect is likely driven by the platform's capacity to facilitate tactical coordination, as VK served as a host for the majority of online protest groups.

Social media can also affect the perceptions people form about others' beliefs or behavior. It could have notable implications for political outcomes by affecting social norms, social pressure, or social image concerns. By changing perceptions, social media could increase protests, even under censorship and without explicit coordination. Qin, Strömberg and Wu (2021) show that social media expands the scope of protests in China through its influence on users' beliefs about others' participation. Since users know that social media can affect how others perceive them, social image concerns may drive their behavior on social media. Using data on Russian political protests in 2011-2012 and a survey of protest participants, Enikolopov et al. (2023) show the importance of social image concerns as another driver of protests. Social media amplifies the significance of these concerns because it enables users to signal to larger groups.

Understanding the relevance of different channels through which social media can affect beliefs and behavior has important policy implications. For example, Bursztyn et al. (2019) demonstrate that in Russia, social media contributes to an increase in ethnic hate crimes, by increasing coordination among perpetrators and changing people's attitudes, but it does not reduce (and in fact increases) the perceived stigma associated with xenophobia. Therefore, in this context, interventions targeting hate crime reduction should focus on the persuasion or coordination channels, rather than social image concerns.

This section outlines the various channels through which social media affects economic and societal outcomes. Beyond identifying relevant channels in specific contexts, future research could quantify the magnitude of their impacts. Another area to explore is the emerging role of generative artificial intelligence on social media platforms. As it becomes increasingly difficult to distinguish between AI-generated and human-created content, the dynamics of user communication on these platforms and its subsequent economic and societal impacts may evolve. For

[28] See Cookson, Mullins and Niessner (2024) for a review on social media and finance.

instance, the value of signaling might diminish if users become aware they are not interacting with other humans.

4.2.2. Case Studies: Political Impacts in Democracies

Through the channels outlined above, social media can have broad political effects. We focus on four outcomes as case studies: misinformation and political knowledge, polarization, political participation, and offline violence. For more comprehensive reviews on political outcomes, see Zhuravskaya, Petrova and Enikolopov (2020) and Lorenz-Spreen et al. (2023).[29]

Misinformation and Political Knowledge. The proliferation and potentially persuasive impact of misinformation on social media have garnered considerable concern and public scrutiny. These concerns are not unwarranted. Around half of the users exposed to fake news on social media report believing it (Allcott and Gentzkow, 2017). Furthermore, misinformation disseminated by politicians has been shown to increase their support (Barrera et al., 2020).

A large and growing literature studies interventions to combat online misinformation on the consumption side. These interventions resemble those aimed at deterring the sharing of harmful content discussed in Section 2.2, but their objectives differ as they seek to influence beliefs, as opposed to affecting sharing decisions.

One popular intervention is fact-checking or debunking. While there is some evidence of a backfire or null effect of fact-checking interventions (Nyhan and Reifler, 2010; Batista Pereira et al., 2022), the vast majority of evidence suggests that they mitigate the impact of misinformation on individuals' beliefs (Walter et al., 2020). However, their effectiveness might be short-lived (Nyhan, 2021) and confined to the specifically debunked content (Berger et al., 2023). Furthermore, despite their influence on beliefs, fact-checks can be ineffective at influencing actual attitudes (Barrera et al., 2020; Nyhan, 2020; Nyhan et al., 2020). Similar to interventions on the production side described in Section 2.2, these consumption-side interventions can also have unintended consequences such as increasing the credibility of untagged information (Pennycook et al., 2020).

Beyond fact-checking, light-touch media literacy interventions, such as exposure to tips to spot fake news, have also been shown to reduce the credibility of misinformation (Guess et al., 2020; Berger et al., 2023), in addition to reducing sharing as discussed in Section 2.2. Moreover, Berger et al. (2023) show that they can be more effective than fact-checking in enhancing truth discernment between fake and factual content, with impacts persisting two weeks after the intervention. These interventions can be especially important on social media when individuals are exposed to a stream of unvetted information from many sources.

A related literature studies more broadly how social media affects political knowledge. Since social media often exposes individuals to news, it is perhaps not surprising that deactivating Facebook decreases news knowledge (Allcott et al., 2020). Which Facebook features contribute to this phenomenon? Individuals may get exposed to news shared by their friends on social media. However, friends tend to share like-minded news (as discussed in Section 3.1.1) so it is not clear if such news will increase knowledge. Indeed, researchers studying this question

[29] Zhuravskaya, Petrova and Enikolopov (2020) review the political effects of the Internet and social media, including in autocracies. Lorenz-Spreen et al. (2023) review the causal and correlational evidence on the relationship between digital media use and political outcomes in democracies. In addition, two recent books also discuss the political impact of social media: Persily and Tucker (2020) cover social media and democracy, and Campante, Durante and Tesei (2023) cover the political economy of social media.

in a lab environment found evidence that sharing like-minded news results in less informed users (Pogorelskiy and Shum, 2019; An, Quercia and Crowcroft, 2014). However, a 2020EP experiment found the opposite effect: Completely removing reshared content from participants' feeds decreased political knowledge (Guess et al., 2023*a*). The inconsistency can be explained by the type of selection happening. While participants are more likely to share like-minded content compared to cross-cutting content, they are also more likely to share political content generally. Removing reshares decreases the exposure to political content and that could explain why knowledge decreased.

Future research could explore how platform features, beyond sharing, influence political knowledge and affect the effectiveness of interventions designed to counter misinformation. For example, Carney (2022) finds that political WhatsApp groups in India increased discernment in environments that allowed peer-to-peer interactions, compared to those where users only received messages from a political party without the ability to engage with each other.

Polarization. As discussed in Section 3.1, social media platforms can amplify engagement with like-minded content, and, concurrently, expose users to diverse perspectives. What is the overall impact of this on the polarization of beliefs and attitudes?

Evidence from the U.S. suggests that social media may have increased polarization but its impact is relatively modest. Allcott et al. (2020) find that disconnecting from Facebook for one month reduced political polarization. Leveraging quasi-experimental variation in 3G internet access, Melnikov (2021) finds that mobile applications, including social media, contributed to polarized political views and support for specific candidates and policies. While social media may increase polarization, based on trends in polarization across demographics, the rise in polarization is unlikely to be primarily driven by social media consumption (Boxell, Gentzkow and Shapiro, 2017).

Outside the U.S., the evidence is mixed. In India, exposure to party messaging on WhatsApp had no effect on affective polarization (Carney, 2022). In France, a 3-week break from Facebook during the 2022 presidential election did not affect political or affective polarization (Arceneaux et al., 2023). In contrast, deactivating Facebook in Bosnia and Herzegovina during a period of heightened attention to past conflicts, actually increased ethnic polarization (Asimovic et al., 2021). The effect is almost entirely driven by individuals residing in more ethnically homogeneous areas. The same experiment was repeated in Cyprus in Asimovic, Nagler and Tucker (2023), and the authors found a null effect and posited that this may be driven by limited accessibility to the outgroup online due to language barriers. These mixed findings across different contexts could be driven by different relative levels of segregation on social media compared to offline interactions. In environments where social media offers greater exposure to outgroups than what is experienced offline, it might help reduce polarization and improve attitudes towards outgroups. Conversely, in scenarios where online and offline interactions show similar levels of segregation, or where offline environments are less segregated, social media could either have no impact or potentially exacerbate polarization.

Several studies have examined the impact of specific platform design features on polarization and have found null effects. In particular, reshares have no effects on issue or affective polarization, nor on any other measure of political attitudes (Guess et al., 2023*a*). Furthermore, in another 2020EP experiment, Nyhan et al. (2023) find that a

reduction in exposure to like-minded content has no effect on affective polarization. Finally, Liu et al. (2023) run an experiment with an online interface that resembles YouTube and find that more extreme videos do not increase polarization.

One often suggested intervention to break filter bubbles is exposing users to content they may not see otherwise (Sunstein, 2017). Levy (2021) shows that affective polarization can be reduced in an experiment on Facebook, while Bail et al. (2018) and Di Tella, Gálvez and Schargrodsky (2021) find that exposure to counter-attitudinal content on Twitter leads to higher polarization. One explanation for these different findings is the set of compliers. Levy (2021) nudged people to break filter bubbles but did not require them to do so, while Bail et al. (2018) incentivized compliance. It is possible that individuals who are willing to break filter bubbles can become less polarized when doing so, while individuals who are averse to cross-cutting information experience a backlash effect. Song (2023) presents evidence supporting the heterogeneous effects of social media content through a survey experiment that exposed users to racial justice content from Twitter. Highlighting the role of ideological distance, the paper shows that counter-attitudinal content could be effective when it is not too far or too close to the audience's preexisting beliefs.

Another way to break filter bubbles is to replace algorithmic curation of content with an RCO feed. Since individuals are more likely to click on links distributed by algorithms (see Section 3.1.2), switching the feed may also affect their beliefs and behavior. Interestingly, the 2020EP study replacing Facebook's feed with the RCO feed did not find an effect on issue polarization or affective polarization. This is one of the strongest pieces of evidence on the real-world effects of algorithms as it was conducted in an actual social media platform, with rich data, around an important event—the 2020 U.S. presidential elections. Perhaps because of this, the null results could stem from participants' relatively strong priors and more careful moderation by Facebook. More research is needed to study how algorithms affect political beliefs and attitudes in other contexts.

Political Participation. Social media has been shown to facilitate protests—one form of political participation—at various stages, from their mobilization and coordination to their long-term effect on individual behavior. Beyond the aforementioned evidence from nondemocratic regimes, social media plays a role in the amplification of social and political movements in democratic countries as well. Examples include the MeToo movement, the 2020 Black Lives Matter protests (the largest protests in U.S. history to date), and the 15M movement in Spain (Casanueva Artís, 2023; Casanueva Artís et al., 2023; Levy and Mattsson, 2023). Across movements, Gylfason (2023) shows that Twitter plays a role in facilitating protests in the U.S., particularly those associated with extreme movements. In an analysis of protests across countries (including both autocratic and democratic regimes), Fergusson and Molina (2021) exploit the expansion of Facebook across languages. The title of their paper succinctly summarizes the main conclusion: *Facebook Causes Protests*. The authors find heterogeneous effects by the level of democracy with a U-shaped pattern: The effects of Facebook on protests are largest at low or high levels of democracy.

The evidence on the effect of social media on voting is more mixed. The 2020EP studies have found precisely estimated null effects on voters (including self-reported political participation and turnout) (Guess et al.,

2023*a*,*b*). In contrast, research in other countries and platforms found that social media content affects voting decisions. In Mexico, social media ads informing voters of municipal expenditure irregularities had large and heterogeneous effects on how people voted (Enríquez et al., 2024). In Colombia, Garbiras-Díaz and Montenegro (2022) show that a social media campaign increased reporting of electoral irregularities by facilitating information transmission. This consequently enhanced electoral integrity, leading to a decrease in irregularities from candidates and a lower vote share for those who relied on such irregularities. In the U.S., Fujiwara, Müller and Schwarz (Forthcoming) provide evidence that the prevailing liberal content on Twitter may have swayed moderate voters to vote against Donald Trump.

Offline Violence. There is convincing evidence that social media use—particularly the exposure to toxic content—can lead to offline hate crimes. In line with the aforementioned evidence for Russia (Bursztyn et al., 2019), similar results have been observed in democratic countries. In the U.S., Trump's tweets have been linked to an uptick in anti-Muslim hate crimes (Müller and Schwarz, 2023) and anti-Asian incidents (Cao, Lindo and Zhong, 2023). Beyond the influence of high-profile individuals, hate speech by users with extreme viewpoints has also been linked to attacks on refugees in Germany (Müller and Schwarz, 2021).

There is also some evidence that government regulation akin to a Pigovian tax can mitigate this externality. Jiménez Durán, Müller and Schwarz (2022) analyze the effect of Germany's Network Enforcement Act, which introduced penalties for large platforms that fail to promptly remove hate speech and induced more content moderation efforts. Exploiting the differential exposure of Germany's municipalities to toxic content prior to the policy, the paper documents that the regulation decreased hate crimes. Additional research is needed to study whether these policies have unintended consequences such as the silencing of political dissidents. Lastly, future research could explore how content moderation affects other offline harmful actions besides violence (for example, self-harm), given recent theoretical work that shows that content moderation may improve welfare when it blocks information that enables harmful acts (Kominers and Shapiro, 2024).

4.3. Consumption Across Platforms

Apart from the study of the behaviors that occur *on* platforms, there is a separate strand of research studying competition *across* platforms. In recent years, this area has gained policy relevance amid concerns that the market for social media applications has become too concentrated. The FTC has a monopolization lawsuit against Meta (FTC, 2021) and the competitiveness of this market has been vigorously debated (Scott Morton et al., 2019). In this section, we focus on one dimension of this broader issue: consumer substitution across platforms and its relevance to antitrust concerns.[30]

We consider that social media platforms compete for consumers on both "prices" and quality, but only focus on the price dimensions here. The definition of price in this context requires some nuance, as these services are typically offered at a zero monetary price. While these services are free, the literature typically models the relevant price as the *advertising load*, or the number of paid advertisements as a fraction of observed content, set by the platform. This modeling idea dates back to the seminal paper of Anderson and Coate (2005) and considers it as an im-

[30]We do not focus on the competition for advertisers across platforms, but note that Gentzkow et al. (2024) empirically find that advertisers' willingness to pay in equilibrium is dependent on consumer substitution patterns across platforms.

plicit cost on consumer time.

The interpretation of price competition in this context has been first-order to antitrust debates regarding social media applications. In a typical antitrust investigation, regulators aim to take actions that maximize consumer welfare. Lacking direct estimates of consumer welfare, regulators resort to studying price effects in a "relevant market" of the closest set of substitutes. The naive intuition that social media products are free and hence there is no price competition led regulators in the Facebook-Instagram and Facebook-WhatsApp merger evaluation to focus primarily on whether other applications provided similar functionalities, and not on measurements of demand with respect to time spent (Argentesi et al., 2021). Even when considering that these applications compete for consumer time, there is substantial disagreement between regulators and Meta about what qualifies as substitutes for social media applications (FTC, 2021).

With the provided interpretation of price competition, the relevant substitution patterns are with respect to changes in the advertising load, though measuring this parameter has proved challenging. Aridor (2023) provides evidence on the substitution patterns for YouTube and Instagram by having participants install software that enables restriction of applications on their phones and characterizes substitution patterns at the "choke" advertising load (i.e., a sufficiently high ad load such that no one consumes the product). This design provides a conservative estimate for the relevant set of substitutes. Aridor (2023) finds that when Instagram is restricted, users substitute not only to other social media applications, but also to communication applications, such as WhatsApp, and that when YouTube is restricted, its users substitute to social applications. Furthermore, participants state that they use different social media applications for different purposes that overlap with nonsocial media applications and can partially explain the cross-category substitution. Other papers also find substitution from social media to communication applications (Collis and Eggers, 2022; Agarwal, Ananthakrishnan and Tucker, 2022). However, Aridor (2023) also finds that there is a large amount of substitution to non-digital activities.

Thus, the conclusion from these papers is that characterizing substitution patterns for social media platforms requires careful empirical examination of what consumers use each platform for as consumers' content on each platform is personalized. As a result, substitutes for social media platforms may overlap with applications not traditionally considered social media. Nonetheless, the large substitution to non-digital activities also indicates that these applications hold significant power over consumer's time. As such, an important avenue for future work is determining how these nuanced substitution patterns play a role in determining relevant markets and subsequently characterizing the market power of these applications.

There are several interesting directions for future work. The first is that given the large informational externalities from consumption discussed in Section 4.2.1, an unexplored question is not only to measure market power in terms of time spent, but also to think of media power as Prat (2018) does for traditional media. Of particular interest is understanding whether social media increases or decreases the media power of existing large media organizations. The second is to explore the implications of habit formation (discussed in Section 4.1) for competition among social media platforms. Furthermore, self-control and other problems mentioned in Section 4.1 indicate that time use is not a good proxy for welfare, which also implies that antitrust tests based on "price"

effects may no longer be good proxies for welfare (Rosenquist, Morton and Weinstein, 2021). Thus, future work should explore the dynamic supply-side implications of these aspects of demand.

5. *Conclusions*

In this guide, we have synthesized the existing literature based on the life cycle of social media content. This section briefly discusses possible future research directions.

The empirical literature uses a variety of methods to study social media, including descriptive studies on the media landscape, quasi-experimental designs exploiting the expansion of new technologies or other exogenous variation in access to social media (Sabatini, 2023), and field experiments (Aridor et al., Forthcoming). In particular, field experiments have become more prominent in recent years due to the strong causal identification they provide, the options to randomize features or content at the individual level, and the rich data available. While these experiments have moved the literature forward, they also come with a set of limitations: they typically measure short-run effects, study partial-equilibrium outcomes, and are often limited to individuals who are willing to participate and platforms where experimenting is possible. Future studies can attempt to overcome these obstacles by analyzing complementary observational data, conducting experiments over longer time periods, studying general equilibrium effects either in a more controlled setting or by studying counterfactuals using structural models, and by analyzing whether treatment effects tend to differ for participants who are less willing to participate in experiments.

While this guide demonstrates that social media has received ample attention, it is constantly changing both in terms of the platforms used and the content produced, distributed, and consumed. Facebook remains the most dominant platform (see Figure 1b), but it faces competition from newer platforms such as TikTok, which already has over one billion users. As social media are constantly evolving, more research is needed about other platforms that are growing in usage and which have been underrepresented in the literature.

Studying other platforms is important not only for external validity. As discussed in Section 3.1, the shift toward new platforms, such as TikTok, reflects a transition from content personalized by users (for example, choosing which accounts to follow) to content that is completely algorithmically curated. Furthermore, this pattern also reflects a coinciding transition in the type of content that gets shared: from primarily text on Facebook and Twitter to primarily photos and videos on Instagram and TikTok. Future work should study the economics behind this transition: Does the emphasis on algorithmically curated videos reflect a technological shock that has made platforms substantially better at producing real-time recommendations? Or does this change reflect a maturing segmented social media market where different platforms offer different algorithms? This transition also raises questions about the downstream implications of the new content consumed: Does the declining importance of social connections affect labor markets and well-being (Section 4.1.2)? Does this change entail an increase of entertainment at the expense of news on social media, diminishing the magnitude of off-platform political effects (Section 4.2.2)?

Beyond shifts in the type and distribution of content, the business models of social media platforms have also begun to change. Facebook and Twitter now offer users the ability to pay for ad-free versions and decentralized, ad-free, platforms

such as Mastodon have grown in popularity. The economic implications of this transition are ripe questions for future work, especially in light of the concern discussed in all three sections, namely that optimizing for engagement—as in advertising-based business models—does not always coincide with optimizing for user utility or social welfare. Will the rise in subscription-based business models decrease the prevalence of harmful content (Section 2.2)? Will it lead to platform changes that reduce the negative effects on mental health (Section 4.1.2)? What will the effects be on small businesses that benefit from social media advertising (Section 3.2)?

Some of the recent changes on social media platforms are a response to government regulations. Indeed, as social media are growing in importance and regulatory policy takes shape, one fruitful direction for future work is to study and inform policy debates. While our guide does not center on specific policies, the broader economics we cover is relevant for the evaluation and design of regulations targeting social media. For instance, the debate over Section 230 in the United States regarding whether platforms should be held liable for their content is informed by the economics of content moderation (Section 2.2), algorithmic distribution (Section 3.1.2), and the platforms' market power (Section 4.3). Furthermore, for EU regulations such as the DSA and the GDPR which provide users with more control over content personalization and their shared data, it is crucial to understand the economic tension between the value of consumer privacy and targeted advertising (Section 3.2) as well as the externalities associated with personalized content (Section 3.1.3). As this discussion illustrates, while social media has continued to evolve, one thing that has not changed is that social media remains a central part of people's lives.

REFERENCES

Abou El-Komboz, Lena, Anna Kerkhof, and Johannes Loh. 2023. "Platform partnership programs and content supply: Evidence from the YouTube 'Adpocalypse'."

Abreu, Luis, and Doh-Shin Jeon. 2020. "Homophily in social media and news polarization."

Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asu Ozdaglar. 2022. "Too much data: Prices and inefficiencies in data markets." *American Economic Journal: Microeconomics*, 14(4): 218–256.

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius. Forthcoming. "A model of online misinformation." *Review of Economic Studies*.

Acemoglu, Daron, Tarek A Hassan, and Ahmed Tahoun. 2018. "The power of the street: Evidence from Egypt's Arab Spring." *The Review of Financial Studies*, 31(1): 1–42.

Acquisti, Alessandro, Curtis Taylor, and Liad Wagman. 2016. "The economics of privacy." *Journal of Economic Literature*, 54(2): 442–492.

Adomavicius, Gediminas, and Alexander Tuzhilin. 2005. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions." *IEEE Transactions on Knowledge & Data Engineering*, 17(6): 734–749.

Agan, Amanda Y, Diag Davenport, Jens Ludwig, and Sendhil Mullainathan. 2023. "Automating automaticity: How the context of human choice affects the extent of algorithmic bias."

Agarwal, Saharsh, Uttara M Ananthakrishnan, and Catherine E Tucker. 2022. "Content Moderation at the Infrastructure Layer: Evidence from Parler."

Aggarwal, Minali, Jennifer Allen, Alexander Coppock, Dan Frankowski, Solomon Messing, Kelly Zhang, James Barnes, Andrew Beasley, Harry Hantman, and Sylvan Zheng. 2023. "A 2 million-person, campaign-wide field experiment shows how digital advertising affects voter turnout." *Nature Human Behaviour*, 1–10.

Ahmad, Wajeeha, Ananya Sen, Charles E Eesley, and Erik Brynjolfsson. Forthcoming. "The role of advertisers and platforms in monetizing misinformation: Descriptive and experimental evidence." *Nature*.

Ajzenman, Nicolas, Bruno Ferman, and Pedro C Sant'Anna. 2023. "Rooting for the same team: On the interplay between political and social identities in the formation of social ties."

Allcott, Hunt, and Matthew Gentzkow. 2017. "Social media and fake news in the 2016 election." *Journal of Economic Perspectives*, 31(2): 211–236.

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. "The welfare effects of social media." *American Economic Review*, 110(3): 629–676.

Allcott, Hunt, Matthew Gentzkow, and Lena Song. 2022. "Digital addiction." *American Economic Review*, 112(7): 2424–63.

Allen, Jennifer, Baird Howland, Markus Mobius, David Rothschild, and Duncan J Watts. 2020. "Evaluating the fake news problem at the scale of the information ecosystem." *Science Advances*, 6(14): eaay3539.

Allen, Jennifer, Cameron Martel, and David G Rand. 2022. "Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program." 1–19.

Andersen, Steffen, Glenn W Harrison, Morten I Lau, and E Elisabet Rutström. 2008. "Eliciting risk and time preferences." *Econometrica*, 76(3): 583–618.

Anderson, Simon P, and Stephen Coate. 2005. "Market provision of broadcasting: A welfare analysis." *The Review of Economic Studies*, 72(4): 947–972.

Andres, Raphaela, and Olga Slivko. 2021. "Combating online hate speech: The impact of legislation on Twitter." *ZEW-Centre for European Economic Research Discussion Paper*.

Angelucci, Charles, Julia Cagé, and Michael Sinkinson. Forthcoming. "Media competition and news diets." *American Economic Journal: Microeconomics*.

An, Jisun, Daniele Quercia, and Jon Crowcroft. 2014. "Partisan sharing: Facebook evidence and societal consequences." 13–24.

Ansolabehere, Stephen, Shanto Iyengar, Adam Simon, and Nicholas Valentino. 1994. "Does attack advertising demobilize the electorate?" *American Political Science Review*, 88(4): 829–838.

Aral, Sinan. 2021. *The Hype Machine: How Social Media Disrupts Our elections, Our Economy, and Our Health–and How We Must Adapt.* Currency.

Arceneaux, Kevin, Martial Foucault, Kalli Giannelos, Jonathan Ladd, and Can Zengin. 2023. "The effects of Facebook access during the 2022 French presidential election: Can we incentivize citizens to be better informed and less polarized?"

Arechar, Antonio A, Jennifer Allen, Adam J Berinsky, Rocky Cole, Ziv Epstein, Kiran Garimella, Andrew Gully, Jackson G Lu, Robert M Ross, Michael N Stagnaro, et al. 2023. "Understanding and combatting misinformation across 16 countries on six continents." *Nature Human Behaviour*, 7(9): 1502–1513.

Argentesi, Elena, Paolo Buccirossi, Emilio Calvano, Tomaso Duso, Alessia Marrazzo, and Salvatore Nava. 2021. "Merger policy in digital markets: an ex post assessment." *Journal of Competition Law & Economics*, 17(1): 95–140.

Aridor, Guy. 2023. "Measuring substitution patterns in the attention economy: An experimental approach."

Aridor, Guy, Duarte Gonçalves, Daniel Kluver, Ruoyan Kong, and Joseph Konstan. 2022. "The economics of recommender systems: Evidence from a field experiment on MovieLens."

Aridor, Guy, Rafael Jiménez-Durán, Ro'ee Levy, and Lena Song. Forthcoming. "Experiments with Social Media." In *Handbook of Experimental Methods in the Social Sciences*. Edward Elgar Publishing.

Aridor, Guy, Yeon-Koo Che, and Tobias Salz. 2023. "The effect of privacy regulation on the data industry: Empirical evidence from GDPR." *RAND Journal of Economics*.

Aridor, Guy, Yeon-Koo Che, Brett Hollenbeck, Maximilian Kaiser, and Daniel McCarthy. 2024. "Evaluating The Impact of Privacy Regulation on E-Commerce Firms: Evidence from Apple's App Tracking Transparency."

Armona, Luis. 2019. "Online social network effects in labor markets: Evidence From Facebook's entry into college campuses."

Asimovic, Nejla, Jonathan Nagler, and Joshua A Tucker. 2023. "Replicating the effects of Facebook deactivation in an ethnically polarized setting." *Research & Politics*, 10(4): 20531680231205157.

Asimovic, Nejla, Jonathan Nagler, Richard Bonneau, and Joshua A Tucker. 2021. "Testing the effects of Facebook usage in an ethnically polarized setting." *Proceedings of the National Academy of Sciences*, 118(25): e2022819118.

Athey, Susan, Christian Catalini, and Catherine Tucker. 2017. "The digital privacy paradox: Small money, small costs, small talk."

Athey, Susan, Kristen Grabarz, Michael Luca, and Nils Wernerfelt. 2023*a*. "Digital public health interventions at scale: The impact of social media advertising on beliefs and outcomes related to COVID vaccines." *Proceedings of the National Academy of Sciences*, 120(5): e2208110120.

Athey, Susan, Matias Cersosimo, Kristine Koutout, and Zelin Li. 2023*b*. "Emotion- versus reasoning-based drivers of misinformation sharing: A field experiment using text message courses in Kenya."

Bagwell, Kyle. 2007. "The economic analysis of advertising." *Handbook of Industrial Organization*, 3: 1701–1844.

Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. "Exposure to opposing views on social media can increase political polarization." *Proceedings of the National Academy of Sciences*, 115(37): 9216–9221.

Bakshy, Eytan, Dean Eckles, Rong Yan, and Itamar Rosenn. 2012. "Social influence in social advertising: evidence from field experiments." 146–161.

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

Barberá, Pablo. 2015. "How social media reduces mass political polarization. Evidence from Germany, Spain, and the US." *Paper Prepared for the 2015 APSA Conference*, 46: 1–46.

Barrera, Oscar, Sergei Guriev, Emeric Henry, and Ekaterina Zhuravskaya. 2020. "Facts, alternative facts, and fact checking in times of post-truth politics." *Journal of Public Economics*, 182: 104123.

Baslandze, Salome, Jeremy Greenwood, Ricardo Marto, and Sara Moreira. 2023. "The expansion of varieties in the new age of advertising." *Review of Economic Dynamics*.

Batista Pereira, Frederico, Natália S Bueno, Felipe Nunes, and Nara Pavão. 2022. "Fake news, fact checking, and partisanship: The resilience of rumors in the 2018 Brazilian elections." *The Journal of Politics*, 84(4): 2188–2201.

Becker, Gary S, and Kevin M Murphy. 1993. "A simple theory of advertising as a good or bad." *The Quarterly Journal of Economics*, 108(4): 941–964.

Becker, Gordon M, Morris H DeGroot, and Jacob Marschak. 1964. "Measuring utility by a single-response sequential method." *Behavioral Science*, 9(3): 226–232.

Beknazar-Yuzbashev, George, Rafael Jiménez Durán, and Mateusz Stalinski. Forthcoming. "A model of harmful yet engaging content on social media." *AEA Papers and Proceedings*.

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."

Bénabou, Roland, and Jean Tirole. 2006. "Incentives and prosocial behavior." *American Economic Review*, 96(5): 1652–1678.

Benjamin, Daniel J, Ori Heffetz, Miles S Kimball, and Alex Rees-Jones. 2012. "What do you think would make you happier? What do you think you would choose?" *American Economic Review*, 102(5): 2083–2110.

Bergemann, Dirk, Alessandro Bonatti, and Tan Gan. 2022. "The economics of social data." *The RAND Journal of Economics*, 53(2): 263–296.

Bergemann, Dirk, and Alessandro Bonatti. 2011. "Targeting in advertising markets: implications for offline versus online media." *The RAND Journal of Economics*, 42(3): 417–443.

Berger, Lara Marie, Anna Kerkhof, Felix Mindl, and Johannes Munster. 2023. "Debunking 'fake news' on social media: Short-term and longer-term effects of fact checking and media literacy interventions."

Bianchi, Francesco, Roberto Gomez Cram, and Howard Kung. Forthcoming. "Using social media to identify the effects of congressional viewpoints on asset prices." *The Review of Financial Studies*.

Bianchi, Francesco, Roberto Gómez-Cram, Thilo Kind, and Howard Kung. 2023. "Threats to central bank independence: High-frequency identification with Twitter." *Journal of Monetary Economics*.

Blair, Robert A, Jessica Gottlieb, Brendan Nyhan, Laura Paler, Pablo Argote, and Charlene J Stainfield. 2023. "Interventions to counter misinformation: Lessons from the global north and applications to the global south." *Current Opinion in Psychology*, 101732.

Blake, Thomas, Chris Nosko, and Steven Tadelis. 2015. "Consumer heterogeneity and paid search effectiveness: A large-scale field experiment." *Econometrica*, 83(1): 155–174.

Bonatti, Alessandro, Dirk Bergemann, and Nicholas Wu. 2023. "How do digital advertising auctions impact product prices?"

Bond, Robert M, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. 2012. "A 61-million-person experiment in social influence and political mobilization." *Nature*, 489(7415): 295–298.

Boxell, Levi, Matthew Gentzkow, and Jesse M Shapiro. 2017. "Greater Internet use is not associated with faster growth in political polarization among US demographic groups." *Proceedings of the National Academy of Sciences*, 114(40): 10612–10617.

Brady, William J, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. "Emotion shapes the diffusion of moralized content in social networks." *Proceedings of the National Academy of Sciences*, 114(28): 7313–7318.

Braghieri, Luca, Ro'ee Levy, and Alexey Makarin. 2022. "Social media and mental health." *American Economic Review*, 112(11): 3660–3693.

Brailovskaia, Julia, Fabienne Ströse, Holger Schillack, and Jürgen Margraf. 2020. "Less Facebook use — More well-being and a healthier lifestyle? An experimental intervention study." *Computers in Human Behavior*, 108: 106332.

Breza, Emily, Fatima Cody Stanford, Marcella Alsan, Burak Alsan, Abhijit Banerjee, Arun G Chandrasekhar, Sarah Eichmeyer, Traci Glushko, Paul Goldsmith-Pinkham, and Kelly Holland. 2021. "Effects of a large-scale social media advertising campaign on holiday travel and COVID-19 infections: A cluster randomized controlled trial." *Nature Medicine*, 27(9): 1622–1628.

Brown, Megan A, James Bisbee, Angela Lai, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. 2022. "Echo chambers, rabbit holes, and algorithmic bias: How YouTube recommends content to real users."

Brynjolfsson, Erik, Avinash Collis, and Felix Eggers. 2019. "Using massive online choice experiments to measure changes in well-being." *Proceedings of the National Academy of Sciences*, 116(15): 7250–7255.

Brynjolfsson, Erik, Avinash Collis, Asad Liaqat, Daley Kutzman, Haritz Garro, Daniel Deisenroth, Nils Wernerfelt, and Jae Joon Lee. 2023. "The digital welfare of nations: New measures of welfare gains and inequality."

Brynjolfsson, Erik, Seon Tae Kim, and Joo Hee Oh. Forthcoming. "The attention economy: Measuring the value of free goods on the internet." *Information Systems Research*.

Bursztyn, Leonardo, Benjamin R Handel, Rafael Jimenez, and Christopher Roth. 2023*a*. "When product markets become collective traps: The case of social media."

Bursztyn, Leonardo, Georgy Egorov, Ingar Haaland, Aakaash Rao, and Christopher Roth. 2023*b*. "Justifying dissent." *The Quarterly Journal of Economics*, 138(3): 1403–1451.

Bursztyn, Leonardo, Georgy Egorov, Ruben Enikolopov, and Maria Petrova. 2019. "Social media and xenophobia: evidence from Russia."

Burtch, Gordon, Qinglai He, Yili Hong, and Dokyun Lee. 2022. "How do peer awards motivate creative content? Experimental evidence from Reddit." *Management Science*, 68(5): 3488–3506.

Cagé, Julia, Nicolas Hervé, and Béatrice Mazoyer. 2022. "Social media influence mainstream media: Evidence from two billion tweets."

Cagé, Julia, Nicolas Hervé, and Marie-Luce Viaud. 2020. "The production of information in an online world." *The Review of Economic Studies*, 87(5): 2126–2164.

Campante, Fillipe, Ruben Durante, and Andrea Tesei. 2023. *The Political Economy of Social Media.* CEPR Press.

Cao, Andy, Jason M Lindo, and Jiee Zhong. 2023. "Can social media rhetoric incite hate incidents? Evidence from Trump's "Chinese Virus" tweets." *Journal of Urban Economics*, 137: 103590.

Carney, Kevin. 2022. "The effect of social media on voters: Experimental evidence from an Indian election."

Casanueva Artís, Annali. 2023. "Can chants in the street change politics' tune? Evidence from the 15M movement in Spain."

Casanueva Artís, Annali, Vladimir Avetian, Sulin Sardoschau, and Kritika Saxena. 2023. "Going Viral in a Pandemic: Social Media and the Broadening of the Black Lives Matter Movement."

Chen, Annie Y, Brendan Nyhan, Jason Reifler, Ronald E Robertson, and Christo Wilson. 2023. "Subscriptions and external links help drive resentful users to alternative and extremist YouTube channels." *Science Advances*, 9(35): eadd8080.

Chen, Yan, F Maxwell Harper, Joseph Konstan, and Sherry Xin Li. 2010. "Social comparisons and contributions to online communities: A field experiment on Movielens." *American Economic Review*, 100(4): 1358–1398.

Chetty, Raj, Matthew O Jackson, Theresa Kuchler, Johannes Stroebel, Nathaniel Hendren, Robert B Fluegge, Sara Gong, Federico Gonzalez, Armelle Grondin, Matthew Jacob, et al. 2022. "Social capital I: Measurement and associations with economic mobility." *Nature*, 608(7921): 108–121.

Choi, Jay Pil, Doh-Shin Jeon, and Byung-Cheol Kim. 2019. "Privacy and personal data collection with information externalities." *Journal of Public Economics*, 173: 113–124.

Collis, Avinash, Alex Moehring, Ananya Sen, and Alessandro Acquisti. 2021. "Information frictions and heterogeneity in valuations of personal data."

Collis, Avinash, and Felix Eggers. 2022. "Effects of restricting social media usage on wellbeing and performance: A randomized control trial among students." *PLOS One*, 17(8): e0272416.

Cookson, J Anthony, William Mullins, and Marina Niessner. 2024. "Social Media and Finance."

Coppock, Alexander, Donald P Green, and Ethan Porter. 2022. "Does digital advertising affect vote choice? Evidence from a randomized field experiment." *Research & Politics*, 9(1): 20531680221076901.

Coppock, Alexander, Seth J Hill, and Lynn Vavreck. 2020. "The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments." *Science Advances*, 6(36): eabc4046.

D'Amico, Leonardo, and Guido Tabellini. 2022. "Disengaging from Reality: Online Behavior and Unpleasant Political News."

Deolankar, Varad, Jessica Fong, and S Sriram. 2023. "Content generation on social media: The role of negative peer feedback."

Dessaint, Olivier, Thierry Foucault, and Laurent Frésard. 2021. "Does alternative data improve financial forecasting? the horizon effect." *The Journal of Finance*.

Di Tella, Rafael, Ramiro H Gálvez, and Ernesto Schargrodsky. 2021. "Does social media cause polarization? Evidence from access to Twitter echo chambers during the 2019 Argentine presidential debate."

Donati, Dante, Ruben Durante, Francesco Sobbrio, and Dijana Zejcirovic. 2022. "Lost in the net? Broadband Internet and youth mental health."

Dvir-Gvirsman, Shira. 2019. "I like what I see: Studying the influence of popularity cues on attention allocation and news selection." *Information, Communication & Society*, 22(2): 286–305.

Eady, Gregory, Jonathan Nagler, Andy Guess, Jan Zilinsky, and Joshua A Tucker. 2019. "How many people live in political bubbles on social media? Evidence from linked survey and Twitter data." *Sage Open*, 9(1): 2158244019832705.

Eckles, Dean, René F Kizilcec, and Eytan Bakshy. 2016. "Estimating peer effects in networks with peer encouragement designs." *Proceedings of the National Academy of Sciences*, 113(27): 7316–7322.

Ehrmann, Michael, and Alena Wabitsch. 2022. "Central bank communication with non-experts – A road to nowhere?" *Journal of Monetary Economics*, 127: 69–85.

Ehrmann, Michael, and Paul Hubert. 2023. "Information acquisition ahead of monetary policy announcements."

Enikolopov, Ruben, Alexey Makarin, and Maria Petrova. 2020. "Social media and protest participation: Evidence from Russia." *Econometrica*, 88(4): 1479–1514.

Enikolopov, Ruben, Alexey Makarin, Maria Petrova, and Leonid Polishchuk. 2023. "Social image, networks, and protest participation."

Enikolopov, Ruben, Maria Petrova, and Konstantin Sonin. 2018. "Social media and corruption." *American Economic Journal: Applied Economics*, 10(1): 150–174.

Enikolopov, Ruben, Maria Petrova, Gianluca Russo, and David Yanagizawa-Drott. 2024. "Socializing Alone: How Online Homophily Has Undermined Social Cohesion in the US."

Enríquez, José Ramón, Horacio Larreguy, John Marshall, and Alberto Simpser. 2024. "Mass political information on social media: Facebook ads, electorate saturation, and electoral accountability in Mexico." *Journal of the European Economic Association*, jvae011.

Ershov, Daniel, and Juan S Morales. 2024. "Sharing news left and right: Frictions and misinformation on Twitter." *The Economic Journal*, ueae027.

Ershov, Daniel, and Matthew Mitchell. Forthcoming. "The effects of influencer advertising disclosure regulations: Evidence from Instagram." *RAND Journal of Economics*.

Ershov, Daniel, Yanting He, and Stephan Seiler. 2023. "How much influencer marketing is undisclosed? Evidence from Twitter."

Fergusson, Leopoldo, and Carlos Molina. 2021. "Facebook causes protests."

Filippas, Apostolos, John J Horton, and Elliot Lipnowski. 2021. "The production and consumption of social media."

Flaxman, Seth, Sharad Goel, and Justin M Rao. 2016. "Filter bubbles, echo chambers, and online news consumption." *Public Opinion Quarterly*, 80(S1): 298–320.

Fowler, Erika Franklin, Michael M Franz, Gregory J Martin, Zachary Peskowitz, and Travis N Ridout. 2021. "Political advertising online and offline." *American Political Science Review*, 115(1): 130–149.

FTC. 2021. "FTC vs. Facebook 1:20-cv-03590-JEB." `https://www.ftc.gov/system/files/documents/cases/ecf_75-1_ftc_v_facebook_public_redacted_fac.pdf`.

Fujiwara, Thomas, Karsten Müller, and Carlo Schwarz. Forthcoming. "The effect of social media on elections: Evidence from the United States." *Journal of the European Economic Association*.

Gans, Joshua S, Avi Goldfarb, and Mara Lederman. 2021. "Exit, tweets, and loyalty." *American Economic Journal: Microeconomics*, 13(2): 68–112.

Garbiras-Díaz, Natalia, and Mateo Montenegro. 2022. "All eyes on them: A field experiment on citizen oversight and electoral integrity." *American Economic Review*, 112(8): 2631–2668.

Garrett, R Kelly. 2009. "Echo chambers online?: Politically motivated selective exposure among Internet news users." *Journal of Computer-Mediated Communication*, 14(2): 265–285.

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. 2011. "Ideological segregation online and offline." *The Quarterly Journal of Economics*, 126(4): 1799–1839.

Gentzkow, Matthew, Jesse M Shapiro, Frank Yang, and Ali Yurukoglu. 2024. "Pricing power in advertising markets: Theory and evidence." *American Economic Review*, 114(2): 500–533.

Gillespie, Tarleton. 2018. *Custodians of the Internet.* Yale University Press.

González-Bailón, Sandra, David Lazer, Pablo Barberá, Meiqing Zhang, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Deen Freelon, Matthew Gentzkow, Andrew M. Guess, Shanto Iyengar, Young Mie Kim, Neil Malhotra, Devra Moehler, Brendan Nyhan, Jennifer Pan, Carlos Velasco Rivera, Jaime Settle, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker. 2023. "Asymmetric ideological segregation in exposure to political news on Facebook." *Science*, 381(6656): 392–398.

Gordon, Brett R, Robert Moakler, and Florian Zettelmeyer. 2023. "Close enough? A large-scale exploration of non-experimental approaches to advertising measurement." *Marketing Science*, 42(4): 768–793.

Guess, Andrew, Brendan Nyhan, Benjamin Lyons, and Jason Reifler. 2018. "Avoiding the echo chamber about echo chambers." *Knight Foundation*, 2(1): 1–25.

Guess, Andrew M. 2021. "(Almost) everything in moderation: New evidence on Americans' online media diets." *American Journal of Political Science*, 65(4): 1007–1022.

Guess, Andrew M, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. "A digital media literacy intervention increases discernment between mainstream and false news in the United States and India." *Proceedings of the National Academy of Sciences*, 117(27): 15536–15545.

Guess, Andrew M, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, et al. 2023*a*. "Reshares on social media amplify political news but do not detectably affect beliefs or opinions." *Science*, 381(6656): 404–408.

Guess, Andrew M., Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, Sandra González-Bailón, Edward Kennedy, Young Mie Kim, David Lazer, Devra Moehler, Brendan Nyhan, Carlos Velasco Rivera, Jaime Settle, Daniel Robert Thomas, Emily Thorson, Rebekah Tromble, Arjun Wilkins, Magdalena Wojcieszak, Beixian Xiong, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A. Tucker. 2023*b*. "How do social media feed algorithms affect attitudes and behavior in an election campaign?" *Science*, 381(6656): 398–404.

Guriev, Sergei, Emeric Henry, Theo Marquis, and Ekaterina Zhuravskaya. 2023. "Curtailing false news, amplifying truth."

Guriev, Sergei, Nikita Melnikov, and Ekaterina Zhuravskaya. 2021. "3G Internet and Confidence in Government." *The Quarterly Journal of Economics*, 136(4): 2533–2613.

Gylfason, Gísli. 2023. "From Tweets to the Streets: Twitter and Extremist Protests in the United States."

Halberstam, Yosh, and Brian Knight. 2016. "Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter." *Journal of Public Economics*, 143: 73–88.

Hangartner, Dominik, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Buse Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, et al. 2021. "Empathy-based counterspeech can reduce racist hate speech in a social media field experiment." *Proceedings of the National Academy of Sciences*, 118(50): e2116310118.

Hatte, Sophie, Etienne Madinier, and Ekaterina Zhuravskaya. 2023. "Reading Twitter in the newsroom: Web 2.0 and traditional-media reporting of conflicts." *CEPR Discussion Paper No. DP16167*.

Henry, Emeric, Ekaterina Zhuravskaya, and Sergei Guriev. 2022. "Checking and sharing alt-facts." *American Economic Journal: Economic Policy*, 14(3): 55–86.

Holder, Patrick, Haaris Mateen, Anya Schiffrin, and Haris Tabakovic. 2023. "Paying for news: What Google and Meta owe US publishers."

Holtz, David, Benjamin Carterette, Praveen Chandar, Zahra Nazari, Henriette Cramer, and Sinan Aral. 2020. "The engagement-diversity connection: Evidence from a field experiment on Spotify." 75–76.

Hoong, Ruru. 2021. "Self control and smartphone use: An experimental study of soft commitment devices." *European Economic Review*, 140: 103924.

Hosseinmardi, Homa, Amir Ghasemian, Aaron Clauset, Markus Mobius, David M Rothschild, and Duncan J Watts. 2021. "Examining the consumption of radical content on YouTube." *Proceedings of the National Academy of Sciences*, 118(32): e2101967118.

Huang, Justin T, and Sridhar Narayanan. 2020. "Effects of attention and recognition on engagement, content creation and sharing: Experimental evidence from an image sharing social network."

Huang, Shan, Sinan Aral, Yu Jeffrey Hu, and Erik Brynjolfsson. 2020. "Social advertising effectiveness across products: A large-scale field experiment." *Marketing Science*, 39(6): 1142–1165.

Hunt, Melissa G, Rachel Marx, Courtney Lipson, and Jordyn Young. 2018. "No more FOMO: Limiting social media decreases loneliness and depression." *Journal of Social and Clinical Psychology*, 37(10): 751–768.

Huszár, Ferenc, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt.

2022. "Algorithmic amplification of politics on Twitter." *Proceedings of the National Academy of Sciences*, 119(1): e2025334119.

Jacobson, Gary C. 2015. "How do campaigns matter?" *Annual Review of Political Science*, 18: 31–47.

Jiménez Durán, Rafael. 2022. "The economics of content moderation: Theory and experimental evidence from hate speech on Twitter."

Jiménez Durán, Rafael, Karsten Müller, and Carlo Schwarz. 2022. "The effect of content moderation on online and offline hate: Evidence from Germany's NetzDG."

Johnson, Garrett A, Scott K Shriver, and Samuel G Goldberg. 2023. "Privacy and market concentration: intended and unintended consequences of the GDPR." *Management Science*.

Kalla, Joshua L, and David E Broockman. 2018. "The minimal persuasive effects of campaign contact in general elections: Evidence from 49 field experiments." *American Political Science Review*, 112(1): 148–166.

Katsaros, Matthew, Kathy Yang, and Lauren Fratamico. 2022. "Reconsidering tweets: Intervening during tweet creation decreases offensive content." Vol. 16, 477–487.

Kemp, Simon. 2023. "Digital 2023: Global Overview Report." *Available at* `https://datareportal.com/reports/digital-2023-global-overview-report`.

Kerkhof, Anna. 2020. "Advertising and content differentiation: Evidence from YouTube."

Kim, Jin Woo, Andrew Guess, Brendan Nyhan, and Jason Reifler. 2021. "The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity." *Journal of Communication*, 71(6): 922–946.

Kominers, Scott Duke, and Jesse M Shapiro. 2024. "Content Moderation with Opaque Policies." National Bureau of Economic Research.

Kozyreva, Anastasia, Philipp Lorenz-Spreen, Stefan Herzog, Ullrich Ecker, Stephan Lewandowsky, and Ralph Hertwig. Forthcoming. "Toolbox of individual-level interventions against online misinformation." *Nature Human Behaviour*.

Kreps, Sarah E, and Douglas L Kriner. 2022. "The COVID-19 infodemic and the efficacy of interventions intended to reduce misinformation." *Public Opinion Quarterly*, 86(1): 162–175.

Kross, Ethan, Philippe Verduyn, Gal Sheppes, Cory K Costello, John Jonides, and Oscar Ybarra. 2021. "Social media and well-being: Pitfalls, progress, and next steps." *Trends in Cognitive Sciences*, 25(1): 55–66.

Lazer, David. 2015. "The rise of the social algorithm." *Science*, 348(6239): 1090–1091.

Lee, Dokyun, Kartik Hosanagar, and Harikesh S Nair. 2018. "Advertising content and consumer engagement on social media: Evidence from Facebook." *Management Science*, 64(11): 5105–5131.

Lerner, Josh, and Jean Tirole. 2002. "Some simple economics of open source." *The Journal of Industrial Economics*, 50(2): 197–234.

Leung, Tin Cheuk, and Koleman S Strumpf. 2023. "All the headlines that are fit to change: Analysis of headline changes in the media industry."

Levy, Gilat, and Ronny Razin. 2019. "Echo chambers and their effects on economic and political outcomes." *Annual Review of Economics*, 11: 303–328.

Levy, Ro'ee, and Martin Mattsson. 2023. "The effects of social movements: Evidence from# MeToo."

Levy, Ro'ee. 2021. "Social media, news consumption, and polarization: Evidence from a field experiment." *American Economic Review*, 111(3): 831–870.

Lewis, Randall A, and Justin M Rao. 2015. "The unfavorable economics of measuring the returns to advertising." *The Quarterly Journal of Economics*, 130(4): 1941–1973.

Lin, Hause, Haritz Garro, Nils Wernerfelt, Jesse Shore, Adam Hughes, Daniel Deisenroth, Nathaniel Barr, Adam Berinsky, Dean Eckles, Gordon Pennycook, et al. 2024. "Reducing misinformation sharing at scale using digital accuracy prompt ads."

Lin, Tesary, and Avner Strulov-Shlain. 2023. "Choice architecture, privacy valuations, and selection bias in consumer data."

Liu, Naijia, Matthew A Baum, Adam J Berinsky, Allison JB Chaney, Justin de Benedictis-Kessner, Andy Guess, Dean Knox, Christopher Lucas, Rachel Mariman, and Brandon M Stewart. 2023. "Algorithmic recommendations have limited effects on polarization: A naturalistic experiment on YouTube."

Liu, Yi, Pinar Yildirim, and Z John Zhang. 2022. "Implications of revenue models and technology for content moderation strategies." *Marketing Science*, 41(4): 831–847.

Lorenz-Spreen, Philipp, Lisa Oswald, Stephan Lewandowsky, and Ralph Hertwig. 2023. "A systematic review of worldwide causal and correlational evidence on digital media and democracy." *Nature Human Behaviour*, 7(1): 74–101.

Luca, Michael. 2015. "User-generated content and social media." In *Handbook of Media Economics*. Vol. 1, 563–592. Elsevier.

Lucas, Elizabeth, Cecilia O Alm, and Reynold Bailey. 2019. "Understanding human and predictive moderation of online science discourse." 1–5, IEEE.

Madio, Leonardo, and Martin Quinn. 2023. "Content moderation and advertising in social media platforms."

Marotta, Veronica, and Alessandro Acquisti. 2018. "Interrupting interruptions: A digital experiment on social media and performance."

Martel, Cameron, and David G Rand. 2023*a*. "Misinformation warning labels are widely effective: A review of warning effects and their moderating features." *Current Opinion in Psychology*, 101710.

Martel, Cameron, and David Rand. 2023*b*. "Fact-checker warning labels are effective even for those who distrust fact-checkers."

Martel, Cameron, Jennifer Allen, Gordon Pennycook, and David G Rand. 2024. "Crowds can effectively identify misinformation at scale." *Perspectives on Psychological Science*, 19(2): 477–488.

McClain, Colleen, Michelle Faverio, Monica Anderson, and Eugenie Park. 2023. "How Americans view data privacy." `https://www.pewresearch.org/internet/2023/10/18/how-americans-view-data-privacy/`.

McDool, Emily, Philip Powell, Jennifer Roberts, and Karl Taylor. 2020. "The internet and children's psychological wellbeing." *Journal of Health Economics*, 69: 102274.

Melnikov, Nikita. 2021. "Mobile internet and political polarization."

Messing, Solomon. 2023. "Are algorithmic bias claims supported?" *Science*, 381(6665): 1420–1420.

Messing, Solomon, and Sean J Westwood. 2014. "Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online." *Communication Research*, 41(8): 1042–1063.

Moehring, Alex. 2023. "Personalized rankings and user engagement: An empirical evaluation of the Reddit news feed."

Mosleh, Mohsen, Cameron Martel, Dean Eckles, and David G Rand. 2021. "Shared partisanship dramatically increases social tie formation in a Twitter field experiment." *Proceedings of the National Academy of Sciences*, 118(7): e2022761118.

Mosquera, Roberto, Mofioluwasademi Odunowo, Trent McNamara, Xiongfei Guo, and Ragan Petrie. 2020. "The economic effects of Facebook." *Experimental Economics*, 23: 575–602.

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.

Müller, Karsten, and Carlo Schwarz. 2022. "The effects of online content moderation: Evidence from President Trump's account deletion."

Müller, Karsten, and Carlo Schwarz. 2023. "From hashtag to hate crime: Twitter and anti-minority sentiment." *American Economic Journal: Applied Economics*.

Mummalaneni, Simha, Hema Yoganarasimhan, and Varad Pathak. 2023. "How do content producers respond to engagement on social media platforms?"

Munger, Kevin. 2017. "Tweetment effects on the tweeted: Experimentally reducing racist harassment." *Political Behavior*, 39: 629–649.

Munger, Kevin. 2021. "Don't@ me: Experimentally reducing partisan incivility on Twitter." *Journal of Experimental Political Science*, 8(2): 102–116.

Neiman, Brent, and Joseph Vavra. 2023. "The rise of niche consumption." *American Economic Journal: Macroeconomics*, 15(3): 224–264.

Newman, Nic, Richard Fletcher, Kirsten Eddy, Craig T Robertson, and Rasmus Kleis Nielsen. 2023. "Reuters institute digital news report 2023." *Reuters Institute for the Study of Journalism*.

Nyhan, Brendan. 2020. "Facts and myths about misperceptions." *Journal of Economic Perspectives*, 34(3): 220–236.

Nyhan, Brendan. 2021. "Why the backfire effect does not explain the durability of political misperceptions." *Proceedings of the National Academy of Sciences*, 118(15): e1912440117.

Nyhan, Brendan, and Jason Reifler. 2010. "When corrections fail: The persistence of political misperceptions." *Political Behavior*, 32(2): 303–330.

Nyhan, Brendan, Ethan Porter, Jason Reifler, and Thomas J Wood. 2020. "Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability." *Political Behavior*, 42: 939–960.

Nyhan, Brendan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, et al. 2023. "Like-minded sources on Facebook are prevalent but not polarizing." *Nature*, 1–8.

Ofcom. 2022. "Ipsos Iris passive monitoring data analysis."

Papanastasiou, Yiangos. 2020. "Fake news propagation and detection: A sequential model." *Management Science*, 66(5): 1826–1846.

Pariser, Eli. 2011. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin.

Pennycook, Gordon, Adam Bear, Evan T Collins, and David G Rand. 2020. "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings." *Management Science*, 66(11): 4944–4957.

Pennycook, Gordon, and David G Rand. 2022. "Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation." *Nature Communications*, 13(1): 2333.

Persily, Nathaniel, and Joshua A Tucker. 2020. *Social media and democracy: The state of the field, prospects for reform*. Cambridge University Press.

Peterson, Erik, Sharad Goel, and Shanto Iyengar. 2021. "Partisan selective exposure in online news consumption: Evidence from the 2016 presidential campaign." *Political Science Research and Methods*, 9(2): 242–258.

Petrova, Maria, Ananya Sen, and Pinar Yildirim. 2021. "Social media and political contributions: The impact of new technology on political competition." *Management Science*, 67(5): 2997–3021.

Peukert, Christian, Stefan Bechtold, Michail Batikas, and Tobias Kretschmer. 2022. "Regulatory spillovers and data governance: Evidence from the GDPR." *Marketing Science*, 41(4): 746–768.

Pogorelskiy, Kirill, and Matthew Shum. 2019. "News we like to share: How news sharing on social networks influences voting outcomes."

Prat, Andrea. 2018. "Media power." *Journal of Political Economy*, 126(4): 1747–1783.

Prat, Andrea, and Tommaso Valletti. 2022. "Attention oligopoly." *American Economic Journal: Microeconomics*, 14(3): 530–57.

Prince, Jeffrey T, and Scott Wallsten. 2022. "How much is privacy worth around the world and across platforms?" *Journal of Economics & Management Strategy*, 31(4): 841–861.

Qin, Bei, David Strömberg, and Yanhui Wu. 2021. "Social media and collective action in China."

Rajkumar, Karthik, Guillaume Saint-Jacques, Iavor Bojinov, Erik Brynjolfsson, and Sinan Aral. 2022. "A causal test of the strength of weak ties." *Science*, 377(6612): 1304–1310.

Rambachan, Ashesh, Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. 2020. "An economic perspective on algorithmic fairness." Vol. 110, 91–95.

Ribeiro, Manoel Horta, Justin Cheng, and Robert West. 2022. "Automated content moderation increases adherence to community guidelines."

Ridout, Travis N, and Michael M Franz. 2011. *The persuasive power of campaign advertising*. Temple University Press.

Rochet, Jean-Charles, and Jean Tirole. 2003. "Platform competition in two-sided markets." *Journal of the European Economic Association*, 1(4): 990–1029.

Roozenbeek, Jon, Sander Van Der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky. 2022. "Psychological inoculation improves resilience against misinformation on social media." *Science Advances*, 8(34): eabo6254.

Rosenquist, James Niels, Fiona M Scott Morton, and Samuel N Weinstein. 2021. "Addictive technology and its implications for antitrust enforcement." *North Carolina Law Review*, 100: 431.

Sabatini, Fabio. 2023. "The Behavioral, Economic, and Political Impact of the Internet and Social Media: Empirical Challenges and Approaches."

Scott Morton, Fiona, Pascal Bouvier, Ariel Ezrachi, Bruno Jullien, Roberta Katz, Gene Kimmelman, A Douglas Melamed, and Jamie Morgenstern. 2019. "Committee for the study of digital platforms: Market structure and antitrust subcommittee report." *Chicago: Stigler Center for the Study of the Economy and the State, University of Chicago Booth School of Business*, 36.

Sen, Ananya, and Pinar Yildirim. 2015. "Clicks and editorial decisions: How does popularity shape online news coverage?"

Shearer, Elisa. 2021. "More than eight-in-ten Americans get news from digital devices." *Pew Research Center*, 12.

Siegel, Alexandra A, and Vivienne Badaan. 2020. "# No2Sectarianism: Experimental approaches to reducing sectarian hate speech online." *American Political Science Review*, 114(3): 837–855.

Smith, Ben. 2021. "How TikTok reads your mind." *New York Times.*

Smith, Ben. 2023. *Traffic: Genius, Rivalry, and Delusion in the Billion-Dollar Race to Go Viral.* Penguin Press.

Song, Lena. 2023. "Closing the distance: The effects of social media content on support for racial justice."

Srinivasan, Karthik. 2023. "Paying Attention."

Steinert-Threlkeld, Zachary C., Delia Mocanu, Alessandro Vespignani, and James Fowler. 2015. "Online social networks and offline protest." *EPJ Data Science*, 4(1): 1–9.

Stroud, Natalie Jomini. 2008. "Media use and political predispositions: Revisiting the concept of selective exposure." *Political Behavior*, 30: 341–366.

Sun, Monic, and Feng Zhu. 2013. "Ad revenue and content commercialization: Evidence from blogs." *Management Science*, 59(10): 2314–2331.

Sunstein, Cass R. 2017. *# Republic: Divided Democracy in the Age of Social Media.* Princeton University Press.

Surgeon General. 2023. "Social media and youth mental health: the US surgeon general's advisory."

Tabellini, Guido, Marco Manacorda, and Andrea Tesei. 2023. "Mobile internet and the rise of communitarian politics."

Tadelis, Steven, Christopher Hooton, Utsav Manjeer, Daniel Deisenroth, Nils Wernerfelt, Nick Dadson, and Lindsay Greenbaum. 2023. "Learning, sophistication, and the returns to advertising: Implications for differences in firm performance."

Tech For Campaigns. 2021. "2020 Political Digital Advertising Report." `https://www.techforcampaigns.org/impact/2020-political-digital-advertising-report`.

Thaler, Michael. 2021. "The supply of motivated beliefs."

Thomas, Daniel Robert, and Laila A Wahedi. 2023. "Disrupting hate: The effect of deplatforming hate organizations on their online audience." *Proceedings of the National Academy of Sciences*, 120(24): e2214080120.

Thompson, Alex. 2020. "Why the right wing has a massive advantage on Facebook." *Politico.*

Tufekci, Zeynep. 2018. "YouTube, the great radicalizer." *The New York Times*, 10(3): 2018.

Valkenburg, Patti M. 2022. "Social media use and well-being: What we know and what we need to know." *Current Opinion in Psychology*, 45: 101294.

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

Walter, Nathan, Jonathan Cohen, R Lance Holbert, and Yasmin Morag. 2020. "Fact-checking: A meta-analysis of what works and for whom." *Political Communication*, 37(3): 350–375.

Weinstein, Emily, and Carrie James. 2022. "School-based initiatives promoting digital citizenship and healthy digital media use." *Handbook of Adolescent Digital Media Use and Mental Health; Cambridge University Press: Cambridge, UK*, 365.

Wernerfelt, Nils, Anna Tuchman, Bradley Shapiro, and Robert Moakler. 2022. "Estimating the value of offsite data to advertisers on Meta." *University of Chicago, Becker Friedman Institute for Economics Working Paper.*

Werner, Geyser. 2022. "The state of the direct-to-consumer (DTC) industry." `https://influencermarketinghub.com/direct-to-consumer-industry/`.

Wojcik, Stefan, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, MB Hunzaker, Keith Coleman, and Jay Baxter. 2022. "Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation."

Wylie, Christopher. 2019. *Mindf* ck: Cambridge Analytica and the Plot to Break America.* Random House.

Yildirim, Mustafa Mikdat, Jonathan Nagler, Richard Bonneau, and Joshua A Tucker. 2021. "Short of suspension: How suspension warnings can reduce hate speech on Twitter." *Perspectives on Politics*, 1–13.

Zeng, Zhiyu, Hengchen Dai, Dennis J Zhang, Heng Zhang, Renyu Zhang, Zhiwei Xu, and Zuo-Jun Max Shen. 2022. "The impact of social nudges on user-generated content for social network platforms." *Management Science.*

Zhang, Xiaoquan, and Feng Zhu. 2011. "Group size and incentives to contribute: A natural experiment at Chinese Wikipedia." *American Economic Review*, 101(4): 1601–1615.

Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov. 2020. "Political effects of the internet and social media." *Annual Review of Economics*, 12: 415–438.