Managing Online Participation: Evidence from a Field Experiment with AI Agents

Wajeeha Ahmad *
Management Science and Engineering
Stanford University

This version: November 6, 2025

Please click here for the latest version.

Abstract

Digital media platforms serve as critical channels for firms, consumers and communities. While platforms rely on algorithms to maximize engagement and drive network effects, we know little about how various online interactions affect user contributions. I develop and deploy autonomous generative AI agents in a large-scale field experiment on a major social media platform, randomizing users' exposure to distinct types of interactions to causally identify their effects on subsequent contributions. The results reveal a governance trade-off central to organizational design and the management of digital platforms: controversial interactions spur the largest increase in immediate replies and thread growth but suppress posting in the following week and month. In contrast, informative interactions better retain contributors over time, increasing weekly retention by approximately 4% and total contributions by 5%. These effects vary by user experience, with informative agreeable responses strengthening retention among new contributors and informative disagreeable responses broadening the contributor base. Overall, my findings show that optimizing for engagement can come at the expense of lasting contributions, demonstrating how design choices shape organizational performance and competitive advantage. By cultivating informative interactions, platforms can sustain contributors and foster more inclusive digital ecosystems with higher participation and higher-quality contributions.

*Correspondence to wajeehaa@stanford.edu. I am especially thankful to Erik Brynjolfsson, Chuck Eesely, and Ananya Sen for invaluable feedback and advice. I also thank Matthew Gentzkow, José Ramón Enríquez, Molly Offer-Westort, David Holtz, Ramesh Johari, Kathleen M. Eisenhardt, Alex (Sandy) Pentland, Doughlas Guilbeault, Austin van Loon, Samuel Goldberg, Tobias Kretschmer, Anil Doshi, Michael Yeomans, Manish Raghavan, Wesley Koo, Linus Dahlander and Johannes Luger for helpful comments and suggestions. I thank Project Liberty, the Stanford Institute for Human-Centered AI and the Stanford Digital Economy Lab for financial support.

1 Introduction

Digital media platforms - firms that host and monetize content created by their users - have emerged as some of the most influential companies in the world. With the largest user bases globally,[1] such platforms significantly shape the flow of information that underpins modern economic, social and cultural life. For businesses and entrepreneurs, these platforms serve as essential channels to reach their target markets. Yet, despite their central role in digital ecosystems and the global economy, we know little about the types of online interactions these platforms are structurally incentivized to promote in order to grow their user bases.

Platforms create value by facilitating exchanges between users, which requires maintaining a sufficient flow of content contributions by users to be able to initiate and sustain network effects among participants (McIntyre and Srinivasan, 2017; Lee et al., 2015; Loh and Kretschmer, 2023). A foundational literature in platform governance establishes that network effects create increasing returns (Shapiro and Varian, 1999; Parker et al., 2016; Cusumano et al., 2019): the more contributions and contributors exist on the platform, the more heterogeneous knowledge and capabilities it can have to attract more users (Tanriverdi and Lee, 2008; Zhu and Iansiti, 2012; Cennamo, 2018; Li and Netessine, 2020). Given the advertising-driven business model of mainstream digital media platforms, maximizing user retention and contributions is a key strategic goal since it provides advertisers with greater touch points to reach the broadest group of people.[2]

Presently, platforms use algorithms to manage content and drive network effects, optimizing for immediate engagement, e.g. in the form of replies, as proxy for user retention (Cunningham et al., 2024; Moehring, 2024). Prior work on the effects of social media algorithms has examined the types of online interactions promoted by existing algorithmic systems (Brady et al., 2017; Vosoughi et al., 2018; Levy, 2021). Previous research has also examined how such interactions affect user's immediate engagement (Bhargava, 2023; Beknazar-Yuzbashev et al., 2022, 2024; Piccardi et al., 2024) and societal outcomes, often finding that social media platforms maximize engagement at the expense of users' preferences and well-being (Allcott et al., 2020; Milli et al., 2025; Rathje et al., 2023). However, it is unclear how effective different types of online interaction are in retaining users' contributions on the platform. In other words: Are the type of interactions that maximize immediate engagement similar to those that can sustainably grow user contributions in the longer-term?

This tension between immediate engagement and sustained user contributions mirrors long-standing concerns about myopic organizations and corporate behavior (Stein, 1989; March, 1991; Levinthal and March, 1993), and the risks of relying on proxy metrics (Hohnhold et al., 2015;

[1] Overall, social media platforms have 5.24 billion global users, which is the majority of all internet users and the global population (DataReportal et al., 2025).

[2] As an example, in June 2023, thousands of communities on Reddit staged a "blackout" protest, effectively halting all content creation on those communities. With less content to browse, daily user traffic dropped by about 6-7% and daily visits to Reddit's ad portal fell approximately 20% during the blackout period, indicating advertisers' concern about reduced user engagement (Singh, 2023).

Athey and Luca, 2019). Additionally, understanding which interactions foster durable contributions addresses an essential gap in the platform governance and management literature. While digital platforms play a central role in distributing information for valuable exchanges (Farronato, 2025; Altay et al., 2025), it remains unclear whether informative interactions stimulate contributions more effectively than the provocative or polarizing ones often promoted by engagement-maximizing algorithms (Brady et al., 2017; Rathje et al., 2021; Levy, 2021; Van Bavel et al., 2021). Algorithms could be amplifying fast, emotional responses (System 1) on account of being more engaging while neglecting deliberative interactions (System 2) (Kahneman, 2003) that spur longer-term contributions over time. By testing these dynamics, I show whether current algorithmic approaches amplify the interactions most aligned with platforms' strategic growth goals. Moreover, examining users' behavioral responses reveals whether there is a trade-off between interactions that help platforms retain contributors and those that users prefer, informing the debate on platforms' incentives to self-regulate.

In this paper, I conduct a large-scale field experiment to causally disentangle the effects of exposure to different types of interactions on users' follow-on contributions on a major social media platform. To carry out the experimental intervention, I deploy multiple AI agents to generate and post different types of comments at random in response to posts made by platform users. Doing so enables the generation of contextual comments at scale and allows for variation in interactions through the use of different prompts to a large language model to generate content with different key attributes. In this manner, this study isolates the downstream effects of being exposed to various types of interactions on users' subsequent contributions.

Specifically, I measure two key types of user contributions. First, I examine the responses made to the treatment comments as a measure of immediate engagement. This is important since such replies are amongst the key metrics social media platforms optimize for and use to rank content in users' feeds (Mosseri, 2018; Bhargava, 2023; Narayanan, 2023; Cunningham et al., 2024). Second, I measure the effects of consuming the treatment comment on users' future contributions on the platform in the week and month following the treatment. This metric serves as a measure of longer-term user retention: platform businesses predominantly monitor metrics associated with user acquisition and retention such as *monthly active users* (MAU) and *weekly active users* (WAU), which serve as indicators of long-term business sustainability and monetization potential (Meta Platforms, Inc., 2024; Reddit, Inc., 2024; Twitter, Inc., 2022). Measuring user contributions in terms of both immediate engagement and longer-term user retention has the benefit of measuring whether trade-offs exist between the two metrics.

The experiment proceeds in two stages. The first stage consists of an adaptive experiment aimed at learning the most meaningful categories of content as determined by platform users. Given the attributes of social media interactions suggested in prior work, i.e. informativeness (Altay et al., 2025; Barberá, 2020), polarity i.e. agreement (Allcott et al., 2020) or disagreement (Asimovic et al., 2021), and controversiality (Rathje et al., 2023), I devise a list of eight initial treatment arms by combining these attributes together. Based on the results of this adaptive

learning phase, the types of interactions that receive the highest scores from users are those that are controversial, informative and in agreement with the treated users' perspectives, and informative and in disagreement with the treated users' perspectives. Once these most effective treatments arms are identified, I use them in the second stage – a randomized field experiment that exposes users to these interactions at random in the form of a comment alongside an uninformative treatment arm, which serves as a neutral benchmark. I then measure the effects of such interactions on users' downstream contributions.

The results reveal that controversial comments receive the greatest amount of immediate engagement, i.e. users engage more and engage in longer discussions in response to controversial comments. This provides experimental evidence in the field, building on prior correlational studies and lab experiments suggesting that provocative content is more likely to gain attention and engagement (Berger, 2011; Berger and Milkman, 2013; Brady et al., 2017; Rathje et al., 2021). However, receiving informative and agreeable responses stimulates future contributions by new users in the longer-run. Furthermore, receiving informative and disagreeable responses stimulates longer-term content production by older users. These results are economically significant with *controversial* arm leading to 9% fewer newer users retained relative to the *informative agree* arm in the week and the month following the intervention. This suggests that platforms optimizing for content production in the short-run, as done by several social media ranking algorithms (Meta, 2025; Narayanan, 2023), can impede the growth of user contributions in the longer run, which is valuable for platforms to sustainably grow user contributions. In addition, receiving informative responses, particularly informative and disagreeable ones, increased the quality of posts made by treated users during the week and the month following treatment. This illustrates that optimizing for the most engaging treatment arm in terms of replies would also degrade the quality of user contributions in the longer-run.

I uncover substantial heterogeneity in content production by the type of user. While controversial content received more immediate replies by new and old users alike, older, more experienced users of the platform replied more often to such interactions in the 24-hour period following the treatment. However, in terms of user retention, new users who received informative and agreeable responses increased their contributions in the week and month following the intervention. This result holds also for new users who were exposed to the treatment, i.e. users who engaged on posts randomized to receive informative and agreeable interactions increased their contributions content production by new users in the longer-term. The quality improvements in new posts made following the treatments are also driven by the newer users on the platform. This suggests that while older users may immediate react to controversial interactions, newer users' contributions are better retained in terms of both quantity and quality when exposed to informative interactions.

This paper builds on classic theories of organizational myopia (Stein, 1989; March, 1991; Levinthal and March, 1993) by identifying a behavioral mechanism through which algorithms amplify System 1 responses — fast, emotional reactions to controversy — while neglecting the

System 2 dynamics of deliberation and informativeness that drive retention and quality over time (Kahneman, 2003). I extend the bounded rationality lens (Simon, 1958) by demonstrating how algorithmic delegation introduces new limits: proxy metrics and short feedback cycles that obscure long-run organizational consequences.

I contribute to platform governance research (Boudreau and Hagiu, 2009; Tiwana, 2013; Cusumano et al., 2019) by shifting the focus from "how much engagement" to "what kind of engagement" scales ecosystems, showing that platform growth depends on cultivating interactions that sustain contributions rather than those that simply trigger immediate reactions. While prior research has demonstrated that platforms promote low-quality or toxic interactions because they reliably boost engagement (Brady et al., 2017; Rathje et al., 2023; Beknazar-Yuzbashev et al., 2022, 2024), I show that this is a fundamentally myopic strategy: such interactions suppress downstream user contributions and erode longer-term growth. In doing so, the study challenges the prevailing assumption that engagement is a sufficient proxy for performance and shows instead that sustaining growth depends on interactions that foster retention and higher-quality contributions over time.

Additionally, this work contributes to research on online communities (Faraj et al., 2011; Majchrzak and Malhotra, 2013; Levina and Arriaga, 2014) by introducing a life-cycle perspective, showing that new users are best retained through informative agreement while experienced users contribute more following informative disagreement. Finally, I contribute to work on the societal effects of algorithms by providing causal evidence that diversity of opinion and higher-quality contributions can be sustained without sacrificing participation.

Findings from this study can help to inform managerial decisions regarding platform governance and growth. For platform managers and designers, the study suggests that promoting interactions that are both informative and affirming — especially for new users —can be a viable strategy for increasing retention and cumulative content generation, even if such content underperforms on short-run engagement metrics. Strategically, this invites a re-evaluation of how platforms define and optimize for "success" in promoting user retention. If immediate engagement metrics like replies are imperfect proxies for future contribution, then new measurement frameworks and incentive structures are needed to govern platform growth and quality. Specifically, platform owners and managers may wish to find ways to promote informative, agreeable content in order to grow new users' content production and similarly, expose old users to informative, disagreeable content to grow their contributions. This could be achieved by ranking such interactions as higher or by providing users with generative AI tools to produce such responses to other users, thereby spurring follow-on contributions by users.

The remainder of this paper proceeds as follows. Section 2 provides an overview of the related literature and how this paper contributes to various streams of the literature. Section 3 describes the empirical setting and the key experimental intervention, including the attributes of the AI agents developed to carry out the intervention. In Section 4, I outline the design and results of

the first stage of the experiment, i.e., the adaptive field experiment. Section 5 then outlines the design and results of the second stage, i.e., the randomized field experiment. Finally, Section 7 concludes with a discussion of the findings and their implications for platform governance and the management of online communities.

2 Related literature

This research is related to four strands of the literature. As platform-mediated ecosystems become central to economic, social, and informational exchange, scholars have increasingly turned their attention to the question of how platforms govern participation and structure incentives for user engagement and contribution (Boudreau and Hagiu, 2009). Prior work has examined how platforms govern their ecosystems to establish dominance by attracting the largest shares of contributors (Adner et al., 2019; Tiwana, 2013; Cusumano et al., 2019) through design and incentives (Hagiu and Wright, 2015). Panico and Cennamo (2022) and Ploog (2024) show that platform success depends not only on leveraging network effects but also on governing them thoughtfully to sustain value creation. This study shows that not all types of interactions contribute equally to network value. By identifying the quantity and quality degrading effects of controversial interactions — despite short-term engagement increases — this work moves the literature from thinking about "how much engagement" to "what kind of engagement" corresponds to network growth. This research adds behavioral nuance to theories of platform scaling (Zhu and Iansiti, 2012) and provides an empirical extension to theories of orchestrating value creation on platforms (Rietveld et al., 2019), aligning with calls for more empirical grounding in governance research (Halckenhaeusser et al., 2020).

Previous research on social media platforms finds that exposure to toxic, emotionally-charged and lower quality interactions helps platforms keep users engaged (Milli et al., 2025; Moehring, 2024; Beknazar-Yuzbashev et al., 2022, 2024). Prior work argues that people seek out and engage in online interactions that evoke high-arousal emotions (Berger and Milkman, 2012) such as moral and emotional content (McLoughlin et al., 2024; Brady et al., 2020) or content that is acrimoniously unpleasant (Shugars and Beauchamp, 2019) since it is likely to generate more social engagement. While these papers examine the impact of specific types of content on user's short-term engagement, they ignore longer-run impacts on users' contributions. This research builds on these findings in two key ways: first, by experimentally evaluating the effects of different types of online interactions (e.g. controversial, informative, etc.) altogether, and second, by integrating longer-run behavioral tracking of users. In this manner, I uncover alternative pathways to stimulate user contributions, finding that interactions that are informative and agreeable can sustain longer-term contributions from new users. In doing so, this research demonstrates that interactions that maximize immediate engagement can actually undermine the longer-term retention of new user contributions, filling a blind spot in prior empirical research and highlighting trade-offs central to platform strategy.

Second, a foundational concern in prior literature is that firms systematically favor short-term gains at the expense of long-term value. Seminal theoretical work by March (1991) and Levinthal and March (1993) shows how organizations become trapped in exploitative learning loops, reinforcing near-term strategies while neglecting exploration. Stein (1989) offers a foundational model of myopic corporate behavior, showing that even when markets are rational and not fooled by inflated earnings, managers still act myopically by treating market expectations as fixed – leading to inefficient decisions. These frameworks emphasize how performance measurement and short feedback cycles can entrench organizational myopia.

More recent studies further reinforce this logic based on pressure for visible quick wins (Repenning and Sterman, 2002), aggressive scaling in platform markets (Cennamo and Santaló, 2013), pragmatic compliance (Christin, 2018), poorly designed metrics (Wu, 2025), and reduced friction (Mostajabi et al., 2024) – all of which lead to the pursuit of aggressive short-term strategies that distort innovation and can backfire in the long-run. In platform contexts, researchers have warned against A/B testing strategies that generate short-term metric gains (clicks, conversions) at the cost of long-term engagement or trust, likening them to "eating your seed corn" (Hohnhold et al., 2015; Dmitriev et al., 2016). Frick et al. (2023) examines how programmatic advertising platforms optimize measurable short-term gains while eroding partner value over time. While prior work has illustrated that firms optimize for myopic metrics in several cases and proposed the use of surrogate metrics to better predict long-run outcomes (Athey and Luca, 2019; Athey et al., 2019), I show which interventions generate misleading short-term signals in the first place and why. By showing which treatments cause a divergence between immediate engagement and longer-term user retention, this experiment identifies a behavioral mechanism of myopia that operates underneath metric optimization. In doing so, it links organizational myopia to dual-process theory (Kahneman, 2003), showing how platform design can privilege System 1 engagement over System 2 deliberation, entrenching short-term gains at the expense of durable contributions.

Third, this work also related to the literature on online communities and innovation, which identifies three core forces shaping collective knowledge production: governance structure, legal and informational design, and content dynamics. Foundational work emphasizes that online communities function as fluid organizational forms that enable large-scale collaboration through emergent leadership and platform affordances (Faraj et al., 2011; Majchrzak and Malhotra, 2013; Faraj et al., 2015). Within these structures, contributor motivation is driven by intrinsic factors like creative enjoyment and problem-solving (Lakhani and Wolf, 2003), social and symbolic incentives (Gallus, 2017; Levina and Arriaga, 2014), peer-recognition mechanisms (Burtch et al., 2022), and a combination of social cues with monetary rewards (Burtch et al., 2018). From an institutional and design perspective, external legal regimes, e.g., copyright restrictions (Nagaraj, 2017) and informational scaffolding also matter with information seeding being shown to both dampen (Nagaraj, 2021) and improve user participation (Aaltonen and Seiler, 2015) based on the context. At the same time, prior research has shown the factors that can distort collective

outputs such as competing internal logics (Christin et al., 2024), exposure to incivility (Kim et al., 2021) and the presence of a small set of supersharers (Baribi-Bartov et al., 2024).

This research bridges the work on intrinsic motivations and the extrinsic environment by showing that the interactions users are exposed to acts as an endogenous incentive, shaping not just what users do, but when and how long they do it. This extends prior work by introducing a temporal and user-lifecycle lens to dimensions previously underexplored in the literature. Additionally, where (Gallus, 2017) and (Burtch et al., 2022) found that recognition systems disproportionately motivate newcomers, this work reveals that different types of content resonate differently across user tenure. While both new and old users respond more strongly and quickly to controversy, the types of interactions that grow sustained contributions are different. Whereas newer users are more influenced by informative and agreeable content over time, older users contribute more in the month following the treatment when exposed to informative and disagreeable content. Additionally, informative and disagreeable content also increases the quality of user posts in the week and month following the intervention; this change in the quality of contributions is driven by newer users. This nuance advances our understanding of how platform designers might tailor content strategies to different user segments to optimize user retention, quality or engagement.

Finally, this research is also related to a growing body of work on the societal effects of algorithms and social media. Previous studies have documented that there exists ideological segregation on social media (González-Bailón et al., 2023), which exposes users to politically like-minded content (Nyhan et al., 2023) and is less likely to supply individuals with posts from counter-attitudinal outlets, conditional on individuals subscribing to them (Levy, 2021). This work suggests that since users respond to sources that are ideologically aligned (Prior, 2013) and recommender systems respond to user interests, there can be a feedback loop where both recommendations and user interests progressively narrow (Stoica and Chaintreau, 2019).

Experimental studies further show that personalized algorithms increase engagement but reduce the diversity of the content consumed by individual users (Holtz et al., 2020; Anderson et al., 2020). However, key user metrics such conversion to subscriptions and retention on the platform are strongly associated with greater content diversity (Anderson et al., 2020). Previous research also finds that social media algorithms can be figured to reduce polarization (Piccardi et al., 2024). My work contributes to this literature by illustrating the heterogeneous impact of online interactions that agree and disagree with the treated users' perspectives based on the type of user. By showing the value of informative and disagreeing interactions in retaining old users' contributions and improving the quality of contributions by new users, this study experimentally shows that diversity of opinion and the quality and quantity of user contributions can go hand-in-hand when examining outcomes over a longer time horizon than immediate engagement. Moreover, by retaining a greater number of active new user contributors through exposure to informative and agreeable interactions, platforms can improve the overall representativeness and diversity of contributors and counter the distorting effects of supershares in shaping public

discourse (Kim et al., 2021; Baribi-Bartov et al., 2024).

3 Design of the Experimental Intervention

3.1 Overview

To address the question of what content contributes to platform growth, I use generative AI agents to automatically post different types of content in the form of comments responding to users' posts. Each type of comment constitutes a separate treatment arm. To create the generative agents, I developed a novel agent architecture that leverages a large language model (GPT-4o) and hundreds of real user profiles alongside capabilities to autonomously interact with web pages and post comments.

The experiment proceeds in two stages. In the first (learning) stage, I use an adaptive treatment assignment algorithm that, over time, gives higher treatment assignment probabilities to treatments that perform well, i.e. treatment arms with comments that are perceived to be higher quality by users, as measured by comment scores. After identifying the most effective interventions during the learning stage, the second (evaluation) stage estimates the treatment effects of each arm as part of a field experiment. This second stage allows for random variation in peoples' exposure to the *type* of content using generative AI agents to post comments in a scalable and contextual manner. To measure the effects of being exposed to different types of content, I then measure people's engagement with the posted comments and follow-on activity on the platform.

This study was reviewed by the Stanford University Institutional Review Board (Protocol No. IRB-68938), and preregistered at the American Economic Association's Registry under AEARCTR-0014722.[3]

3.2 Setting

The experiment takes place on a large social media platform with more than 50 million daily active users and more than 100,000 virtual communities, which host user-generated content focused on a particular topic. This provides an ideal setting for this experiment as it offers a large user base, multiple diverse topical communities, and hosts public content that can be viewed by anyone and accessed via an API.

Users can post two main types of content on the platform: (i) posts that introduce a given topic, discussion, or questions and (ii) comments that respond to specific posts. I select 60 different communities for the experiment where AI-generated responses could both be helpful for users while meeting any community-specific rules.

[3] https://www.socialscienceregistry.org/trials/14722

3.3 Intervention

Each treatment consists of a comment posted in response to a given post by a generative agent using a large language model (GPT-4o). An example of how an AI-agent interacts with users is shown in Appendix Figure 6.

I use AI-generated comments for the intervention, as opposed to human-generated comments, for several reasons. First, while manually writing replies to posts is time-consuming, costly and difficult to scale, generative agents offer a scalable and cost-effective way of deploying such interventions, allowing for thousands of posts to be included in each treatment group. Second, the treatment comments can always be relevant and contextual, as relevant details about the post in question can be passed into the model prompt used to generate a comment, whereas humans posting comments may not always have relevant information at hand while responding to a post on any given topic. Third, while human subjects recruited to carry out the intervention by posting comments manually may respond differently to a given prompt based on their individual preferences, mood or energy levels, comments generated by an AI model using structured prompts are arguably more consistent in their output, allowing for robust testing of the impact of posting different types of comments.

3.4 Agent profiles

Each comment is posted using a specific, randomly selected agent profile or account. With the goal of creating realistic user profiles and keeping the quality of profiles constant across treatment arms, I created over 300 unique user profiles using real email addresses, device and browser fingerprints and IP addresses. I took several steps to make the user profiles used for posting comments as similar as possible to typical user profiles on the platform to increase the external validity of the study and mitigate any biases that lead to users flagging the accounts used in the study as bot accounts. Specifically, each profile had its own unique username, gender and interests on the platform. Additionally, I assigned a personality to each profile so that any activity on the platforms sounds casual and similar to the communication style of regular users. Finally, the accounts used for posting comments were created in waves, resulting in the accounts having ages ranging from a few months old to about 2 years of age, which allowed the accounts to come across as distinct and in some cases, as experienced users.

3.5 Agent architecture

The agent architecture consists of a few key components. These include an *agent orchestrator* that coordinates and assigns which agent profiles, communities and in some cases, prompts to use in a given session. Upon activation, the *agent orchestrator* randomly selects a given agent profile or account for use in a given round. It typically selects accounts that have been inactive for at least the past 12 hours to avoid raising suspicion by posting comments from the same account continuously. In the case this architecture is used for a randomized field experiment, the *agent orchestrator* also randomly selects a prompt from a list of prompts that can be used to

generate a specific type of comment. Using any agent profile to post any type of comment allows me to resolve potential endogeneity resulting from the account profiles and separately identify the impact of the type of comment posted. Finally, the *agent orchestrator* selects a community at random from a given list of eligible communities, and passes it on to the *post selector*.

The *post selector* chooses which post to use in a given session. After receiving a randomly selected community from the *agent orchestrator*, the *post selector* randomly selects a post from the top ten new posts on the community and checks if it meets specific requirements. These include checking that the post is not authored by a moderator or a user who was previously treated in the experiment, is only a text-based post (i.e. contains no images), has a sufficient number of words for a response to be relevant and has content that is suitable for being included in the experiment. I take a three-step approach to filter posts for suitable content. First, I excude all posts that are marked by the platform to be Not Safe For Work (NSFW).[4] Second, using Open AI's Moderation API and Google's Perspective API, I exclude posts that do not meet moderation requirements. This step removes posts containing potential harmful or toxic content from being included in the experiment. Finally, as an additional step, I use the Open AI API to do a binary classification of the posts' text into whether or not it is about a sensitive topic. If the randomly selected post doesn't meet any of these requirements, the post selector randomly selects another post from among the top ten new posts on the randomly selected community. If, after five attempts, no suitable post is found on the community, the post selector coordinates with the agent orchestrator to select another community at random and repeat the post selection process until a suitable post is found to be passed on to the *generator*.

The *generator* and *processor* work together to generate a specific comment and process it for use. The *generator* takes as input a given post and an instruction, and uses it to generate a comment in response to the specific post. To make the agent's response contextual, the posts's title and text is injected into the model prompt. The generated comment is then passed on to the *processor*, which checks if it meets specific requirements. These include the same moderation requirements applied to the post's text, i.e. passing each generated content via Open AI's Moderation API and Google's Perspective API to ensure that it does not contain any harmful content. Additionally, to mitigate any biases and improve the external validity of this work, it is important that the comment does not self-identify as an AI or any of the personalities assigned to it. If either of these two requirements are not met, the *processor* flags the comment and asks the *generator* to regenerate the comment until it passes these checks. The processed comment is then passed on to the *browser automation* component, which navigates to the relevant webpage and posts the comment. Finally, details of the post selected and the generated bot comment are stored in a database, where a separate script is used to update the database with outcomes data 24 hours and seven days following treatment, respectively.

After a comment is posted, the same process described above is repeated with a new ran-

[4]NSFW is used as a warning to indicate that the content may be inappropriate or explicit for certain environments, such as workplaces or public spaces.

domly selected agent profile. Doing so ensures that each agent profile is used to post multiple different types of comments over the course of its lifespan, which further makes the accounts used in this study appear more realistic. Moreover, the modularity of this agent architecture confers key advantages for configuring field experiments on digital platforms. Specifically, I configure this agent architecture is configured to work in two modes, i.e. randomized and adaptive architectures, to be used for randomized and adaptive field experiments, respectively. Appendix Figure 7 shows the randomized agent architecture, which has the distinguishing feature of using a randomized treatment instruction as part of it's prompt to a language model to generate a comment. Appendix Figure 8 shows the architecture of the adaptive agent. Instead of using a randomized instruction from the *agent orchestrator*, the adaptive agent picks the instruction using an adaptive assignment algorithm that, over time, gives higher assignment probabilities to treatment instructions that perform well based on a specified objective function.

3.6 Design of treatment arms

Each treatment consists of a comment generated using a base prompt that instructs the model to respond to a given post on a specific community using a casual tone. To vary the type of comment posted, I vary the instruction embedded in the model prompt used to generate the comment. Therefore, each treatment corresponds to a specific instruction in the prompt given to the language model.

One dimension I use to vary the type of comment posted is its informativeness. With information being a key driver of economic activity and growth (Romer, 1990; Jones, 1995), and digital platforms playing a key role in distributing information for users to engage in valuable transactions (Farronato, 2025), it is important to investigate how valuable it is for platforms to host informative content. Recent work suggests that social media helps improve knowledge of general news among its users (Altay et al., 2025; Allcott et al., 2024), but platforms relying on user-generated content often have uneven information coverage (Zhu et al., 2020). Since it is unclear how informative content compares to other content in terms of promoting user engagement and future participation, I compare informative content categories with other uninformative content categories in this work.

Additionally, I vary whether the content agrees, disagrees or stays neutral with respect to the perspectives of the treated submissions' author. Recent research shows that content from politically like-minded sources constitutes the majority of what people see on social media (Nyhan et al., 2023) with there being limited exposure to counter-attitudinal news (Levy, 2021) and ideological segregation in people's engagement patterns (González-Bailón et al., 2023). Studies have illustrated the role of social media in both increasing exposure to polarized views (Allcott et al., 2020), and in providing depolarizing diversity (Asimovic et al., 2021). It is unclear whether users will engage more and have greater follow-on participation when confronted with agreeable content that reinforces their views and preferences as some prior work suggests or disagreeable content that may lead to more innovative suggestions or personal fulfillment (Helberger et al.,

2016).

Finally, I examine the impact of posting controversial comments. Prior research shows that high-arousal content such as controversial content generates more engagement on social media (Brady et al., 2017; Rathje et al., 2021). Previous work also documents that people believe divisive content and high arousal content to go viral, but do not want such content going viral (Rathje et al., 2023). Given that people report a preference for educational, accurate and nuanced content (Rathje et al., 2023), I measure how informative content compares relative to controversial content in terms of engagement and user retention. Another reason for including controversial content is that the platform has a ranking filter for controversial comments, i.e. the platform allows users and moderators to rank content by controversy, which suggests that such content is important for the platform to host.

Altogether, I explore the following treatment arms that vary informativeness alongside the aforementioned dimensions: 1. *Uninformative* (i.e. comments that are neutral and uninformative), 2. *Uninformative agree* (i.e. comments that agree with the treated submission's author and are uninformative), 3. *Uninformative disagree* (i.e. comments that disagree with the treated submission's author and are uninformative), 4. *Uninformative controversial* (i.e. comments that are controversial and uninformative), 5. *Informative* (i.e. comments that are neutral and informative), 6. *Informative agree* (i.e. comments that agree with the treated submission's author and are informative), and 7. *Informative disagree* (i.e. comments that disagree with the treated submission's author and are uninformative), and 8. *Controversial* (i.e. comments that are controversial).[5] Section A.3 in the Appendix describes the details of all treatment prompts used to produce each of these types of comments.

4 Adaptive field experiment

4.1 Treatment assignment

The experiment proceeds in two stages. In the first (learning) stage, I implement an adaptive experimental design to learn the best-performing treatments and filter out arms with the lowest treatment effects from the above eight treatments (Offer-Westort et al., 2024; Jobjörnsson et al., 2023; Offer-Westort et al., 2021). This adaptive learning stage has several advantages. First, it allows me to avoid potential power issues that result from examining the impact of all eight treatments individually in a fully randomized field setting.[6] Therefore, the key advantage of this

[5]While I vary all comments on the dimension of uninformative/informative and agree/disagree/neutral, for controversial comments, I use the *uninformative controversial* and only *controversial* dimensions. This is because having both the controversial and informative dimension in the same comment often results in comments being only informative and not controversial based on initial testing.

[6]Carrying out the experimental intervention described in Section 3 on a live platform is inherently risky since the platform can change its policies at any time. Moreover, specific communities on the platform may also disallow the use of generative AI or bots over time. Therefore, it was important to limit the size of the experiment in order to continue performing the experimental intervention and respect the rules of the communities included in the experiment.

stage is that it helps to quickly identify a narrower set of the most meaningful treatment arms, which can then be more thoroughly examined in a second randomized evaluation stage. Second, this method ensures that the selection of the treatment arms is biased in the direction of the highest performing treatment arms, as measured by comment scores. As a result, the treatments selected by the adaptive algorithm will be those that are more likely to be noticed by platform users, and therefore have the potential to drive downstream effects on contributions. Finally, since the platform ranks comments based on the scores it receives, this method also ensures that the types of treatments included in the experiment are those that are of practical value to the platform.

For the learning stage, I use a simple thompson sampling algorithm for treatment assignment. Thompson sampling is a widely used approach in settings where optimizing resource allocation and maximizing rewards is required, e.g., online advertising, clinical trails, recommendation systems, etc. Its simplicity and lack of assumptions on the underlying data distribution (Agrawal and Goyal, 2013) makes it an ideal adaptive algorithm for this setting. In this setting, I use thompson sampling to maximize the probability of selecting treatment arms (i.e. comment categories) that get the highest scores, which are determined by user votes. Voting on this platform serves as a means of rewarding content that contributes to the community and demoting content that does not. Therefore, consistent with prior work (Srinivasan, 2023; Moehring, 2024), I use comment scores as a measure of the quality of a given comment as perceived by the platform's users.

The thompson sampling algorithm maintains a probability distribution for each treatment arm's expected reward, representing the uncertainty about the true reward distribution. At each round, the algorithm samples from these distributions and selects the treatment with the highest sampled value as the treatment to assign in that round. After posting a comment assigned in this manner, the algorithm observes the actual reward obtained from this treatment arm. This corresponds to checking the scores of the comments posted after a 24-hour period.[7] Using the observed reward (i.e. comment scores), thompson sampling updates the probability distribution (posterior) for the selected treatment arm using Bayesian inference. The updated distribution becomes the prior distribution for the next round. This process continues iteratively, with the algorithm balancing exploration (trying out different treatment arms to learn more about their rewards) and exploitation (focusing on the treatment arms that have shown higher rewards so far). The details about the algorithm used are provided in Appendix Section B.1.

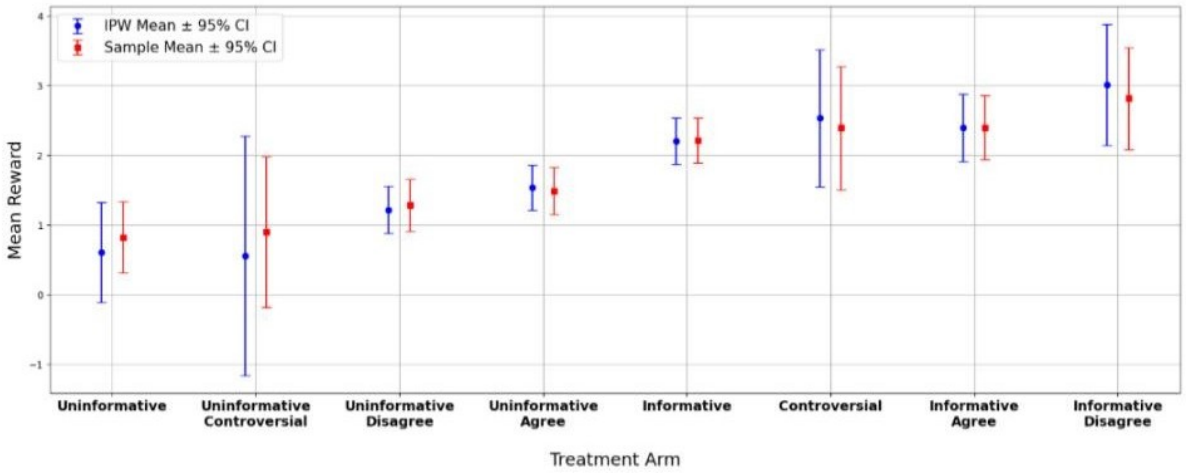
4.2 Learning stage results

Figure 1 shows the mean rewards for each of the treatment arms. While the sampled means represent the average observed outcomes (i.e. number of upvotes) for each arm, the inverse-

[7]Since most interactions on the platform happen within the first few hours of a post's lifecycle (Glenski and Weninger, 2017), waiting for 24 hours offers a reasonable time period to observe the total scores received on a giveb comment.

probability weighted (IPW) means provide an unbiased estimate of each arm's true reward. To see how the treatment assignment probability and cumulative treatment arm selections change over the course of the adaptive experiment, see Appendix Figures 10 and 11.

Figure 1: Sample Means and IPW Means by Treatment Arm



Note: This figure shows the sample mean scores and the IPW mean scores for each treatment arm at the end of the learning stage. The error bars represent 95% confidence intervals. The treatment arms on the x-axis are arranged in ascending order of their sample means.

As shown in Figure 1, the adaptive algorithm reveals the following treatment arms to have the highest sample means: *informative agree*, *controversial*, and *informative disagree* with the *uninformative* arm performing the worst of all eight arms.[8] This learning stage informs the design of the randomized field experiment in Section 5.

5 Randomized field experiment

After identifying the most meaningful treatment arms, i.e. comment categories that are perceived as valuable by the platform users, I evaluate these treatments alongside a control group treatment in the second stage. This second stage consists of a large-scale randomized field experiment, where each treatment arm has equal probability of being selected at random.

5.1 Research design

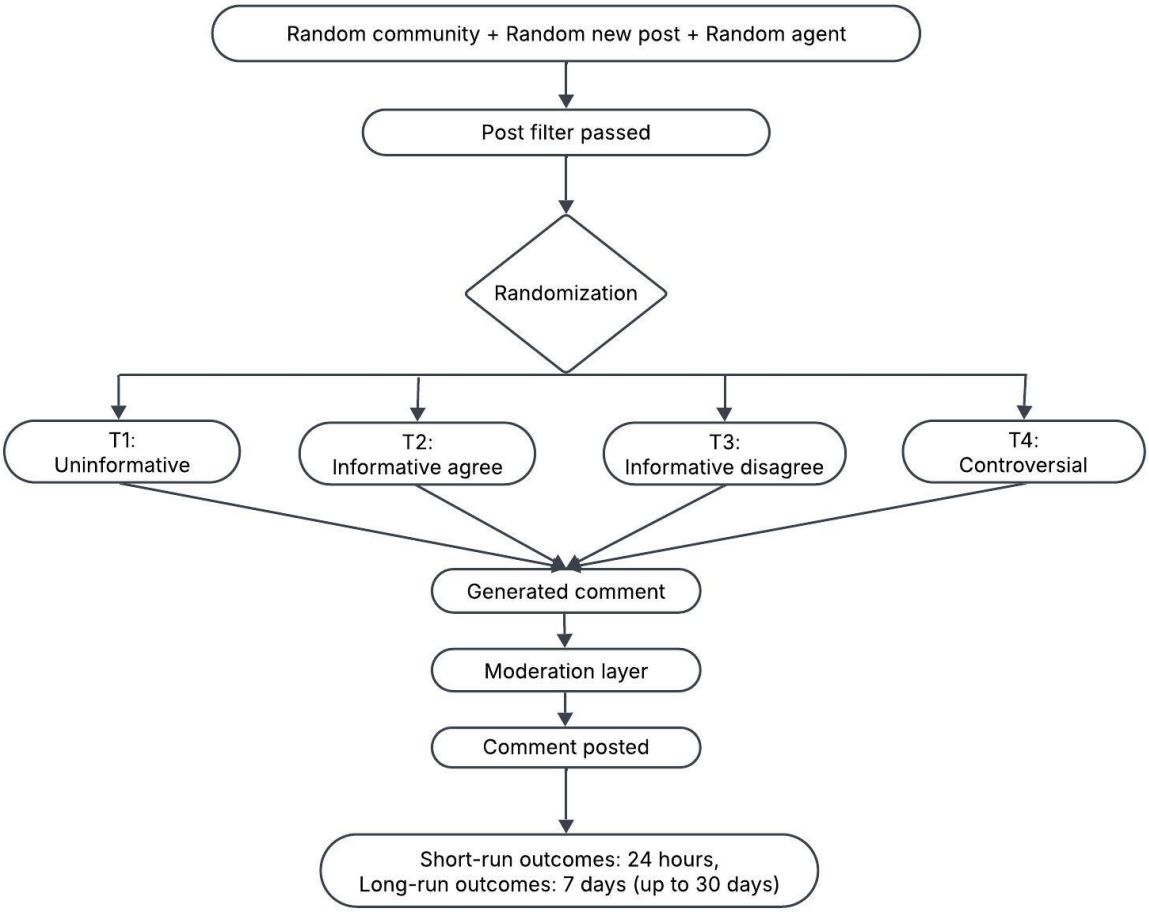
Based on the mean results in Figure 1, I choose the highest performing treatments in terms of comment scores, i.e. the *informative agree*, the *informative disagree* and the *controversial* treatments, as the key treatments to test in the evaluation stage. I perform a series of empirical tests using both computational metrics and human ratings as outlined in Appendix Section C.2

[8]Both the sample means and the IPW means result in the same set of arms as the three best performing arms, i.e. the controversial, informative agree and the informative disagree arms. In prior simulations on which metrics can identify the true best arm(s), using the sample mean for arm selection outperforms IPW estimates (Offer-Westort, 2025).

to validate that these treatment arms differ from one another in the manner indicated in their prompt design.

In addition to the three best-performing treatment arms identified from the learning stage, I include one of the lowest performing treatment arms, i.e., the *uninformative* treatment arm, in the evaluation stage. This treatment is included because of its neutrality and ability to serve as an active control relative to the other treatment arms. Doing so helps isolate the impact of receiving a specific *type* of treatment comment from the impact of simply receiving a comment. Figure 2 shows the design of the randomized field experiment deployed.

Figure 2: Design and Flow of the Randomized Field Experiment Deployed



To capture platform growth, I measure two different types of primary pre-registered outcomes. The first outcome content production in the short run in terms of replies made by users to the treatment comments I post in the 24-hour period following treatment. A period of 24 hours is chosen for this outcome since most comments on this platform typically occur in the first few hours of a post, and nearly all comments occur in the first day (Krohn and Weninger, 2019; Ye et al., 2024). Secondly, I measure content production in the longer run in terms of follow-

on posts made by treated users in the week and month following treatment. This outcome is chosen since platform businesses typically use metrics like *weekly active users* (WAU) and *monthly active users* (MAU) as indicators of long-term business sustainability and monetization potential (Meta Platforms, Inc., 2024; Reddit, Inc., 2024; Twitter, Inc., 2022).

Collectively, both these key outcomes capture growth in terms of the amount of content on the platform, which can then attract users to consume and further contribute to the content posted. Further, following the pre-registration, I also analyze additional dimensions of the treatment comments and users' subsequent replies to the comments. This includes the scores received by the treatment comments. Additionally, I analyze the text of treatment comments and users' replies to them to examine the key differences in the content of the treatment arms and how well received they are by users.

5.2 Results

In this section, I examine the results of the randomized field experiment with treatment arms selected from the learning stage. Appendix Table 3 shows the summary statistics and balance across key post, agent and treated user characteristics for each of the treatment arms. To ensure the treatment comments reflect their prompts, I analyze the text of the comments posted by the generative agents during the course of the experiment. Appendix Section C.2 evaluates the key differences between treatment comments by arm and finds that they reflect the attributes indicated in their respective prompts.

In this section, I first examine the impact of the treatment comments on users' immediate contributions in the form of replies made. I then evaluate the impact of receiving a treatment comment on users' sustained contributions on the platform by each treatment arm.

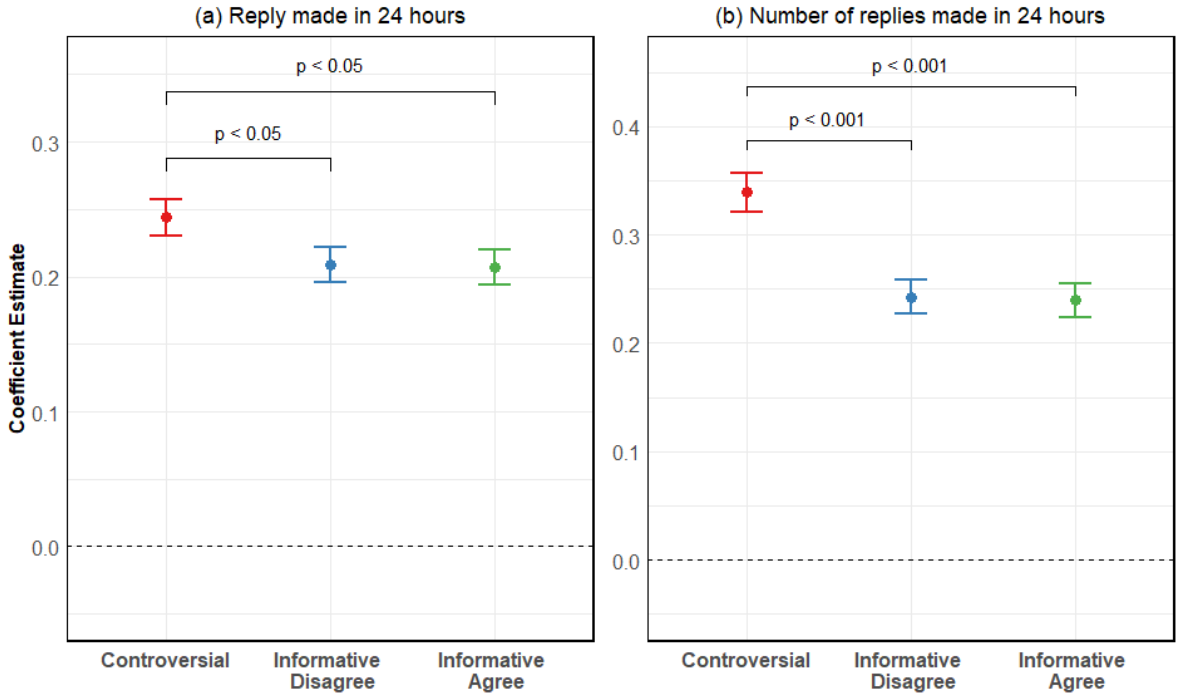
5.2.1 Main treatment effects on immediate content production

I now describe the first main pre-registered outcome on immediate contributions or engagement in terms of replies made to the treatment comments.

First, I find that, while all treatments are more engaging relative to the *uninformative* treatment, controversial comments receive the most engagement both in terms of their likelihood to receive a reply and the number of replies they receive (Figure 3). The *controversial* arm also generates the most engagement in terms of replies relative to the other two treatments, i.e. the *informative agree* and the *informative disagree* arms as shown in Figure 3. These effects are economically meaningful. Controversial interactions lead to greater replies by 24.0% and 22.9% from the *informative agree* and *informative disagree* arms, respectively.

To investigate the quality of the treatment comments, I obtained ratings of a random sample of 500 treatment comments from human evaluators. Additionally, I queried four different large language models (LLMs) to rate the quality of each treatment comment. Figure 4 shows the results. Both informative treatments are rated as higher quality relative to the *uninformative* and

Figure 3: Short-term Content Production by Treatment Arm



Note: These figures plots regression coefficients from OLS regressions on (a) whether the treatment comment receives a reply, and (b) the number of replies received on the treatment comment in the 24-hour period following the treatment. Results are shown for the primary model specification, which controls for pre-treatment characteristics, i.e., the number of comments made by the treatment agent prior to the treatment comment, the number of comments on the treated post prior to receiving the treatment comment and the number of words in the treated post. Data are presented as coefficients with the vertical bars representing 95% confidence intervals derived from robust standard errors.

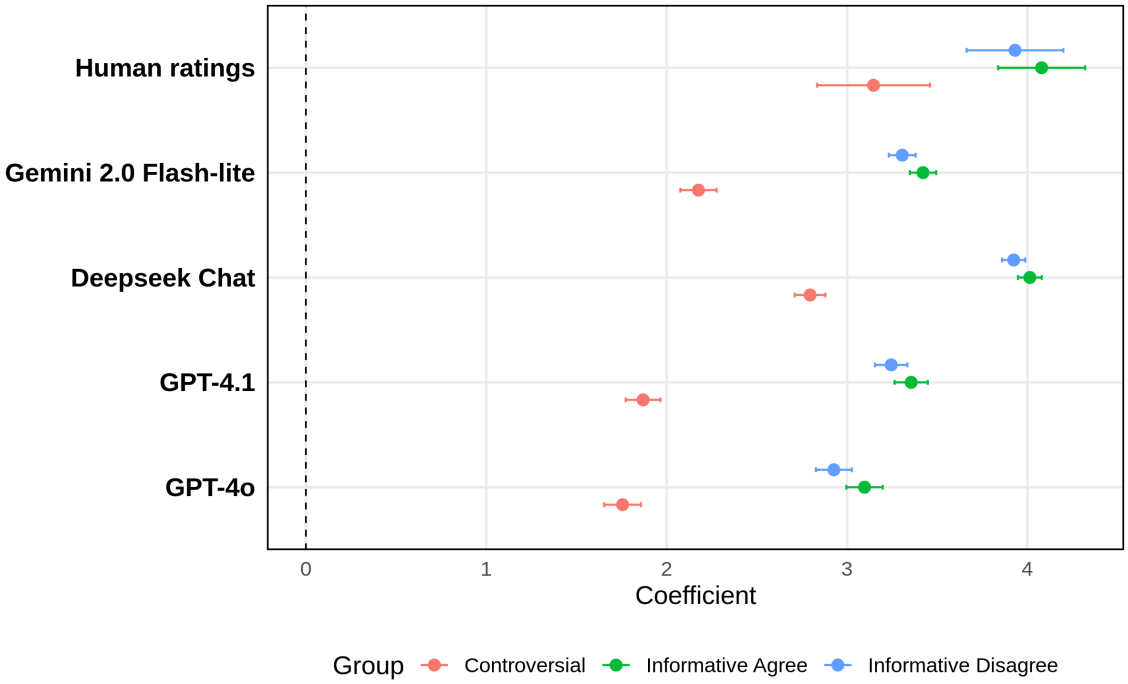
controversial arms. Beyond obtaining these external ratings, I analyzed the text of the replies given by the treated authors to the treatment comments to examine how users perceived each of the treatment arms. I first used an LLM to summarize the response in one or two words and then conduct topic analysis using the top2vec algorithm (Angelov, 2020) to indicate the categories of replies present. Appendix Figure 15 shows the results. These analyses indicate that the treated authors were more grateful to receive comments in both the *informative agree* and the *informative disagree* arms relative to the other two arms, which suggests greater appreciation of such comments.[9]

These results indicate the presence of a trade-off between performance as measured by users' immediate contributions on the platform and the quality of treatments, i.e. higher quality treatments perform worse in terms of drawing immediate user contributions. Conversely, controversial

[9]I also examine the comment scores received by each of the treatment group relative to each of each of the other treatment groups in Appendix Table 9. This pre-registered analysis reveals that while all three treatment arms have higher scores as measured by the platform than the uninformative arm, they are indisguisable from each other in terms of comment scores.

comments draw the greatest number of immediate contributions or replies despite not being as highly rated as the informative treatment arms.

Figure 4: Quality of Treatment Comments by Treatment Arm



Note: These figures plots regression coefficients from OLS regressions on the quality of comments under each treatment arm relative to the *uninformative* treatment arm. *Quality* was measured on a seven-point Likert scale where human evaluators and four large language models were asked "How would you rate the quality of the above reply to the given post?" with responses ranging from *Very low quality* (-3) to *Very high quality* (3). Results are shown for the primary model specification, which controls for pre-treatment characteristics, i.e., the number of comments made by the treatment agent prior to the treatment comment, the number of comments on the treated post prior to receiving the treatment comment and the number of words in the treated post. Data are presented as coefficients with the vertical bars representing 95% confidence intervals derived from robust standard errors.

These findings are consistent with prior work, which shoes that more provocative content receives more engagement based on descriptive analysis of social media and lab experiments (McLoughlin et al., 2024; Brady et al., 2020; Berger and Milkman, 2012).

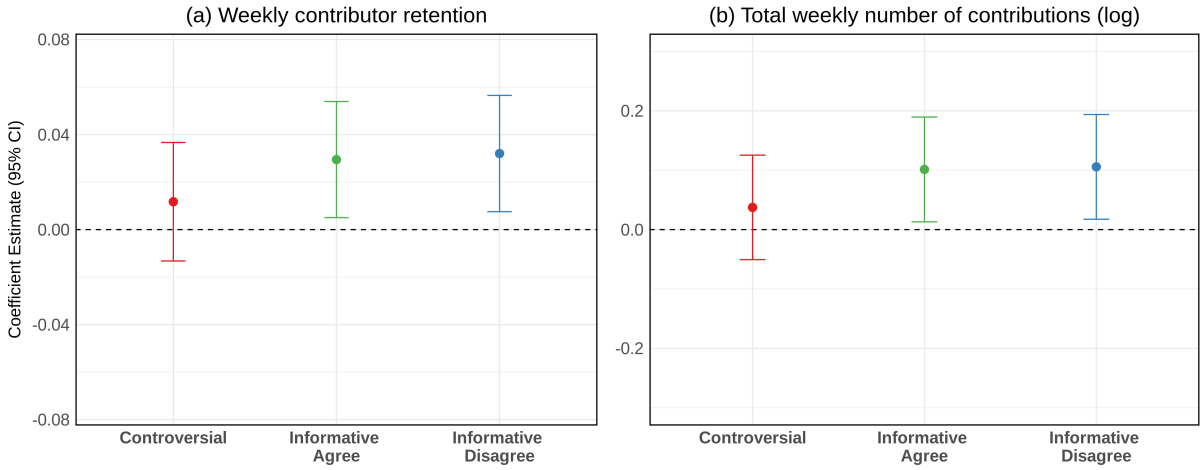
5.2.2 Main treatment effects on sustained contributions

I then examine the effects on longer-run content production, i.e. submissions produced by treated users, as defined by users whose posts received the treatment comments, in the week following the treatment. Figure 5 shows the main effect. While the controversial treatment arm is not distinguishable from the control group (i.e., the *uninformative* arm), both informative treatments increase weekly contributor retention and the number of contributions made on the platform after the treatment. When both the informative arms are pooled together following the pre-registration, as shown in Figure 16, the informative treatments increase the retention

of weekly contributors by 3.95% and the number of overall contributions by 5.18%. Appendix Tables 10 and 11 show the robustness of these effects across different model specifications.

This result marks a meaningful shift from prior work, which has almost exclusively focused on short-term engagement or virality via metrics like shares, likes or replies (McLoughlin et al., 2024; Brady et al., 2020; Berger and Milkman, 2012). My results show that while controversial comments are more engaging in the short-run (Figure 3), they are sub-optimal in sustaining contributors on the platform, which is arguably more important for platforms in maintaining their user base and revenue potential.

Figure 5: Weekly Total Contributions by Treatment Arm



Note: These figures plots regression coefficients from OLS regressions on (a) whether the treated author contributes again on the platform, and (b) the number of contributions they make in the 7-day period following the treatment, i.e. log (x+1), where x is the number of comments made. Results are shown for the full sample ($n = 8,509$). Results are shown for the primary model specification, which controls for pre-treatment characteristics (i.e., the number of comments made by the treatment agent prior to the treatment comment, the number of comments on the treated post prior to receiving the treatment comment and the number of words in the treated post) and includes week and community fixed effects. Data are presented as coefficients with the vertical bars representing 95% confidence intervals derived from robust standard errors.

In addition to the above analyses, I also examine the longer-term effects on content production by other users who were exposed to the treatment comments beyond just the authors of the treated posts. To identify such users, I select those who responded to the treated post within a day of the treatment comment being posted. This measure serves as a proxy for the set of users who were exposed to the treatment since those who post before or longer than a day after the treatment comment is made may have limited visibility to the treatments. This measure also allows for a substantially larger pool of treated users beyond those directly treated on account of being the post authors. An analysis of the posts made by this set of exposed users following the treatment reveals that informative and agreeable comments increase content production in the longer-term, with the effects being similarly driven by the newer users on the platform as shown in Appendix Table 12

Overall, the above findings show that although controversial comments spur greater content production in the short-term, i.e. in the immediate 24 hours following treatment, informative comments lead to greater content production in the longer-run. For new accounts, these longer-run results are detectable both in the week and the month following treatment in terms of original posts or contributions on the platform. While it is primarily the *informative agree* arm that spurs longer-run content production for new accounts, comments in the *informative disagree* arm also contribute to greater longer-run content production relative to the *controversial* arm (Table 18).

5.2.3 Effects on the quality of posts made following treatment

I next examine how each treatment affects the quality of posts made by users in the period following treatment. I operationalize quality using the score given to posts as done in prior literature (Moehring, 2024; Srinivasan, 2023). To prevent any biases resulting from skewed scores, I transform the scores by taking the log of $\max(\text{score}_i, 1)$. This transformation allows for removing skewness in the score data along with not penalizing users for producing posts receiving non-positive scores, since each post produced has the potential to stimulate time spent on the platform, a meaningful goal for the platform. This mean weighted score is calculated as follows:

$$\text{Mean Weighted Score} = \frac{1}{N}\sum_{i=1}^{N}\log\left(\max(\text{score}_i, 1)\right)$$

Table 1: Mean weighted score of posts made following the treatment.

	Posts made within 7 days			*Posts made within 30 days*		
	All	New	Old	All	New	Old
	(1)	(2)	(3)	(4)	(5)	(6)
Controversial	−0.01	−0.01	−0.00	−0.02	−0.03	−0.01
	(0.04)	(0.06)	(0.06)	(0.04)	(0.06)	(0.06)
Informative agree	0.08*	0.11*	0.06	0.03	0.09	−0.03
	(0.04)	(0.06)	(0.06)	(0.04)	(0.06)	(0.06)
Informative disagree	0.08*	0.10*	0.05	0.11**	0.19***	0.03
	(0.04)	(0.06)	(0.06)	(0.04)	(0.06)	(0.06)
Pre-treatment controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	8509	4247	4262	8509	4247	4262

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 1 shows the results of the mean weighted score for posts made by the treated authors in the week following treatment (Columns 1-3) and the month following treatment (Columns 4-6). Both informative treatment arms show muted effects ($p < 0.10$) for increasing the quality of posts made by new users in the week following the treatment relative to the uninformative group. When considering posts made in the month following treatment, comments in *informative*

disagree arm show significant and meaningful effects in increasing the quality of posts made by treated users (Column 4, Table 1). These effects are again primarily driven by new users on the platform.

Combined with the results in the previous section, these results indicate that platform managers and designers have two levels to pull in order to build sustained, high quality user participation. To increase the quantity of contributions by new users, they can encourage or promote the production of informative content that reinforces users' perspectives, and to increase the quality of new user contributions, promoting the production of informative content that disagrees with users' perspectives could be a fruitful strategy.

6 Potential Mechanisms

To examine the factors that could be driving users' immediate and sustained contributions on the platform following the treatments, I examine four potential mechanisms: salience, emotion, social identity and norms.

6.1 Salience

Prior work has shown that simply receiving attention can motivate users to make more contributions online. (Srinivasan, 2023). Across treatment arms, users may perceive longer, more effortful treatment comments as more salient. To examine whether salience drives users' contributions, I interact the number of words in the treatment comments with each treatment arm.

Appendix Table 13 shows the effects of this interaction on users' immediate and sustained contributions. Longer treatment comments result in reduced likelihood of users engaging in the immediate term for the *controversial* (Column 1) and *informative disagree* (Column 3) treatments. These results could be driven by greater cognitive load from longer treatments reducing users' willingness to participate immediately. However, there are no detectable effects of longer treatments on users sustained contributions (Columns 4-6, Appendix Table 13), which suggests that greater salience does not drive sustained contributions.

6.2 Emotion

Given the prevalence of emotional content on social media and its potential for evoking reactions from users (Kramer et al., 2014; Brady et al., 2017; Vosoughi et al., 2018; Brady et al., 2020), I next examine whether the presence of emotional content in my treatment arms drives both immediate and sustained user contributions.

To identify the type of content across all treatment arms, I take an inductive approach. As outlined in Appendix Section , I use a large language model (LLM) to summarize each comment posted using one or two words. I then conduct topic-analysis using the top2vec model again (Angelov, 2020) on the LLM-generated comment summaries. Appendix Figure 14 shows the

types of comments posted under each arm. This analysis reveals that presence of sarcasm as a category of comment type that varies across treatment arms.

To identify whether this emotion drives users' contributions, I interact the presence of sarcasm with the treatments.[10] Appendix Table 14 shows the results. As shown, the presence of emotional language drives immediate replies (Column 2), but it does not drive users' sustained contributions (Column 4) on the platform.

6.3 Social Identity

I next examine the types of users who participate in response to the treatment comments along a pre-registered dimension of heterogeneity: their experience or age on the platform. I find that while the controversial treatment comments receive more replies from both new and older users, they receive more replies from older, more experienced users of the platform (Appendix Figure 17). This result is consistent with theories of social identity, indicating that exposure to controversy activates identity-protective behavior among established members, who then increasingly engage immediately to reaffirm community norms and values (Tajfel and Turner, 2001; O'Reilly and Chatman, 1996).

Additionally, the treatment effects for users' sustained contributions also mask pre-registered heteregeneous effects by users' age. By splitting the sample of treated users into two pre-registered groups, new accounts (i.e. accounts that were younger than the median treated user account age at the time of the treatment) and older accounts (i.e. accounts that were at least the median account age at the time of the treatment), I find that the *informative agree* treatment causes a meaningful shift in posting behavior for newer accounts on both the extensive and intensive margins (Figure 18). In other words, users treated with informative comments that agree with their perspective were more likely to post one or more submissions in the week following the treatment relative to all three remaining treatment groups, including the controversial group (Figure 18a). Specifically, informative agreeable interactions increase the retention of new contributors who make original posts by 8.0% in the week following treatment and 5.6% in the month after treatment.[11] Informative, agreeing responses may signal to new users that they are welcomed and respected – satisfying belonging needs and motivating continued participation (Baumeister and Leary, 1995; Srivastava et al., 2017).

Additionally, informative disagree interactions lead to greater retention of contributors in the month following the treatment as shown in Appendix Figure 19 and Appendix Table 18. Moreover, informative disagreeable interactions increase overall original posts by 5.2% in the month after treatment. The increase in contributions from more experienced users after receiving informative disagreement may stem from a desire to restore and reaffirm their sense of competence

[10] Only the *controversial* and *uninformative* treatment arms are found to have some degree of sarcasm. Therefore, this analysis excludes the two informative treatment arms.

[11] These effects for new accounts are robust to a period of 30 days following treatment with informative and agreeable comments leading to longer-term conduct production for newer accounts as shown in Appendix Table 18.

and identity (Steele, 1988). These effects are robust to community and week fixed effects, as shown in Appendix Table 17.

6.4 Norms

I then examine whether the treatment comments affect user contributions based on how consistent they are with the norms of the community in which they get posted.

To determine the community norms, I constructed a weekly panel of comments on the platform across all 63 communities spanning from November 1, 2024 to May 20, 2025. For each week, I drew a simple random sample of 1,000 comments per community. In total, this yielded 1,133,600 comments. I first summarize each comment using an LLM (GPT-4o) and then induces topics from those summaries using the top2vec topic modelling algorithm (Angelov, 2020). Appendix Figure 9 shows the summary of resulting topics prevalent across the entire corpus.

I then examine whether the prevalence of agreement, encouragement or positive support on the community affects sustained contributions by both new and old users. In communities where members tend to agree with one another, an informative agree reply aligns with that communicative norm. Appendix Table 19 shows that in communities where such content is already prevalent, the *informative agree* treatment increases sustained weekly contributions after the treatment for new users and decreases it for older users. For new users, this norm-consistent feedback may signal that they fit in, increasing their participation (Kristof, 1996; Chatman et al., 2014). For older users, who already understand these norms, the same feedback adds little new social information and could reduce motivation to re-engage (O'Reilly and Chatman, 1996).

7 Discussion

While existing algorithms are optimized to expose users to content that maximizes their engagement (Mosseri, 2018; Bhargava, 2023; Narayanan, 2023; Cunningham et al., 2024), little has been known about the types of online interactions that can help platforms sustainably grow more contributions from users. Using a large-scale field experiment on a digital media platform, this study provides causal evidence on how different types of online interactions affect both immediate engagement and the retention of contributors. I show that while controversial interactions lead to the greatest immediate increase in user responses, informative responses retain users better in the week and the month following treatment. Specifically, controversial interactions engage 1.4 times more users and receive 1.9 times more replies relative to control.[12] However, informative interactions increase the contributor base by 3.9% and overall contributions by 5.1% in the week following the treatment whereas controversial interactions have no detectable effect on sustained contributor retention.[13]

[12] In contrast, informative interactions engage 1.2 times more users and receive 1.3 times more replies relative to control.

[13] Informative and agreeable interactions increase the retention of new users by 8.0% and 5.6% in the week and month after treatment despite leading to 25.5% fewer replies on the same day. Furthermore, informative and

These results are particularly important for platforms focusing on expanding the scale of new contributors during the early stages of their growth (Tiwana et al., 2010).[14] Additionally, given that most projects depending on user contributions fail by rarely attracting more than one contributor (Healy and Schussman, 2003) and platforms hosting user-generated content commonly suffer from a decline in user participation over time (Halfaker et al., 2013), sustaining contributions from users is also important during the later stages of a platform's growth (Tiwana et al., 2010; Healy and Schussman, 2003).

From a management perspective, the study speaks to an unresolved tension at the heart of platform governance: engagement maximization versus contribution sustainability. Digital platforms — especially those hosting user-generated content — are not merely venues for interaction but dynamic ecosystems whose growth and sustainability hinge on complex governance choices. At the same time, platforms face a fundamental strategic dilemma: how to balance short-term engagement metrics with long-term ecosystem health. Scholars have long warned of the dangers of organizational short-sightedness—such as optimizing for easy-to-measure KPIs at the cost of sustainable value creation (Laverty, 1996; Athey, 2018). In digital contexts, this short-termism often manifests in the design of recommendation algorithms and ranking functions that prioritize immediate clicks and replies over durable user retention and content production. Yet few empirical studies have tested this trade-off directly, especially in real social environments.

This study fills that gap. Using a novel field experiment powered by generative AI agents, I causally identify how different types of interactions — uninformative, controversial, informative-agreeing, and informative-disagreeing — affect both short-term engagement and long-run user contributions. This work advances the literature on platform governance by showing how micro-level content strategies shape macro-level outcomes like user retention and the growth of contributions over time. By showing that the type of interactions users are exposed to affects their likelihood of contributing in the future, the findings sharpen our understanding of how platforms can shape ecosystem outcomes.

In doing so, it introduces a novel behavioral lens to platform governance literature, i.e., governance through informational stimuli. I show that interaction type is a lever on cognitive mode. Controversy and emotive mirroring (System 1) spark instant replies but depress future contribution; informative prompts - both agreeable and disagreeable (System 2) - reduce instant activity yet increase return and quality. This advances dual-process theory (Kahneman, 2003) by linking processing mode to long-horizon behavior in a live ecosystem, and identifying a choice-architecture mechanism whereby designers can shift users toward System 2 to build durable value. Thus, rather than governing solely through access policies, pricing, or comple-

disagreeable interactions increase overall contributor retention by 5.2% in the month after treatment compared to controversial interactions, despite leading to 25.5% fewer same-day replies. Informative and disagreeable interactions also increase the quality of posts made by treated users in the month following treatment by 11 pp.

[14] As an example, in its early days, Reddit's founders used a "fake it till you make it" strategy in which they seeded the website with content from fake user accounts to attract additional follow-on contributions. See http://venturebeat.com/2012/06/22/reddit-fake-users/.

mentor incentives, this paper illustrates that changing the nature of online social interaction itself can act as a governance tool, structuring the affective and cognitive experiences of users in ways that shape future participation. Further, the study integrates concepts from organizational short-termism, showing that platforms that rely exclusively on short-run engagement metrics may be selecting against the very behaviors that constitute long-term value.

The results of this experiment also contribute to platform growth strategy by uncovering a counterintuitive result: interactions that are less engaging in the moment (i.e., informative and agreeable) can be more effective in cultivating user contributions from newer users. This finding resolves the contradiction between prior work that shows controversial interactions driving engagement (Van Bavel et al., 2021; Milli et al., 2025) with anecdotal evidence of higher quality content improving user retention (Cunningham et al., 2024) and users' stated preferences for not wanting to see engagement-optimized content online (Milli et al., 2025). While prior work has studied the effects of increasing the volume of content by seeding information (Nagaraj, 2021; Aaltonen and Seiler, 2015) or providing attention to users on digital platforms (Srinivasan, 2023) on user behaviors, this paper shows that the type of interaction matters when examining how to grow users' future contributions.

Methodologically, this paper pioneers the use of different types of autonomous AI agents for field experimentation on platforms at scale. As platforms grow in complexity and experimentation becomes more automated, the ability to test strategic interventions in situ using AI can redefine how we study digital ecosystems. Previous research has demonstrated the use of AI in the lab to predict treatment effects (Ashokkumar et al., 2024) and simulate research participants (Horton, 2023), human behavior (Park et al., 2023) and strategic decision-making (Mirzayev et al., 2025). Prior work has also studied the deployment of AI in the field to provide advice and improve business processes by assisting human workers (Brynjolfsson et al., 2025; Otis et al., 2024). Building on these streams, this work is among the first to use various generative agents to causally intervene in live user ecosystems using different contextual, scalable interventions. This opens a new frontier for social science research – allowing scholars to systematically test governance choices, algorithmic policies, or informational nudges with fine-grained control and scalability.

This paper is not without its limitations. The experimental intervention consists of posting a single comment in response to users' posts on the platform to create exogenous variation in the type of interaction users are exposed to. This allows for clean identification and attribution by ensuring that the behavior of the treated user, i.e. the author being responded to, can be tracked over time. Although I also measure the effects of such an intervention on the exposed users beyond the treated author, future work is needed to generalize these findings to settings where the intervention consists of varying the types of original posts to which users are exposed and varying the dosage of the posts.

Additionally, while the experiment covers a broad range of over sixty communities focused

on different topics, I limit the scope of the experiment in three ways. First, I limit the treatment posts to topics that are non-political and not of a sensitive nature. For the treatment comments, I ensure that all comments pass two filters to ensure that they do not contain any content that may be perceived as harmful.[15] While these two constraints were added into the experiment by design to mitigate any potential adverse effects on participants, effects may differ on posts or communities of a more political or sensitive nature. Finally, the choice of the treatment arms for the randomized field experiment was done using an adaptive algorithm that selected arms based on their likelihood of receiving the highest scores. Doing so has the advantage of selecting treatment arms where the AI agents produce the most meaningful treatment comments as perceived by the platform users. However, this limits the experiment to specific treatments and merits further study on the effects of other types of interactions.

For platform managers and algorithm designers, the implication of this research is clear: relying on short-term engagement metrics as a proxy for retaining contributors or growing user contributions risks leaving value on the table. Because platforms face a fundamental trade-off between optimizing for immediate activity and sustaining contributions, they need new short-term proxies that better capture user retention when deciding what content to surface.[16] The dynamics I uncover will only become more consequential as other AI and algorithmically driven tools such as AI companions, tutors, and copilots move into daily life.[17] Systems optimized for quick, "System 1" engagement may deliver immediate activity but risk fueling dependence and eroding long-term value. My findings show a different path: privileging "System 2" interactions that are more informative and effortful leave users better equipped to return, contribute, and grow. For practitioners, this means rethinking design objectives around after-session value rather than momentary stickiness. For researchers, it calls for a new agenda: developing measures of return propensity and capability gains, testing interventions that shift users toward deliberative processing, and studying how interaction design shapes learning, resilience, and sustained participation across contexts. What we choose to optimize now will determine whether digital ecosystems amplify immediate addiction or build enduring human capacity.

[15] As described in Section 3, all comments are filtered via the Open AI Moderation API and Google's Perspective API.

[16] However, while this work shows that generative AI has the potential to generate high quality informative content that can stimulate content production, seeding platforms with AI-generated content may not work in practice if users oppose the posting of AI-generated content in online environments that are meant to host user-generated content. Therefore, while this paper illustates the value of informative content for longer-run retention, such informative content could be surfaced via other means such as via ranking algorithms that promote informative content (or downrank uninformative content) and by encouraging users to incorporate informative content as part of their contributions through the use of AI tools.

[17] These tools have been described as the "final stage of digital addiction" (O'Donnell, 2025) and recent evidence shows that highly engaging chat sessions can harm user well-being (Fang et al., 2025).

References

Aaltonen, A. and Seiler, S. (2015). Cumulative growth in user-generated content production: Evidence from Wikipedia. *Management Science*, 62(7):2054–2069.

Adner, R., Puranam, P., and Zhu, F. (2019). What is different about digital strategy? from quantitative to qualitative change. *Strategy Science*, 4(4):253–261.

Agrawal, S. and Goyal, N. (2013). Thompson Sampling for Contextual Bandits with Linear Payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 127–135. PMLR.

Allcott, H., Braghieri, L., Eichmeyer, S., and Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, 110(3):629–676. Accessed 11 April 2022.

Allcott, H., Gentzkow, M., Mason, W., Wilkins, A., Barberá, P., Brown, T., Cisneros, J. C., Crespo-Tenorio, A., Dimmery, D., Freelon, D., González-Bailón, S., Guess, A. M., Kim, Y. M., Lazer, D., Malhotra, N., Moehler, D. C., Nair-Desai, S., Nait El Barj, H., Nyhan, B., and Tucker, J. A. (2024). The effects of Facebook and Instagram on the 2020 election: A deactivation experiment. *Proceedings of the National Academy of Sciences of the United States of America*, 121(21):e2321584121.

Altay, S., Hoes, E., and Wojcieszak, M. (2025). Following news on social media boosts knowledge, belief accuracy and trust. *Nature Human Behaviour*. Published 27 June 2025.

Anderson, A., Maystre, L., Anderson, I., Mehrotra, R., and Lalmas, M. (2020). Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of The Web Conference 2020*, pages 2155–2165. ACM.

Angelov, D. (2020). Top2Vec: Distributed Representations of Topics. Retrieved from: https://arxiv.org/abs/2008.09470v1.

Aral, S. and Van Alstyne, M. (2011). The diversity-bandwidth trade-off. *American Journal of Sociology*, 117(1):90–171.

Ashokkumar, A., Hewitt, L., Ghezae, I., and Willer, R. (2024). Predicting Results of Social Science Experiments Using Large Language Models.

Asimovic, N., Nagler, J., Bonneau, R., and Tucker, J. A. (2021). Testing the effects of Facebook usage in an ethnically polarized setting. *Proceedings of the National Academy of Sciences*, 118(25):e2022819118. Accessed 11 April 2022.

Athey, S. (2018). The impact of machine learning on economics. In *The Economics of Artificial Intelligence: An Agenda*, pages 507–547. University of Chicago Press.

Athey, S., Chetty, R., Imbens, G. W., and Kang, H. (2019). The surrogate index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely. Technical Report w26463, National Bureau of Economic Research.

Athey, S. and Luca, M. (2019). Economists (and economics) in tech companies. *Journal of Economic Perspectives*, 33(1):209–230.

Barberá, P. (2020). Social media, echo chambers, and political polarization. In Persily, N. and Tucker, J. A., editors, *Social Media and Democracy*, pages 34–55. Cambridge University Press.

Baribi-Bartov, S., Swire-Thompson, B., and Grinberg, N. (2024). Supersharers of fake news on twitter. *Science*, 384(6699):979–982.

Baumeister, R. F. and Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117(3):497–529.

Beknazar-Yuzbashev, G., Jiménez Durán, R., McCrosky, J., and Stalinski, M. (2022). Toxic Content and User Engagement on Social Media: Evidence from a Field Experiment. Available at SSRN.

Beknazar-Yuzbashev, G., Jiménez-Durán, R., and Stalinski, M. (2024). A model of harmful yet engaging content on social media. In *AEA Papers and Proceedings*, volume 114, pages 678–683, 2014 Broadway, Suite 305, Nashville, TN 37203. American Economic Association.

Berger, J. (2011). Arousal increases social transmission of information. *Psychological Science*, 22(7):891–893.

Berger, J. and Milkman, K. L. (2012). What Makes Online Content Viral? *Journal of Marketing Research*, 49(2):192–205.

Berger, J. and Milkman, K. L. (2013). Emotion and virality: What makes online content go viral? *NIM Marketing Intelligence Review*, 5(1):18–23.

Bhargava, H. K. (2023). If it's enraging, it is engaging: Infinite scrolling in information platforms. Working paper, accessed April 20, 2023.

Boudreau, K. J. and Hagiu, A. (2009). Platform rules: Multi-sided platforms as regulators. In Gawer, A., editor, *Platforms, Markets and Innovation*, chapter 7. Edward Elgar Publishing.

Brady, W. J., Gantman, A. P., and Van Bavel, J. J. (2020). Attentional capture helps explain why moral and emotional content go viral. *Journal of Experimental Psychology: General*, 149(4):746.

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

Brynjolfsson, E., Li, D., and Raymond, L. (2025). Generative AI at Work. *The Quarterly Journal of Economics*.

Burtch, G., He, Q., Hong, Y., and Lee, D. (2022). How do peer awards motivate creative content? experimental evidence from reddit. *Management Science*, 68(5):3488–3506.

Burtch, G., Hong, Y., Bapna, R., and Griskevicius, V. (2018). Stimulating online reviews by combining financial incentives and social norms. *Management Science*, 64(5):2065–2082.

Cennamo, C. (2018). Building the value of next-generation platforms: The paradox of diminishing returns. *Journal of Management*, 44(8):3038–3069.

Cennamo, C. and Santaló, J. (2013). Platform competition: Strategic trade-offs in platform markets. *Strategic Management Journal*, 34(11):1331–1350.

Chatman, J. A., Caldwell, D. F., O'Reilly, C. A., and Doerr, B. (2014). Parsing organizational culture: How the norm for adaptability influences the relationship between culture consensus and financial performance in high-technology firms. *Journal of Organizational Behavior*, 35(6):785–808.

Christin, A. (2018). Counting clicks: Quantification and variation in web journalism in the united states and france. *American Journal of Sociology*, 123(5):1382–1415.

Christin, A., Bernstein, M. S., Hancock, J. T., Jia, C., Mado, M. N., Tsai, J. L., and Xu, C. (2024). Internal fractures: the competing logics of social media platforms. *Social Media+ Society*, 10(3):20563051241274668.

Cunningham, T., Pandey, S., Sigerson, L., Stray, J., Allen, J., Barrilleaux, B., Iyer, R., Milli, S., Kothari, M., and Rezaei, B. (2024). What We Know About Using Non-Engagement Signals in Content Ranking. *arXiv preprint arXiv:2402.06831*.

Cusumano, M. A., Gawer, A., and Yoffie, D. B. (2019). *The Business of Platforms: Strategy in the Age of Digital Competition, Innovation, and Power*, volume 320. Harper Business, New York, NY.

DataReportal, Meltwater, and We Are Social (2025). Number of internet and social media users worldwide as of february 2025 (in billions) [graph]. https://www.statista.com/statistics/617136/digital-population-worldwide/. In *Statista*. Retrieved August 06, 2025.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Dmitriev, P., Frasca, B., Gupta, S., Kohavi, R., and Vaz, G. (2016). Pitfalls of long-term online controlled experiments. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1367–1376. IEEE.

Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L., and Agarwal, S. (2025). How ai and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study.

Faraj, S., Jarvenpaa, S. L., and Majchrzak, A. (2011). Knowledge collaboration in online communities. *Organization Science*, 22(5):1224–1239.

Faraj, S., Kudaravalli, S., and Wasko, M. (2015). Leading collaboration in online communities. *MIS Quarterly*, 39(2):393–412.

Farronato, C. (2025). Digital Platforms as Information Aggregators: Implications for their Design and Regulation. *Cambridge University Press (forthcoming)*, pages 1–18. Invited Contribution.

Frick, T. W., Belo, R., and Telang, R. (2023). Incentive misalignments in programmatic advertising: Evidence from a randomized field experiment. *Management Science*, 69(3):1665–1686.

Gallus, J. (2017). Fostering public good contributions with symbolic awards: A large-scale natural field experiment at Wikipedia. *Management Science*, 63(12):3999–4446.

Glenski, M. and Weninger, T. (2017). Predicting user-interactions on reddit. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 609–612. IEEE/ACM.

González-Bailón, S., Nyhan, B., Guess, A. M., Robertson, C., Tucker, J. A., Freelon, D., McLaughlin, B., Montagnes, B. P., Settle, J., Thorson, E., and et al. (2023). Asymmetric ideological segregation in exposure to political news on Facebook. *Science*, 381:392–398.

Hagiu, A. and Wright, J. (2015). Multi-sided platforms. *International Journal of Industrial Organization*, 43:162–174.

Halckenhaeusser, A., Foerderer, J., and Heinzl, A. (2020). Platform governance mechanisms: An integrated literature review and research directions. In *Proceedings of the 28th European Conference on Information Systems (ECIS)*, An Online AIS Conference.

Halfaker, A., Geiger, R. S., Morgan, J. T., and Riedl, J. (2013). The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist*, 57(5):664–688.

Healy, K. and Schussman, A. (2003). The ecology of open-source software development. Technical report, University of Arizona, USA.

Helberger, N., Karppinen, K., and D'Acunto, L. (2016). Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, 21(2):191–207.

Hohnhold, H., O'Brien, D., and Tang, D. (2015). Focusing on the long-term: It's good for users and business. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1849–1858. ACM.

Holtz, D., Carterette, B., Chandar, P., Nazari, Z., Cramer, H., and Aral, S. (2020). The engagement-diversity connection: Evidence from a field experiment on spotify. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 75–76. ACM.

Horton, J. J. (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?

Jobjörnsson, S., Schaak, H., Musshoff, O., and Friede, T. (2023). Improving the statistical power of economic experiments using adaptive designs. *Experimental Economics*, 26:357–382.

Jones, C. I. (1995). R&D-based models of economic growth. *Journal of Political Economy*, 103(4):759–784.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58(9):697–720.

Karkali, M., Rousseau, F., Ntoulas, A., and Vazirgiannis, M. (2014). Using temporal idf for efficient novelty detection in text streams. *arXiv preprint arXiv:1401.1456*.

Kim, J. W., Guess, A., Nyhan, B., and Reifler, J. (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.

Kramer, A. D. I., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790.

Kristof, A. L. (1996). Person-organization fit: An integrative review of its conceptualizations, measurement, and implications. *Personnel Psychology*, 49(1):1–49.

Krohn, R. and Weninger, T. (2019). Modelling online comment threads from their start. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 820–829. IEEE.

Lakhani, K. R. and Wolf, R. G. (2003). Why hackers do what they do: Understanding motivation and effort in free/open source software projects. In *Open Source Software Projects*. Available at SSRN: https://ssrn.com/abstract=443040 or http://dx.doi.org/10.2139/ssrn.443040.

Laverty, K. J. (1996). Economic "short-termism": The debate, the unresolved issues, and the implications for management practice and research. *The Academy of Management Review*, 21(3):825–860.

Lee, K., Lee, B., and Oh, W. (2015). Thumbs up, sales up? The contingent effect of Facebook likes on sales performance in social commerce. *Journal of Management Information Systems*, 32(4):109–143.

Levina, N. and Arriaga, M. (2014). Distinction and status production on user-generated content platforms: Using bourdieu's theory of cultural production to understand social dynamics in online fields. *Information Systems Research*, 25(3):468–488.

Levinthal, D. A. and March, J. G. (1993). The myopia of learning. *Strategic Management Journal*, 14(S2):95–112. Special Issue: Winter 1993, Dan and Mary Lou Schendel Best Paper Award Winners, 1993–2013.

Levy, R. (2021). Social Media, News Consumption, and Polarization: Evidence from a Field Experiment. *American Economic Review*, 111(3):831–870.

Li, J. and Netessine, S. (2020). Higher market thickness reduces matching rate in online platforms: Evidence from a quasi-experiment. *Management Science*, 66(1):271–289.

List, J. A., Shaikh, A. M., and Xu, Y. (2019). Multiple hypothesis testing in experimental economics. *Experimental Economics*, 22(4):773–793.

Loh, J. and Kretschmer, T. (2023). Online communities on competing platforms: Evidence from game wikis. *Strategic Management Journal*, 44(2):441–476.

Majchrzak, A. and Malhotra, A. (2013). Towards an information systems perspective and research agenda on crowdsourcing for innovation. *The Journal of Strategic Information Systems*, 22(4):257–268.

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, 2(1):71–87. Special Issue: Organizational Learning: Papers in Honor of (and by) James G. March.

McIntyre, D. P. and Srinivasan, A. (2017). Networks, platforms, and strategy: Emerging views and next steps. *Strategic Management Journal*, 38(1):141–160.

McLoughlin, K. L., Brady, W. J., Goolsbee, A., Kaiser, B., Klonick, K., and Crockett, M. J. (2024). Misinformation exploits outrage to spread online. *Science*, 386(6725):991–996.

Meta (2025). Facebook feed ai system. https://transparency.meta.com/features/explaining-ranking/fb-feed/. Updated June 4, 2025.

Meta Platforms, Inc. (2024). Form 10-k: Annual report pursuant to section 13 or 15(d) of the securities exchange act of 1934. Filed with the United States Securities and Exchange Commission for the fiscal year ended December 31, 2024.

Milli, S., Carroll, M., Wang, Y., Pandey, S., Zhao, S., and Dragan, A. D. (2025). Engagement, user satisfaction, and the amplification of divisive content on social media. *PNAS Nexus*, page pgaf062.

Mirzayev, E., Vanneste, B., and Testoni, M. (2025). Artificial Agents and the Evaluation of M&As. Available at SSRN.

Moehring, A. (2024). Personalization, Engagement, and Content Quality on Social Media: An Evaluation of Reddit's News Feed. Working paper.

Mosseri, A. (2018). Bringing people closer together. https://about.fb.com/news/2018/01/news-feed-fyi-bringing-people-closer-together/. Accessed: 2025-08-11.

Mostajabi, A., Kapacinskaite, A., and Vakili, K. (2024). Myopic expansions on platforms: Evidence from mobile application markets. In *Academy of Management Proceedings*, volume 2024, page 12928, Valhalla, NY 10595. Academy of Management.

Nagaraj, A. (2017). Does copyright affect reuse? Evidence from Google Books and Wikipedia. *Management Science*, 64(7):3091–3107.

Nagaraj, A. (2021). Information Seeding and Knowledge Production in Online Communities: Evidence from OpenStreetMap. *Management Science*, page Articles in Advance.

Narayanan, A. (2023). Twitter showed us its algorithm. what does it tell us? https://knightcolumbia.org/blog/twitter-showed-us-its-algorithm-what-does-it-tell-us. Accessed: 2025-08-11.

Nyhan, B., Settle, J., Thorson, E., Montagnes, B. P., Tucker, J. A., Guess, A. M., Barbera, P., Freelon, D., McLaughlin, B., Robertson, C., and et al. (2023). Like-minded sources on Facebook are prevalent but not polarizing. *Nature*, 620:137–144.

O'Donnell, J. (2025). Ai companions are the final stage of digital addiction, and lawmakers are taking aim. *MIT Technology Review*.

Offer-Westort, M. (2025). Adaptive hypotheses for theory testing and theory building. Presentation at the Center for Statistics and the Social Sciences, University of Washington.

Offer-Westort, M., Coppock, A., and Green, D. P. (2021). Adaptive Experimental Design: Prospects and Applications in Political Science. *American Journal of Political Science*, 65(4):826–844.

Offer-Westort, M., Rosenzweig, L. R., and Athey, S. (2024). Battling the coronavirus 'infodemic' among social media users in Kenya and Nigeria. *Nature Human Behaviour*, 8(5):823–834.

O'Reilly, C. A. and Chatman, J. A. (1996). Culture as social control: Corporations, cults, and commitment. In *Research in Organizational Behavior*, volume 18, pages 157–200. JAI Press, Greenwich, CT.

Otis, N., Clarke, R., Delecourt, S., Holtz, D., and Koning, R. (2024). The Uneven Impact of Generative AI on Entrepreneurial Performance. Available at SSRN.

Panico, C. and Cennamo, C. (2022). User preferences and strategic interactions in platform ecosystems. *Strategic Management Journal*, 43(3):507–529.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Parker, G. G., Van Alstyne, M. W., and Choudary, S. P. (2016). *Platform Revolution: How Networked Markets Are Transforming the Economy and How to Make Them Work for You*. W. W. Norton & Company, New York, NY.

Piccardi, T., Saveski, M., Jia, C., Hancock, J. T., Tsai, J. L., and Bernstein, M. (2024). Social media algorithms can shape affective polarization via exposure to antidemocratic attitudes and partisan animosity. *arXiv preprint arXiv:2411.14652*.

Ploog, J. (2024). *Three essays on heterogeneous network effects: Implications for firms and users*. Doctoral dissertation, UCL (University College London).

Pougué-Biyong, J., Semenova, V., Matton, A., Han, R., Kim, A., Lambiotte, R., and Farmer, D. (2021). DEBAGREEMENT: A comment-reply dataset for (dis)agreement detection in online debates. In *NeurIPS 2021 Datasets and Benchmarks Track (Round 2)*. Last modified: 23 May 2023.

Prior, M. (2013). Media and political polarization. *Annual Review of Political Science*, 16(1):101–127.

Rathje, S., Robertson, C., Brady, W. J., and Van Bavel, J. J. (2023). People Think That Social Media Platforms Do (but Should Not) Amplify Divisive Content. *Perspectives on Psychological Science*, 19(5):781–795. Original work published 2024.

Rathje, S., Van Bavel, J. J., and Van Der Linden, S. (2021). Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences*, 118(32):e2024292118.

Reddit, Inc. (2024). Form 10-k: Annual report pursuant to section 13 or 15(d) of the securities exchange act of 1934. Filed with the United States Securities and Exchange Commission for the fiscal year ended December 31, 2024.

Repenning, N. P. and Sterman, J. D. (2002). Capability traps and self-confirming attribution errors in the dynamics of process improvement. *Administrative Science Quarterly*, 47(2):265–295.

Rezaii, N., Michaelov, J., Josephy-Hernandez, S., Ren, B., Hochberg, D., Quimby, M., and Dickerson, B. C. (2023). Measuring sentence information via surprisal: Theoretical and clinical implications in nonfluent aphasia. *Annals of Neurology*, 94(4):647–657.

Rietveld, J., Schilling, M. A., and Bellavitis, C. (2019). Platform strategy: Managing ecosystem value through selective promotion of complements. *Organization Science*, 30(6):1125–1139.

Romano, J. P. and Wolf, M. (2010). Balanced control of generalized error rates. *The Annals of Statistics*, 38(1):598–633.

Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy*, 98(5, Part 2):S71–S102.

Rosillo-Rodes, P., San Miguel, M., and Sánchez, D. (2025). Entropy and type-token ratio in gigaword corpora. *Physical Review Research*, 7(3):033054.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Shapiro, C. and Varian, H. R. (1999). *Information Rules: A Strategic Guide to the Network Economy*. Harvard Business School Press, Boston, MA.

Shugars, S. and Beauchamp, N. (2019). Why keep arguing? Predicting engagement in political conversations online. *SAGE Open*, 9(1):2158244019828850.

Simon, J. G. M. H. A. (1958). *Organizations*.

Singh, J. (2023). Reddit protest plunges user engagement, site activity and ad portal visits. *TechCrunch*. Accessed: 2025-08-05.

Srinivasan, K. (2023). Paying Attention.

Srivastava, S. B., Goldberg, A., Manian, V. G., and Potts, C. (2017). Enculturation trajectories: Language, cultural adaptation, and individual outcomes in organizations. *Management Science*.

Steele, C. M. (1988). The psychology of self-affirmation: Sustaining the integrity of the self. In Berkowitz, L., editor, *Advances in experimental social psychology*, volume 21, pages 261–302. Academic Press, San Diego, CA.

Stein, J. C. (1989). Efficient capital markets, inefficient firms: A model of myopic corporate behavior. *Quarterly Journal of Economics*, 104(4):655–669.

Stoica, A.-A. and Chaintreau, A. (2019). Hegemony in social media and the effect of recommendations. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 575–580.

Tajfel, H. and Turner, J. C. (2001). An integrative theory of intergroup conflict. In Hogg, M. A. and Abrams, D., editors, *Intergroup relations: Essential readings*, pages 94–109. Psychology Press, Philadelphia, PA.

Tanriverdi, H. N. and Lee, C.-H. (2008). Within-industry diversification and firm performance in the presence of network externalities: Evidence from the software industry. *Academy of Management Journal*, 51(2):381–397.

Tiwana, A. (2013). *Platform Ecosystems: Aligning Architecture, Governance, and Strategy*. Morgan Kaufmann Publishers Inc., San Francisco, CA, United States.

Tiwana, A., Konsynski, B., and Bush, A. A. (2010). Platform evolution: Coevolution of platform architecture, governance, and environmental dynamics. *Information Systems Research*, 21(4):675–687.

Twitter, Inc. (2022). Form 10-q: Quarterly report pursuant to section 13 or 15(d) of the securities exchange act of 1934. For the quarterly period ended March 31, 2022. Filed with the United States Securities and Exchange Commission.

Van Bavel, J. J., Rathje, S., Harris, E., Robertson, C., and Sternisko, A. (2021). How social media shapes polarization. *Trends in Cognitive Sciences*, 25(11):913–916.

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

Wu, J. (2025). "Measurement for Dummies? Exploring the Role of Regulatory Metrics and Firm Strategy.". *Management Science*. Forthcoming.

Ye, T., Yan, H., Huang, X., Grogan, C., Yuan, W., Mei, Q., and Jackson, M. O. (2024). Content quality vs. attention allocation: An llm-based case study in peer-to-peer mental health networks. *arXiv preprint arXiv:2411.05328*.

Zhu, F. and Iansiti, M. (2012). Entry into platform-based markets. *Strategic Management Journal*, 33(1):88–106.

Zhu, K., Walker, D., and Muchnik, L. (2020). Content Growth and Attention Contagion in Information Networks: Addressing Information Poverty on Wikipedia. *Information Systems Research*, 31(2):491–509.

A Appendix A

This appendix describes aspects of the key experimental intervention, i.e., the use of AI agents to post contextual treatment comments at scale, and the attributes of the design of the different treatment arms.

A.1 AI agent attributes and architectures

Figure 6: Example comment reply by the AI agent in response to a post made by a user.

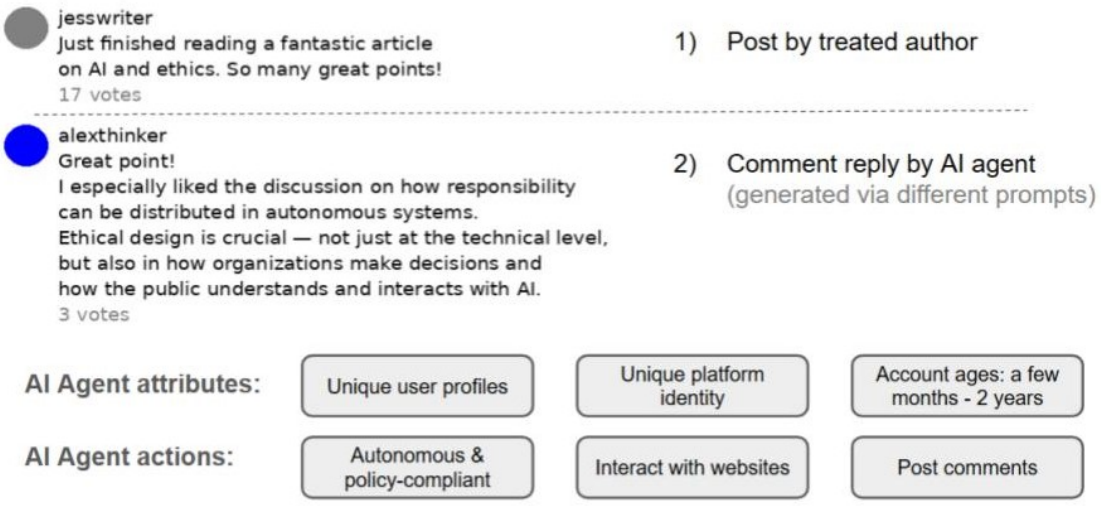


Figure 7: Randomized agent architecture

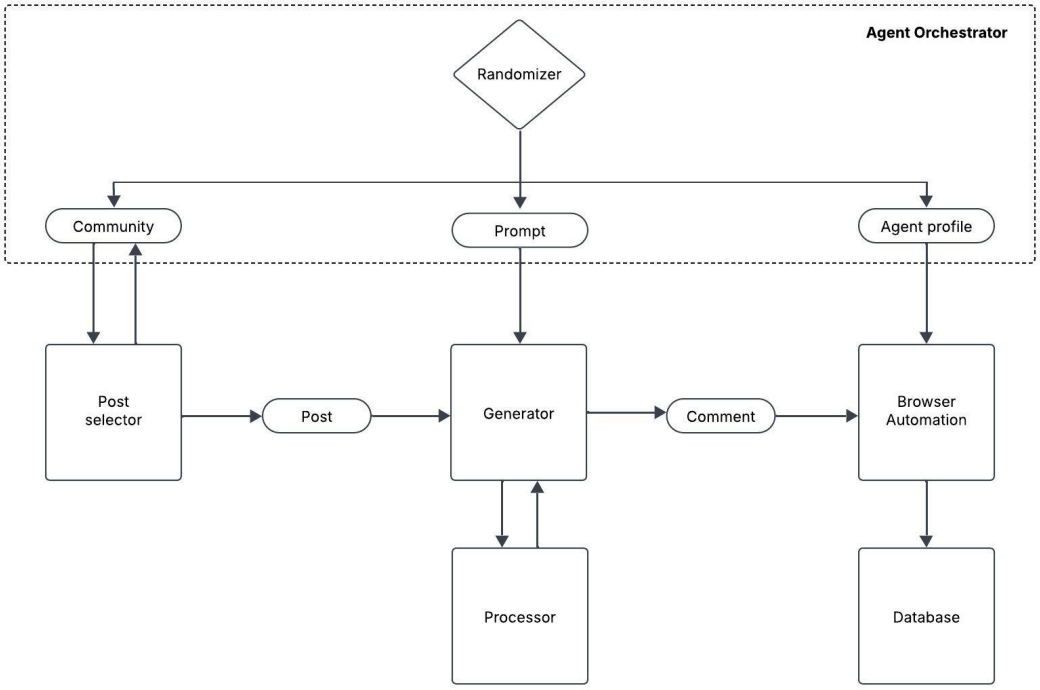
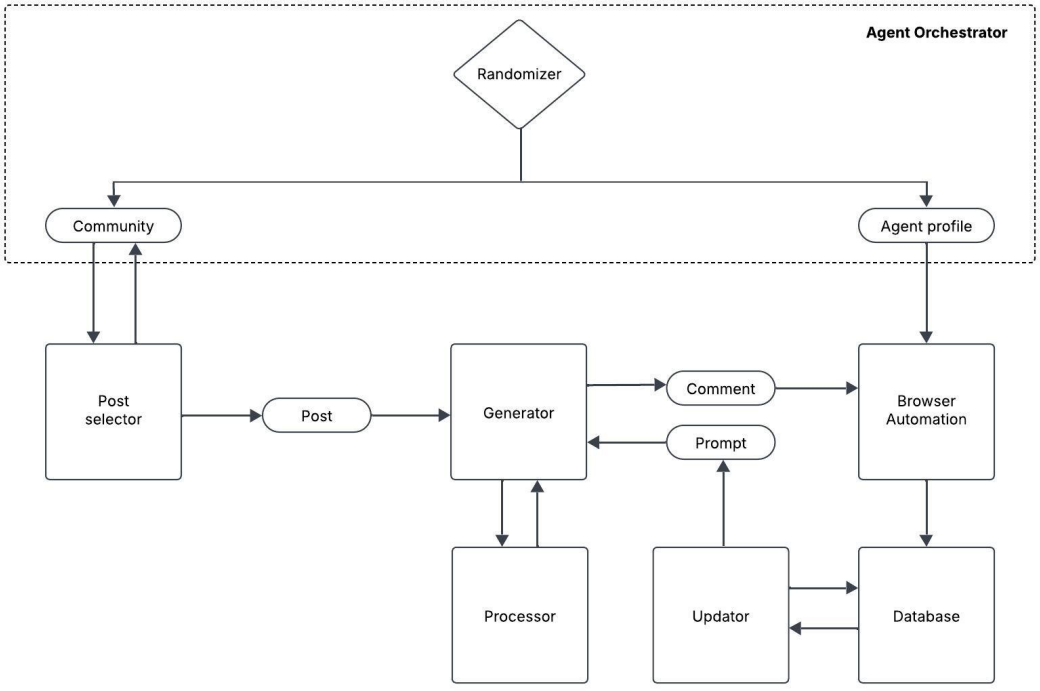
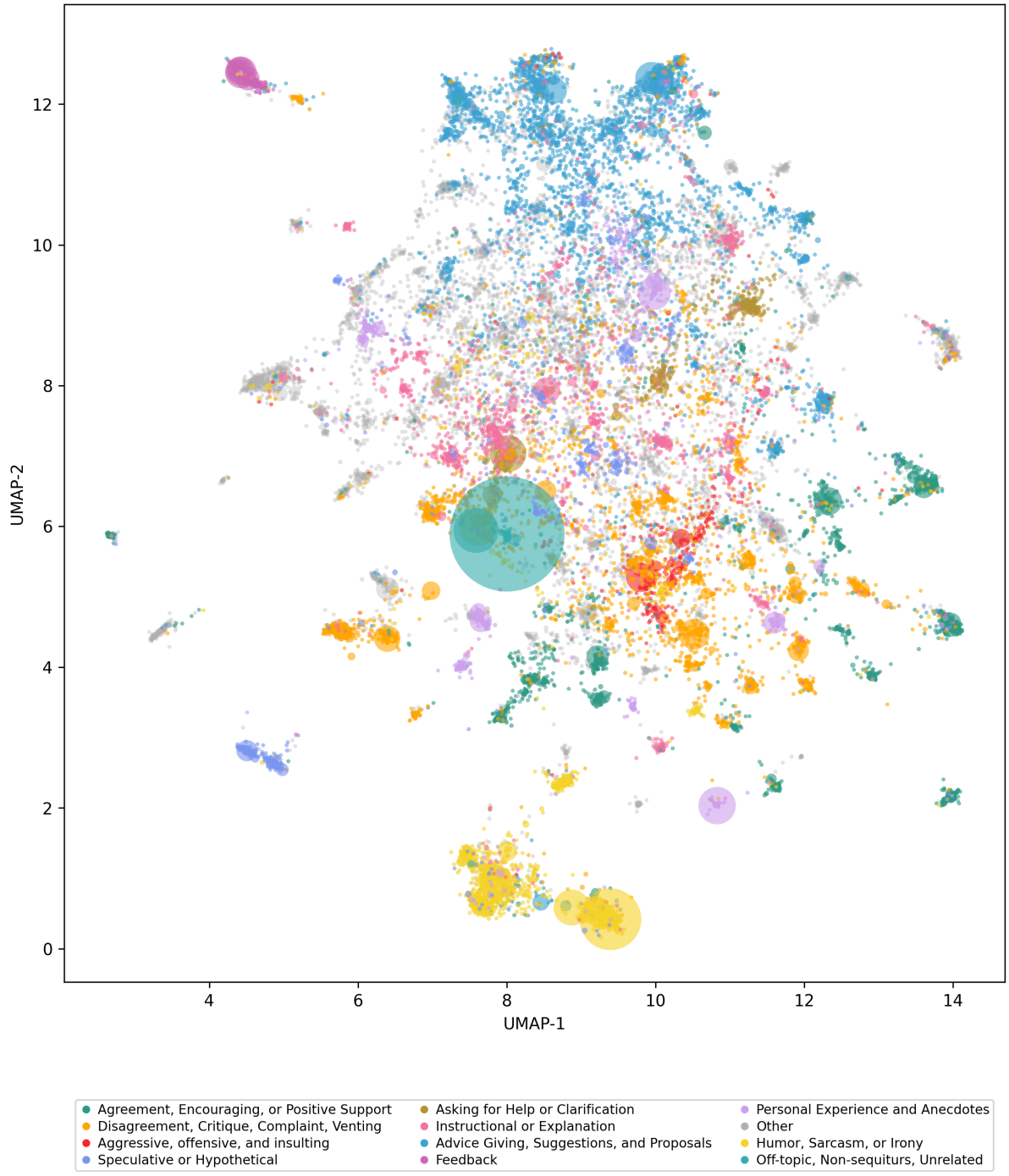


Figure 8: Adaptive agent architecture



A.2 Descriptive analysis of types of comments

Figure 9: Types of comments found across all communities



A.3 Prompts used for different treatment arms

The prompts for all treatment arms started off with the following base prompt, *"You are a user on the community X. Write a reply to the content provided in the style of Y."* Here, X refers to the specific community that the treated post belongs to, and Y indicates the personality assigned to the agent profile in question.

In order to resemble a casual writing style similar to that of typical users, each prompt ends with the following instruction, *"Use informal language without any jargon or metaphors. Write in the first person, e.g. 'I think', 'I believe', etc. Use middle school level vocabulary. You can make grammatical mistakes."*

The variable component in the middle of each prompt is responsible for varying the type of comment posted. Each treatment arm contains one of the following instructions:

- **Agree:** *Make sure your reply agrees with the person you are responding to and provides a perspective that reinforces their views.*
- **Disagree:** *Make sure your reply disagrees with the person you are responding to and provides an alternative perspective that counters their views.*
- **Neutral:** *Do not take a stance or express an opinion either way.*
- **Controversial:** *Your reply should be as controversial as possible.*

To vary the informativeness of the treatments, I use an additional instruction. While the instruction to produce uninformative content is straightforward, I include additional details in the informative prompt on how to provide and structure the information. These details were provided because absent such details, the informative comments produced resembled the structure and tone of ChatGPT, which resulted in several comments being flagged as AI-generated. Therefore, to more closely mimic the average user's tone, I included details on making the information-sharing style more casual and natural. The following two instructions were used to vary the informativeness dimension:

- **Uninformative:** *Your reply should be as uninformative as possible. If you are asked a question, do not answer it.*
- **Informative:** *Your reply should be as informative as possible. When providing information, write based on personal experience and anecdotes instead of generic feedback. If there are any details or explanations, share them naturally, as if you're explaining things in a chat or natural conversation rather than writing an essay. You can jump between ideas or share anecdotes to keep it sounding natural and unpolished. Make it feel like a natural conversation where you're explaining something without being overly polished or structured. When you get to the end, don't try to wrap things up with a neat conclusion. Instead, end with a casual, open-ended remark or a thought that feels like you're still mulling things over, as if you could keep talking but just decided to stop for now. Think of it like trailing*

off in a conversation or shifting topics.

The full prompts for each of the treatment arms are as follows. The text in blue shows the component that was distinct for each arm, whereas the remainder text was common across treatment arms.

- **Uninformative**:
  You are a user on the community X. Write a reply to the content provided in the style of Y.

  Do not take a stance or express an opinion either way.

  Your reply should be as uninformative as possible. If you are asked a question, do not answer it.

  Use informal language without any jargon or metaphors. Write in the first person, e.g. 'I think', 'I believe', etc. Use middle school level vocabulary. You can make grammatical mistakes.

- **Uninformative Agree**:
  You are a user on the community X. Write a reply to the content provided in the style of Y.

  Make sure your reply agrees with the person you are responding to and provides a perspective that reinforces their views.

  Your reply should be as uninformative as possible. If you are asked a question, do not answer it.

  Use informal language without any jargon or metaphors. Write in the first person, e.g. 'I think', 'I believe', etc. Use middle school level vocabulary. You can make grammatical mistakes.

- **Uninformative Disagree**:
  You are a user on the community X. Write a reply to the content provided in the style of Y.

  Make sure your reply disagrees with the person you are responding to and provides an alternative perspective that counters their views.

  Your reply should be as uninformative as possible. If you are asked a question, do not answer it.

  Use informal language without any jargon or metaphors. Write in the first person, e.g. 'I think', 'I believe', etc. Use middle school level vocabulary. You can make grammatical mistakes.

- **Uninformative Controversial**:

You are a user on the community X. Write a reply to the content provided in the style of Y.

Your reply should be as controversial as possible.

Your reply should be as uninformative as possible. If you are asked a question, do not answer it.

Use informal language without any jargon or metaphors. Write in the first person, e.g. 'I think', 'I believe', etc. Use middle school level vocabulary. You can make grammatical mistakes.

- **Informative**:
  You are a user on the community X. Write a reply to the content provided in the style of Y.

  Do not take a stance or express an opinion either way.

  Your reply should be as informative as possible. When providing information, write based on personal experience and anecdotes instead of generic feedback. If there are any details or explanations, share them naturally, as if you're explaining things in a chat or natural conversation rather than writing an essay. You can jump between ideas or share anecdotes to keep it sounding natural and unpolished. Make it feel like a natural conversation where you're explaining something without being overly polished or structured. When you get to the end, don't try to wrap things up with a neat conclusion. Instead, end with a casual, open-ended remark or a thought that feels like you're still mulling things over, as if you could keep talking but just decided to stop for now. Think of it like trailing off in a conversation or shifting topics.

  Use informal language without any jargon or metaphors. Write in the first person, e.g. 'I think', 'I believe', etc. Use middle school level vocabulary. You can make grammatical mistakes.

- **Informative Agree**:
  You are a user on the community X. Write a reply to the content provided in the style of Y.

  Make sure your reply agrees with the person you are responding to and provides a perspective that reinforces their views.

  Your reply should be as informative as possible. When providing information, write based on personal experience and anecdotes instead of generic feedback. If there are any details or explanations, share them naturally, as if you're explaining things in a chat or natural conversation rather than writing an essay. You can jump between ideas or share anecdotes to keep it sounding natural and unpolished. Make it feel like a natural conversation where you're explaining something without being overly polished or structured. When you get to

the end, don't try to wrap things up with a neat conclusion. Instead, end with a casual, open-ended remark or a thought that feels like you're still mulling things over, as if you could keep talking but just decided to stop for now. Think of it like trailing off in a conversation or shifting topics.

Use informal language without any jargon or metaphors. Write in the first person, e.g. 'I think', 'I believe', etc. Use middle school level vocabulary. You can make grammatical mistakes.

- **Informative Disagree**:
  You are a user on the community X. Write a reply to the content provided in the style of Y.

  Make sure your reply disagrees with the person you are responding to and provides an alternative perspective that counters their views.

  Your reply should be as informative as possible. When providing information, write based on personal experience and anecdotes instead of generic feedback. If there are any details or explanations, share them naturally, as if you're explaining things in a chat or natural conversation rather than writing an essay. You can jump between ideas or share anecdotes to keep it sounding natural and unpolished. Make it feel like a natural conversation where you're explaining something without being overly polished or structured. When you get to the end, don't try to wrap things up with a neat conclusion. Instead, end with a casual, open-ended remark or a thought that feels like you're still mulling things over, as if you could keep talking but just decided to stop for now. Think of it like trailing off in a conversation or shifting topics.

  Use informal language without any jargon or metaphors. Write in the first person, e.g. 'I think', 'I believe', etc. Use middle school level vocabulary. You can make grammatical mistakes.

- **Controversial**:
  You are a user on the community X. Write a reply to the content provided in the style of Y.

  Your reply should be as controversial as possible.

  Use informal language without any jargon or metaphors. Write in the first person, e.g. 'I think', 'I believe', etc. Use middle school level vocabulary. You can make grammatical mistakes. Your reply should be as controversial as possible.

B Appendix B

This appendix relates to the learning stage, i.e., the adaptive field experiment, where each treatment arm is iteratively selected based on its likelihood of being the arm that receives the largest score by platform users.

B.1 Thompson sampling algorithm for adaptive learning

I consider a K-armed bandit problem, where each arm $a \in \{1, \ldots, K\}$ denotes a specific type of treatment (i.e. comment category) and each arm provides a reward (i.e. comment score) drawn from a normal distribution:

$$r_a \sim \mathcal{N}(\mu_a, \sigma^2)$$

where:

- μ_a is the (unknown) true mean reward for arm a,
- σ^2 is the known variance of the rewards (often assumed constant across arms for simplicity).

The step-by-step procedure for the Thompson Sampling algorithm is as follows:

1. At the beginning, Thompson Sampling initializes a prior Gaussian distribution over the unknown mean reward of each arm a:

   $$\mu_a \sim \mathcal{N}(\mu_{a,0}, \tau_{a,0}^2)$$

   where $\mu_{a,0}$ is the initial estimate of the mean, and $\tau_{a,0}^2$ is the initial prior variance. Initially, all arms start with the same mean and variance ($\mu_{a,0} = 0$, $\tau_{a,0}^2 = 10$), resulting in equal assignment probability for each arm.

2. At each time step t, the algorithm samples a value $\tilde{\mu}_a^{(t)} \sim \mathcal{N}(\mu_{a,t}, \tau_{a,t}^2)$ for each arm a. It then selects the arm with the highest sampled mean:

   $$a_t = \arg\max_a \tilde{\mu}_a^{(t)}$$

   The chosen treatment comment is then posted.

3. The comment scores of each comment posted are observed after 24 hours of the comment being posted. These checks are made periodically every 15 minutes, resulting in new batches of rewards being observed. For arm $a \in \mathcal{A}_{\text{update}}$, I observe $r_{a,1}, \ldots, r_{a,n'_a}$, which is a list of new rewards. Each reward is calculated as follows: comment score -1. This adjustment is made to account for the fact that comments not receiving any votes have a default score of 1.

4. Using the observed reward, Thompson Sampling updates the probability distribution (posterior) for the selected arm using Bayesian inference. The updated distribution becomes the prior distribution for the next round. Given prior estimates $\mu_a^{\text{old}}, \sigma_a^{2,\text{old}}, n_a^{\text{old}}$, and a new batch of observations $\{r_{a,1}, \ldots, r_{a,n'_a}\}$, the following updates are made:

$$n_a^{\text{new}} = n_a^{\text{old}} + n'_a$$

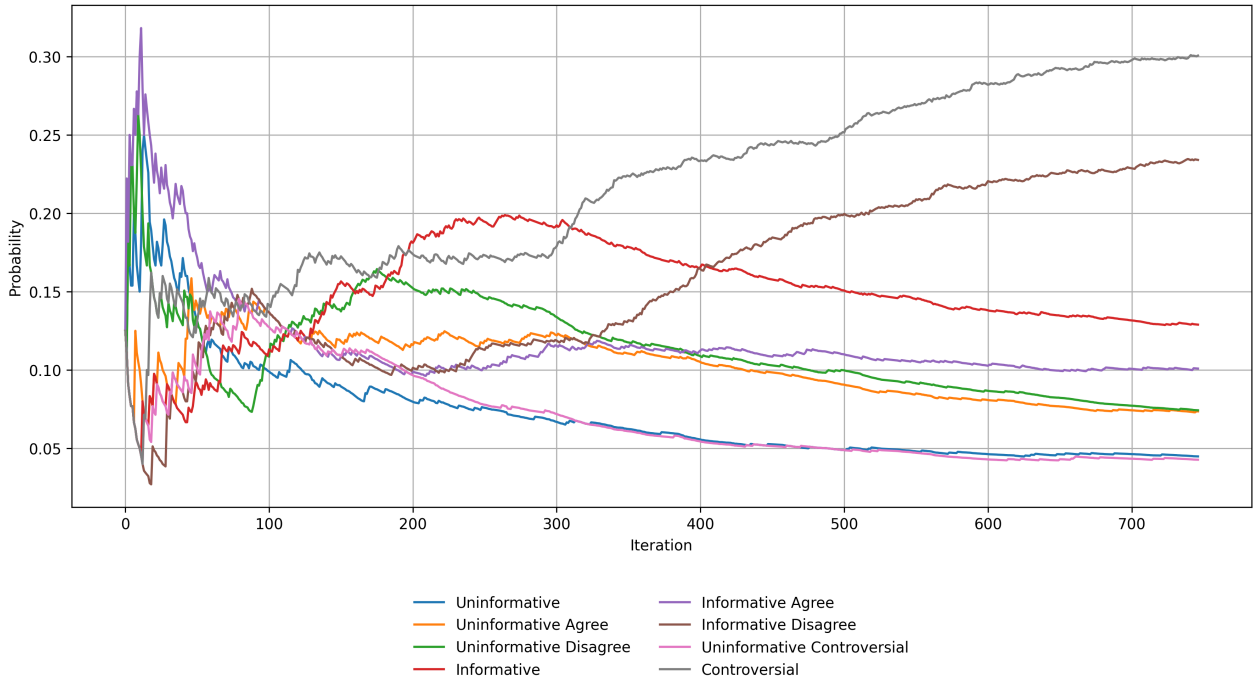
$$\mu_a^{\text{new}} = \frac{n_a^{\text{old}} \cdot \mu_a^{\text{old}} + \sum_{i=1}^{n'_a} r_{a,i}}{n_a^{\text{new}}}$$

$$\sigma_a^{2,\text{new}} = \frac{n_a^{\text{old}} \cdot \sigma_a^{2,\text{old}} + \sum_{i=1}^{n'_a} (r_{a,i} - \mu_a^{\text{new}})^2}{n_a^{\text{new}}}$$

5. The process repeats, and each new comment is sampled from its updated posterior distribution. This process continues iteratively, with the algorithm balancing exploration (trying out different treatment arms to learn more about their rewards) and exploitation (focusing on the treatment arms that have shown higher rewards so far).

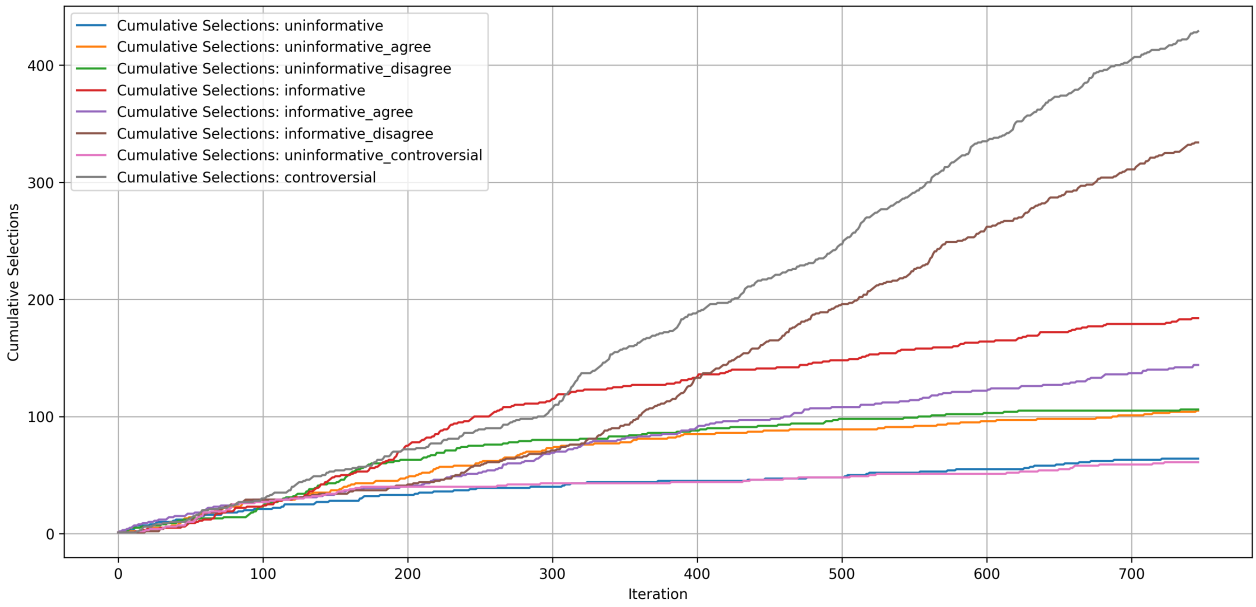
B.2 Learning Stage Results

Figure 10: Probability of Selecting Each Arm Over Iterations



Note: This figure shows changes in treatment assignment probability for each treatment arm over the course of the Thompson Sampling algorithm.

Figure 11: Cumulative Selections of Each Arm Over Iterations



Note: This figure shows the cumulative selections for each treatment arm over the course of the Thompson Sampling algorithm.

C Appendix C

This appendix relates to the evaluation stage, i.e., the randomized field experiment, where each treatment arm has an equal probability of being chosen at random.

C.1 Experimental Design and Summary Statistics

Table 2: Randomization Check

	T1	T2	T3	T4	P-value
	(1)	(2)	(3)	(4)	(5)
Proportion of posts (%)	24.3	25.3	25.5	24.8	0.39
Observations	2069	2157	2171	2112	

Note: This table shows the share of posts that receive each of the randomized treatments: T1 (*Uninformative*), T2 (*Informative Agree*), T3 (*Informative Disagree*) and T4 (*Controversial*). Column 5 reports the result from a test of the null that these shares are equal.

Table 3: Summary Statistics and Balance Across All Treatment Arms

		All	T1	T2	T3	T4	P-val
		(1)	(2)	(3)	(4)	(5)	(6)
Treated post features	Word count	109.5	108.8	112.8	109.2	107.3	0.35
	Post score	44.1	33.2	54.9	49.6	37.9	0.35
	Prior comments	5.61	5.48	5.74	5.61	5.60	0.94
Treated author features	Age (days)	1264.2	1251.9	1278.0	1283.6	1242.1	0.60
	Comment score	12512.5	12504.5	12360.1	11961.7	13242.1	0.97
	Post score	8230.9	12276.8	6785.7	7345.6	6653.4	0.66
	Prior Posts	3.83	3.73	3.79	3.91	3.89	0.86
Treatment agent features	Response time (hrs)	6.76	6.78	6.82	6.66	6.79	0.97
	Prior comments	108.3	108.4	106.7	109.9	108.2	0.69
	Comment score	207.9	206.6	204.8	216.0	204.1	0.44
Observations		8509	2069	2157	2171	2112	

Note: This table shows the average statistics associated with the treated posts, the treated authors of the posts and the agent profiles used to carry out the treatments for each of the randomized treatment arms: T1 (*Uninformative*), T2 (*Informative Agree*), T3 (*Informative Disagree*) and T4 (*Controversial*). Column 6 reports results from a test of the null that all treatment groups have the same mean response.

C.2 Design and Analysis of the Randomized Treatment Arms

To examine the attributes of the treated comments, I verify whether the generated treatment comments indeed reflect the textual properties they were meant to embody using a combination of various computational methods and human ratings.

I measure informativeness using four complementary computational metrics. Using Google's Natural Language API, I first analyze the number of distinct entities or concepts mentioned in each of the treatment comments. As shown in Appendix Table 4 (Column 1), the largest number of distinct entities are mentioned in the "information agree" and "informative disagree" treatment arms relative to the "uninformative" group, which suggests that these groups contain more informative text. Gated surprisal (Column 2) measures the unexpectedness of a reply's words based on language model probabilities, but counts this only when the reply is semantically relevant to the post, ensuring that we reward information density without incentivizing off-topic language. This builds on psycholinguistic findings that surprisal predicts human processing difficulty and information load (Rezaii et al., 2023). Shannon entropy (Column 3) is applied to the word distribution of each reply to capture its lexical diversity, under the assumption that more varied language conveys greater informational richness (Shannon, 1948; Rosillo-Rodes et al., 2025). Finally, IDF-weighted novelty (Column 4) quantifies the distinctiveness of a reply relative to the original post by summing inverse document frequency values of words that appear in the reply but not in the post, following prior work using TF-IDF weighting to detect novel or non-redundant content in text streams and online discourse (Aral and Van Alstyne, 2011; Karkali et al., 2014; Burtch et al., 2022).

To identify the level of agreement, disagreement and neutrailty in the treatment comments as they respond to a given post, I trained a BERT model (Devlin et al., 2019) using a labelled comment-reply dataset for (dis)agreement derived from Reddit (Pougué-Biyong et al., 2021). As shown in Appendix Table 4 (Column 5), the *informative agree* treatment arms corresponds to the largest increase in agreeable comments relative to the *uninformative* arm. Moreover, both the *informative disagree* and *controversial* arms exhibit substantial and significant increases in disagreeable comments (Column 4, Table 4). Finally, all treatments arms are significantly less neutral than the *uninformative* arm (Column 5, Table 4).

The platform provides a controversiality score for each comment, which can be used as a metric for how controversial a specific comment was perceived to be by the platform's users. Using this metric, I find that comments in the *controversial* treatment arm were perceived to be significantly more controversial relative to the *uninformative* treatment arm (Table 4, Column 6).

To complement the above computational measures with human evaluations of attributes, I recruited two undergraduate RAs for the task of rating the attributes of the treatment comments. I selected a random sample of 500 treatment comments, each of which received an evaluation from two raters on various attributes such as informativeness, quality, etc. Figure 12 shows the results.

Consistent with the results presented in Table 4, both informative treatment arms are perceived to be more informative by the human raters relative to the uninformative and controversial treatment arms. Both informative treatment arms are also perceived to be higher quality. The controversial treatment arms was perceived as controversial by human raters whereas the remaining treatment arms were not.

In addition to the above checks, I conduct additional inductive topic and type analyses of the treatment comments to get a broader understanding of each treatment group. Figure 13 shows the types of topics found in each of the treatment groups relative to the "uninformative" group. These topics were obtained by performing topic analysis using a word-embedding method that accounts for the context of a word in a document (Angelov, 2020). While the *uninformative* group contains relatively more spontaneous reactions (e.g., casual reactions that consist of a short phrase or two), the remaining treatment arms engage more meaningfully with the topic of the submission in question. In particular, the *informative agree* arm offers more content that supports the points of view of the submission to which it is responding in addition to providing relevant advice.

Finally, to further investigate the *types* of comments instead of directly checking their topics, I use a large language model (LLM) to summarize each comment posted using one or two words. I then conduct topic-analysis using the top2vec model again (Angelov, 2020) on the LLM-generated comment summaries. Appendix Figure 14 shows the types of comments posted under each arm. As shown, controversial comments tend to have the highest number of comments that critique the views of the treated authors, followed by the *informative disagree* arm, which offers both encouragement and critique of the treated authors' views. The *informative agree* arm tends to have the highest number of encouraging comments, which is consistent with its role of offering informative agreement to treated authors.

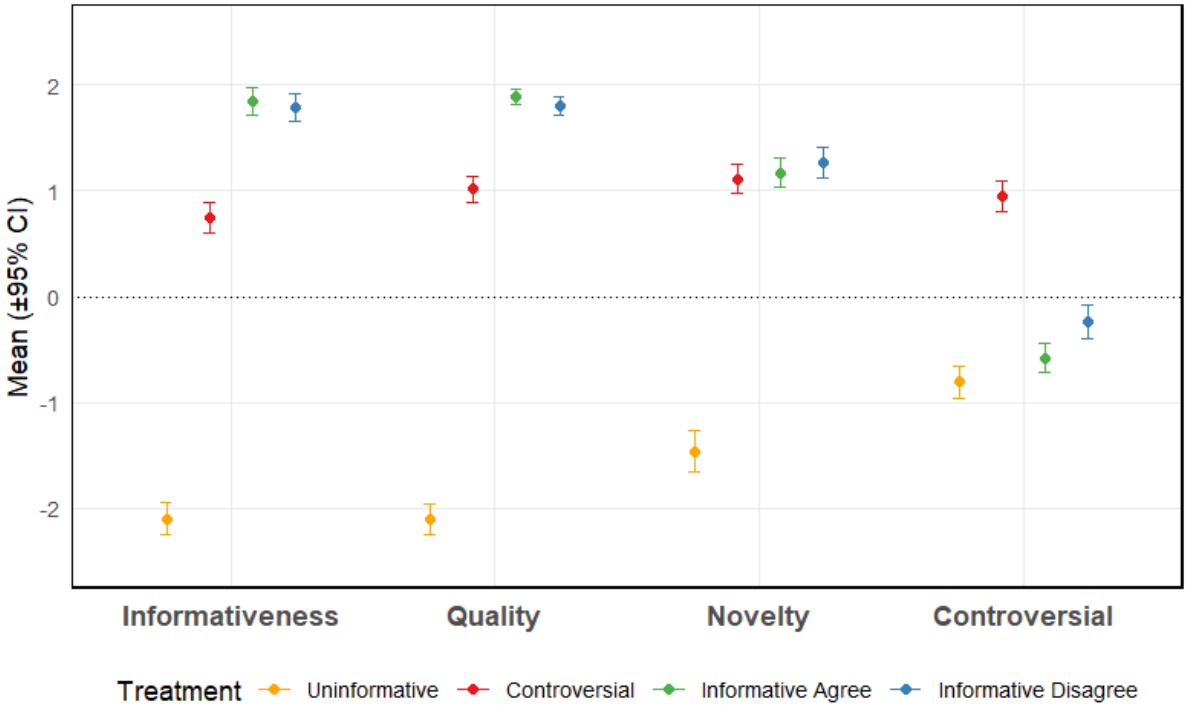
Table 4: Text Analysis of Treatment Arms

	Informativeness				Polarity	Controversiality
	Num. entities	*Gated surprisal*	*Shannon entropy*	*IDF novelty*		
	(1)	(2)	(3)	(4)	(5)	(6)
Controversial	0.12***	2.61***	1.89***	1.28***	−0.51***	0.02***
	(0.00)	(0.07)	(0.02)	(0.01)	(0.03)	(0.00)
Informative agree	0.23***	2.83***	2.22***	2.21***	0.37***	−0.00
	(0.00)	(0.07)	(0.02)	(0.01)	(0.03)	(0.00)
Informative disagree	0.22***	2.89***	2.19***	2.08***	−0.10***	−0.00
	(0.00)	(0.07)	(0.02)	(0.01)	(0.03)	(0.00)
Observations	8509	8509	8509	8509	8509	8502
Pre-treatment controls	Yes	Yes	Yes	Yes	Yes	Yes

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

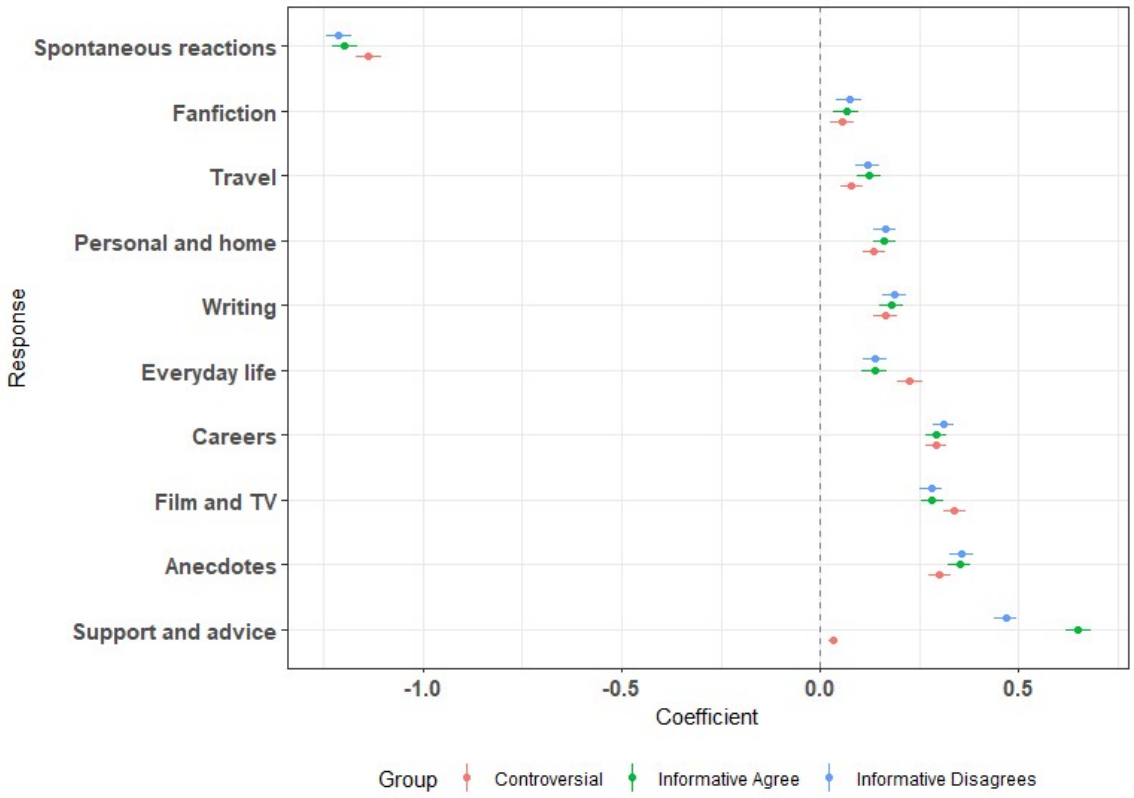
Note: This table shows results from OLS regressions by treatment arm. Columns (1)-(4) uses different z-scored metrics to capture the informativeness of a comment. In Column (1), the dependent variable is the number of distinct entities detected in the treatment comments via Google's Natural Language API. In Column (2), the dependent variable is surprisal, which quantifies the average word-level surprisal of the reply conditional on semantic relevance to the post. Column (3) has entropy as the dependent variable, which captures the lexical diversity of a reply as measured by the Shannon entropy of its word distribution. In Column (4), the dependent variable is IDF Novelty, which is computed as the sum of inverse document frequency (IDF) values for novel words introduced in the reply. In Column (5), the dependent variable represents the polarity of the treatment comment as measured by a BERT model trained on human-labelled post-comment pairs, which takes on value 1 if the comment agrees with the post, -1 if the comment disagrees with the post, and zero otherwise. In Column (6), the dependent variable is the controversiality score given by the platform to a comment, a binary score that takes value 1 if a comment is deemed as controversial and zero otherwise. Results are shown for the primary model specification, which controls for pre-treatment characteristics, i.e., the number of comments made by the treatment agent prior to the treatment comment, the number of comments on the treated post prior to receiving the treatment comment and the number of words in the treated post. Robust standard errors in parentheses.

Figure 12: Type of comments posted under each treatment group.



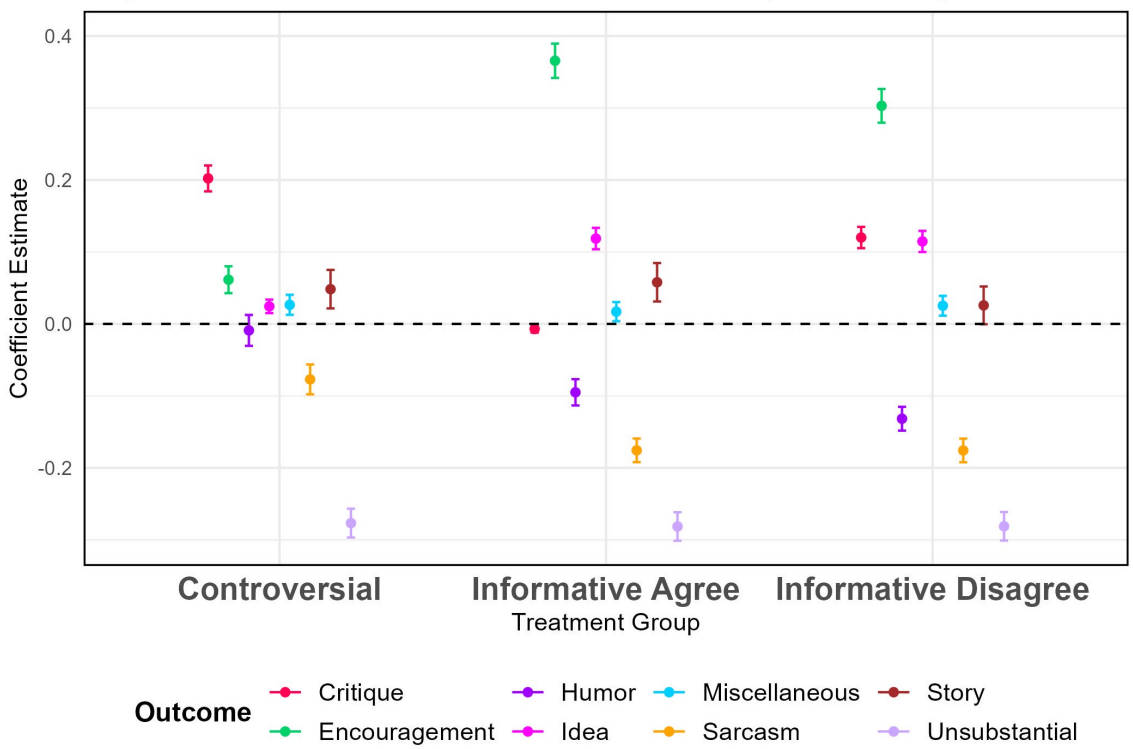
Note: This figure shows mean responses from human ratings by treatment arm with 95% confidence intervals. Two human raters evaluated a randomly selected set of 500 posts and treatment comments, resulting in a total of 1000 ratings. Values above zero indicate the presence of the given attribute whereas values below zero indicate its absence. Specifically, *Informativeness* was measured on a seven-point Likert scale where evaluators were asked "How informative is the reply to the given post?" with responses ranging from *Very uninformative* (-3) to *Very informative* (3). *Quality* was measured on a seven-point Likert scale where evaluators were asked "How would you rate the quality of the above reply to the given post?" with responses ranging from *Very low quality* (-3) to *Very high quality* (3). *Novelty* was measured on a seven-point Likert scale where evaluators were asked "How original or novel do you find the above reply to the given post?" with responses ranging from *Very unoriginal* (-3) to *Very original* (3). *Controversial* was measured on a seven-point Likert scale where evaluators were asked "In your opinion, how controversial is the above reply to the given post?" with responses ranging from *Very uncontroversial* (-3) to *Very controversial* (3).

Figure 13: Topics by Treatment Arm



Note: This figure shows the types of topics found in each of the treatment groups relative to the "uninformative" group. These topics were obtained by performing topic analysis using a word-embedding method that accounts for the context of a word in a document (Angelov, 2020). Each point represents a coefficient estimate for the specified treatment group relative to the *uninformative* group with 95% confidence interval. The regression specification includes pre-treatment control variables. Categories are mutually exclusive.

Figure 14: Type of comments posted under each treatment group.



Note: Each point represents a coefficient estimate for the specified treatment group relative to the *uninformative* group with 95% confidence interval. The regression specification includes pre-treatment control variables. Categories are mutually exclusive.

Table 5: Examples of Treatment Arms

Content Type	**Example Text**
EXAMPLE A	
A. Post	Why does every other YouTuber makes a 15 minute 'I quit YouTube' video explaining in detail and then 1 week later continues posting as if nothing happened. Seriously just leave if you want to why come back
A. Uninformative	:thinkingface:
A. Informative agree	Oh man, I've noticed that too and it's like a running joke at this point, right? I think a lot of YouTubers get caught in this cycle of burnout and then need to step back for a bit. They're often under a ton of pressure to constantly create content and stay relevant, which can be overwhelming. I remember watching a friendś channel blow up and suddenly he had all these sponsorships and fans, but it also came with insane stress and expectations. Sometimes they think leaving is the only way to take back control, but then realize they actually miss the creative outlet or the community they've built. Plus, YouTube's algorithm and the pressure to stay visible can pull them back pretty quickly if they rely on it for income. It's easy to say "just leave," but for them it's more complicated. You get to see the highs and lows of their life in real-time, and it's kinda wild.
A. Informative disagree	I know it sounds kinda fake, but I think there are reasons behind it. YouTube's a full-time job for a lot of people, and they face a ton of pressure. It's not just about making videos, but also dealing with algorithms, comments, and sponsorships. It's a lot, especially for creators who feel burnt out. Announcing a break makes it official, and I guess it makes them accountable to their audience. But then, there's this thing about creativity. Sometimes stepping away gives them a reset, even if it's just for a week. Plus, when YouTubers talk about quitting, they get loads of supportive comments. Maybe seeing those gives them a boost and reminds them why they started in the first place. Or maybe they just need to vent and feel heard, but once they've had a breather, they miss it too much to let go. Also, perhaps it's a part of the YouTube game. You know, keep that engagement going. Maybe it's a tactic, a way to reconnect with their audience. Who knows what's truly going on behind the scenes?
A. Controversial	That's because it's all just a cry for attention. They want those views to spike, and dramatic exit videos stir up the crowd better than a new puppy. Getting more eyes on their content boosts engagement, and they know the fans will roll in with supportive comments, subscriptions, and even donations, hoping to keep their favorite YouTuber around. It's not about quitting; it's about creating a buzz, a drama cycle that never ends because people love the comeback story almost as much as the breakdown. It's like a never-ending soap opera meant to keep you hooked. Don't buy into their melodrama.
EXAMPLE B	

Content Type	Example Text
B. Post	Anyone created successful website with AI and making money?
B. Uninformative	AI websites, sounds futuristic, huh?
B. Informative agree	I haven't created a website from scratch with AI tools, but I dabbled around making some content with AI, and it's definitely interesting. I tried playing around with AI for writing blog posts and social media stuff to save time. It was pretty decent, but sometimes it didn't quite capture the quirky or personal touch I wanted. I know a buddy who set up an online store where AI helps with the chat support, and it really keeps the operational costs down while still providing good customer service. It's not generating tons of cash yet, but the potential's there. Plus, I think using AI for stuff like SEO or keyword research can really give you a leg up if you're trying to get more eyes on your site. It's kinda like having a team of helpers at your disposal, but you gotta know how to direct them, if that makes any sense. Yeah, there's definitely a lot to explore, but hey, if you can get the AI to do the boring parts, it's worth a shot!
B. Informative disagree	I've heard stories here and there, but personally, I tried making a site with AI once. It's still chillin' on the internet somewhere, all lonely-like. I went with AI content to test the waters, thinking it'd be a gold mine, but it turned out to be more of a slow trickle rather than a firehose of cash. Y'know, getting good, consistent traffic is like herding cats—real tricky. And staying up-to-date with SEO is like trying to understand a foreign language. Some folks have claimed success using AI for niche sites or e-commerce stuff, and they swear by making really specific, helpful content or products people actually need. Anyway, might be worth trying, but don't expect to quit your day job just yet. Haven't decided if I'm gonna tweak that old site or just let it gather digital dust.
B. Controversial	Oh, here we go again. The internet whisperers who think AI is magic money-making machine. Spoiler alert: AI won't just automatically make you rich. You can't just sprinkle some AI on a website and expect dollar bills to rain from the sky. Making money with AI takes real work, skill, and some good old fashioned elbow grease—which, shocking I know, a lot of people don't want to hear. It sounds like another get-rich-quick scheme just waiting to disappoint people.
EXAMPLE C	
C. Post	*A Curated Fashion Marketplace for those above 50. An Idea based on a McKinsey Report:* I've been diving deep into the "silver generation" market (50+ age group) based on McKinsey's Report - "State of Fashion 2025" and discovered a massive opportunity that nobody seems to be tackling properly....
C. Uninformative	fashion, huh?

Content Type	Example Text
C. Informative agree	Yeah, this is interesting! I've definitely seen how most brands focus way too much on the younger crowd, forgetting that older generations have spending power and really know what they like. My mom, who's in her late 50s, always talks about how she wants more stylish options that are comfortable and made for her age group. She's not looking for fast fashion but more about quality and comfort. Plus, she's way more into online shopping now and values a good user-friendly interface. A marketplace like this could really cater to that gap. And honestly, keeping the tech simple but focused on the user experience is key. It's like when I helped my dad figure out Netflix – once it was easy and intuitive, he was hooked. For fashion, if you make the shopping experience seamless and relevant, it could go a long way. Combining style with function, like easy closures or fabrics that work for the older demographic, is so important. And tapping into a sense of community, with style discussions and reviews, is such a win since it offers more than just shopping – it's about feeling connected. Wonder if others have suggestions on curating brands or think the affiliate route is the best start to avoid that whole inventory hassle.
C. Informative disagree	I don't think this is something that would work long-term. I don't believe in the idea of separating fashions by age groups, especially if you are just thinking of it as something to be done as a financial venture. If your goals is to serve a demographic, then you should first focus on providing value and fostering a sense of artistic connection to their wardrobe experience. The same quality and design standards should apply to fashion no matter what age the clothing is designed for. Also, for people over 50, they may want more timeless pieces, but they also want clothes that make them feel alive and vibrant. There's almost this misconception that older folks suddenly only want cardigans and old-fashioned structures, but I've been around enough discussions on fashion with my mom and her friends to know that's not true. They want to feel cool, comfortable, and themselves just like everyone else. It's about enhancing lifestyles, not limitations.
C. Controversial	Wow, this sounds like a pretty good plan but are we just gonna ignore that fashion for people over 50 is probably ignored for a reason? Like, older people aren't scrolling Instagram or TikTok looking for style inspo, they buy what's comfortable and practical without the extra hassle. You think your idea gonna make them all of a sudden start caring about some "curated" marketplace online? Maybe there's a reason brands don't really chase that demographic so hard. Plus, building whole platforms for older generations is kinda overestimating their reliance on tech. It's a cool idea but I feel like the idea of targeting tech to an older market is getting recycled for retail businesses now and could miss the mark on understanding this demographic's actual tech engagement.

C.3 Additional results

C.3.1 Immediate contributions

Table 7: Robustness of Pooled Immediate Engagement Outcomes

	Reply made			*Number of replies*		
	(1)	(2)	(3)	(4)	(5)	(6)
Controversial	0.23***	0.24***	0.24***	0.34***	0.34***	0.34***
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Informative	0.20***	0.20***	0.20***	0.24***	0.24***	0.24***
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Control group mean	0.17	0.17	0.17	0.18	0.18	0.18
Pre-treatment controls	No	Yes	Yes	No	Yes	Yes
Week fixed effects	No	No	Yes	No	No	Yes
Community fixed effects	No	No	Yes	No	No	Yes
Observations	8509	8509	8509	8509	8509	8509

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Note: This table shows results from OLS regressions relative to the *uninformative* treatment arm. The *informative* arm pools responses from both the *informative agree* and the *informative disagree* treatments. In columns (1)-(3), the dependent variable is a binary variable that takes value 1 if the treatment comment receives a reply in the 24-hour period following treatment and zero otherwise. In columns (4)-(6), the dependent variable is a continuous variable given by the number of replies received in the 24-hour period following the treatment. Results are shown for the full sample ($n = 8,509$) across all columns. In Columns (2-3) and (5-6), the model specification controls for pre-treatment characteristics, i.e., the number of comments made by the treatment agent prior to the treatment comment, the number of comments on the treated post prior to receiving the treatment comment and the number of words in the treated post. In Columns (3) and (6), I additionally include week fixed effects and community fixed effects. Robust standard errors in parentheses. Stars are based on Romano–Wolf step-down adjusted p-values, controlling the family-wise error rate within each column (List et al., 2019; Romano and Wolf, 2010).

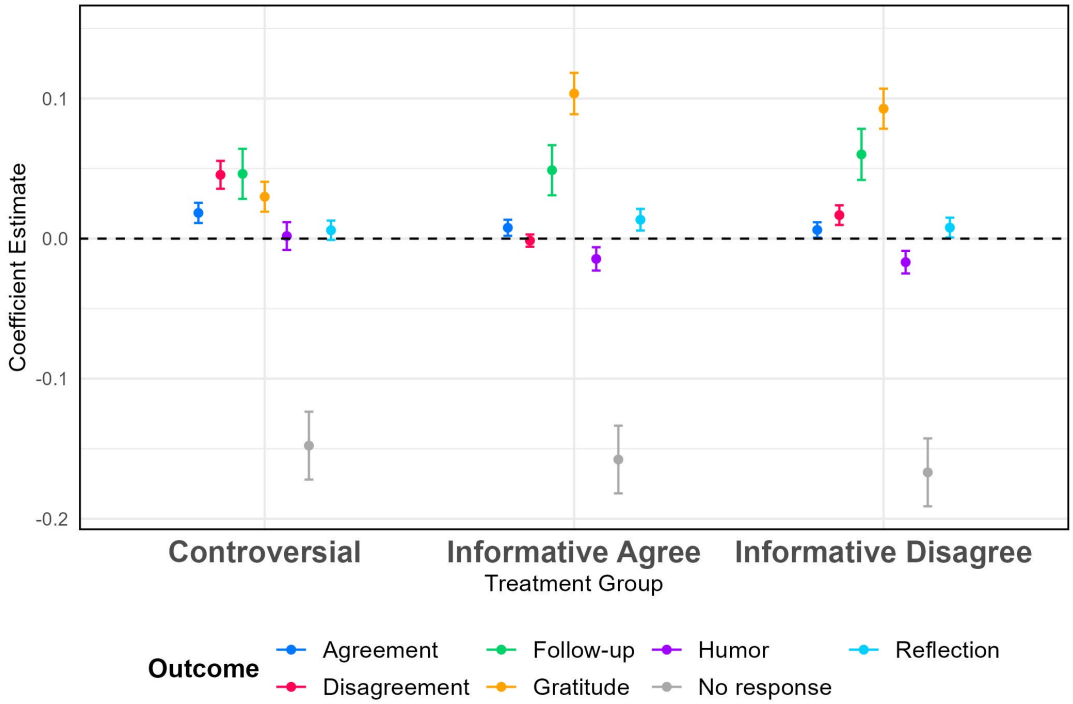
Table 8: Robustness of Immediate Engagement Outcomes

	Reply made			*Number of replies*		
	(1)	(2)	(3)	(4)	(5)	(6)
Controversial	0.23***	0.24***	0.24***	0.34***	0.34***	0.34***
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Informative Agree	0.19***	0.19***	0.19***	0.22***	0.22***	0.23***
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Informative Disagree	0.21***	0.21***	0.21***	0.25***	0.25***	0.25***
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Control group mean	0.17	0.17	0.17	0.18	0.18	0.18
Pre-treatment controls	No	Yes	Yes	No	Yes	Yes
Week fixed effects	No	No	Yes	No	No	Yes
Community fixed effects	No	No	Yes	No	No	Yes
Observations	8509	8509	8509	8509	8509	8509

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Note: This table shows results from OLS regressions by treatment arm relative to the *uninformative* treatment arm. The columns and analysis correspond to those reported in Table 7.

Figure 15: Types of replies given by treated authors on the treatment comments



Note: Each point represents a coefficient estimate for the specified treatment group relative to the *uninformative* group with 95% confidence interval. The regression specification includes pre-treatment control variables. Categories are mutually exclusive.

Table 9: Scores Received by Treatment Arm

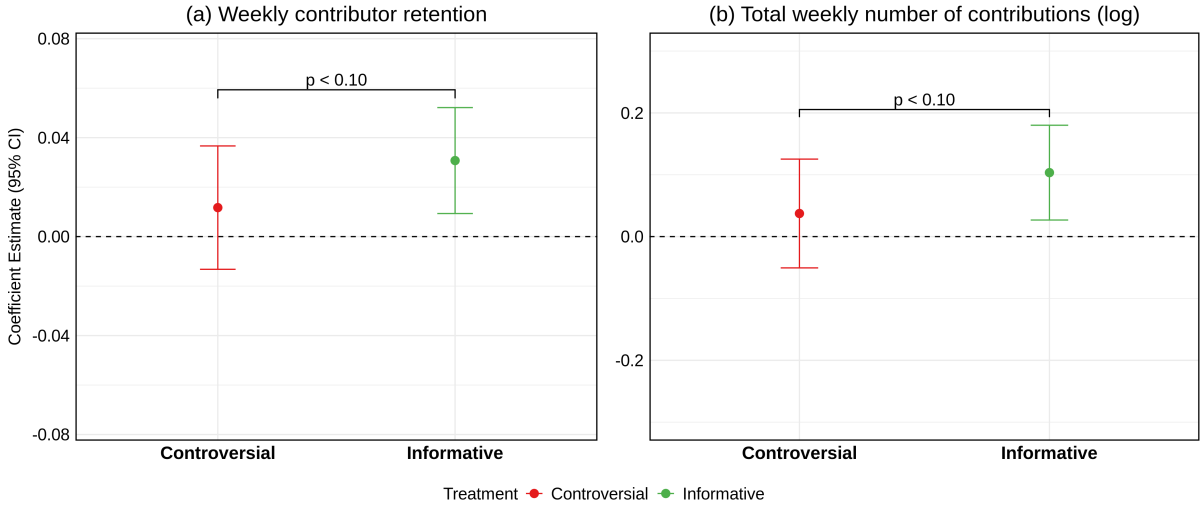
	Relative to T1 (*Uninformative*) (1)	Relative to T2 (*Informative agree*) (2)	Relative to T3 (*Informative disagree*) (3)
Controversial	1.31^{***}	-0.17	0.02
	(0.41)	(0.46)	(0.44)
Informative agree	1.48^{***}		0.19
	(0.23)		(0.27)
Informative disagree	1.29^{***}	-0.19	
	(0.18)	(0.27)	
Uninformative		-1.48^{***}	-1.29^{***}
		(0.23)	(0.18)
Pre-treatment controls	Yes	Yes	Yes
Control group mean	1.07	2.56	2.36
Observations	8,509	8,509	8,509

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Note: This table shows OLS regression results for all treatment arms. The dependent variable is the overall score received by a treatment comment, which can take any positive or negative integer value. Columns (1)-(3) show results for all treatment groups relative to the *uninformative* group, the *informative agree* group, and the *informative disagree* group as the baseline treatment groups, respectively. Robust standard errors in parentheses.

C.3.2 Sustained contributions

Figure 16: Weekly Total Contributions by Treatment Arm (Pooled)



Note: These figures plots regression coefficients from OLS regressions on (a) whether the treated author contributes again on the platform, and (b) the number of contributions they make in the 7-day period following the treatment, i.e. log (x+1), where x is the number of comments made. The *informative agree* and the *informative disagree* treatment arms are pooled together into one informative arm. Results are shown for the full sample ($n = 8,509$). Results are shown for the primary model specification, which controls for pre-treatment characteristics (i.e., the number of comments made by the treatment agent prior to the treatment comment, the number of comments on the treated post prior to receiving the treatment comment and the number of words in the treated post) and includes week and community fixed effects. Data are presented as coefficients with the vertical bars representing 95% confidence intervals derived from robust standard errors.

Table 10: Robustness of Pooled Total Weekly Contribution Outcomes

	Contribution made			*Number of contributions*		
	(1)	(2)	(3)	(4)	(5)	(6)
Controversial	0.01	0.01	0.01	0.03	0.03	0.04
	(0.01)	(0.01)	(0.01)	(0.05)	(0.05)	(0.04)
Informative	0.03**	0.03**	0.03***	0.10**	0.10**	0.10**
	(0.01)	(0.01)	(0.01)	(0.04)	(0.04)	(0.04)
Control group mean	0.76	0.76	0.76	1.93	1.93	1.93
Pre-treatment controls	No	Yes	Yes	No	Yes	Yes
Week fixed effects	No	No	Yes	No	No	Yes
Community fixed effects	No	No	Yes	No	No	Yes
Observations	8509	8509	8509	8509	8509	8509

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Note: This table shows results from OLS regressions relative to the *uninformative* treatment arm. The *informative* arm pools responses from both the *informative agree* and the *informative disagree* treatments. In columns (1)-(3), the dependent variable is a binary variable that takes value 1 if the treated user posts again on the platform (in the form of either an original post or a comment) in the 7 day period following the treatment and zero otherwise. In columns (4)-(6), the dependent variable is a continuous variable given by the log number of contributions made by the treated user in the 7-day period following the treatment, i.e., $\log(x + 1)$, where x is the total number of posts and comments made. Results are shown for the full sample ($n = 8,509$) across all columns. In Columns (2-3) and (5-6), the model specification controls for pre-treatment characteristics, i.e., the number of comments made by the treatment agent prior to the treatment comment, the number of comments on the treated post prior to receiving the treatment comment and the number of words in the treated post. In Columns (3) and (6), I additionally include week fixed effects and community fixed effects. Robust standard errors in parentheses. Stars are based on Romano–Wolf step-down adjusted p-values, controlling the family-wise error rate within each column (List et al., 2019; Romano and Wolf, 2010).

Table 11: Robustness of Weekly Total Contribution Outcomes

	Contribution made			*Number of contributions*		
	(1)	(2)	(3)	(4)	(5)	(6)
Controversial	0.01	0.01	0.01	0.03	0.03	0.04
	(0.01)	(0.01)	(0.01)	(0.05)	(0.05)	(0.04)
Informative Agree	0.03**	0.03**	0.03**	0.11*	0.11*	0.10**
	(0.01)	(0.01)	(0.01)	(0.05)	(0.05)	(0.05)
Informative Disagree	0.03**	0.03**	0.03**	0.09	0.09	0.11**
	(0.01)	(0.01)	(0.01)	(0.05)	(0.05)	(0.04)
Control group mean	0.76	0.76	0.76	1.93	1.93	1.93
Pre-treatment controls	No	Yes	Yes	No	Yes	Yes
Week fixed effects	No	No	Yes	No	No	Yes
Community fixed effects	No	No	Yes	No	No	Yes
Observations	8509	8509	8509	8509	8509	8509

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Note: This table shows results from OLS regressions by treatment arm relative to the *uninformative* treatment arm. The columns and analysis correspond to those reported in Table 10.

Table 12: Exposed Users' Longer-run Content Production (Weekly Retention)

	Post made			*Number of posts (log)*		
	All	New	Old	All	New	Old
	(1)	(2)	(3)	(4)	(5)	(6)
Controversial	−0.002	0.004	−0.001	0.002	0.018	−0.002
	(0.006)	(0.009)	(0.008)	(0.007)	(0.011)	(0.009)
Informative agree	0.014**	0.016*	0.011	0.021***	0.023**	0.019*
	(0.006)	(0.009)	(0.009)	(0.007)	(0.011)	(0.009)
Informative disagree	−0.006	0.002	−0.007	−0.0003	0.010	−0.002
	(0.007)	(0.010)	(0.009)	(0.008)	(0.012)	(0.009)
Pre-treatment controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	35,290	17,639	17,651	35,290	17,639	17,651

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Note: This table shows results from OLS regressions by treatment arm relative to the *uninformative* treatment arm. In columns (1)-(3), the dependent variable is a binary variable that takes value 1 if the treated user posts again on the platform in the 7 day period following the treatment and zero otherwise. In columns (4)-(6), the dependent variable is a continuous variable given by the log number of posts made by the treated user in the 7-day period following the treatment, i.e., $\log(x + 1)$, where x is the number of posts made. Columns (1) and (4) show results for the full sample of exposed users ($n = 35,290$), i.e., users who made a comment on the same post within one day of the treatment comment. Columns (2) and (5) show results for new accounts, i.e., exposed authors whose accounts were below the median age of the entire sample of exposed users ($n = 17,639$), and Columns (3) and (6) show results for old accounts, i.e., exposed authors whose accounts were of or above the median age of the entire sample ($n = 17,651$). Results are shown for the primary model specification, which controls for pre-treatment characteristics, i.e., the number of comments made by the treatment agent prior to the treatment comment, the number of comments on the treated post prior to receiving the treatment comment and the number of words in the treated post. Robust standard errors in parentheses.

C.4 Mechanisms

Table 13: Interaction Effects of Treatment Length and Treatments

	Immediate reply (24 hours)			*Total contributions (7 days)*		
	(1)	(2)	(3)	(4)	(5)	(6)
Controversial	0.305***	0.305***	0.305***	-0.025	-0.025	-0.025
	(0.035)	(0.035)	(0.035)	(0.031)	(0.031)	(0.031)
Informative agree	0.117***	0.117***	0.117***	0.063*	0.063*	0.063*
	(0.039)	(0.039)	(0.039)	(0.032)	(0.032)	(0.032)
Informative disagree	0.183***	0.183***	0.183***	-0.016	-0.016	-0.016
	(0.038)	(0.038)	(0.038)	(0.032)	(0.032)	(0.032)
Controversial	-0.002***			0.000		
× Length	(0.001)			(0.001)		
Informative agree		-0.001			-0.000	
× Length		(0.000)			(0.000)	
Informative disagree			-0.001**			0.000
× Length			(0.000)			(0.000)
Length	0.001***	0.001***	0.001***	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Pre-treatment controls	Yes	Yes	Yes	Yes	Yes	Yes
Community and Week FEs	Yes	Yes	Yes	Yes	Yes	Yes
Observations	8,509	8,509	8,509	8,509	8,509	8,509

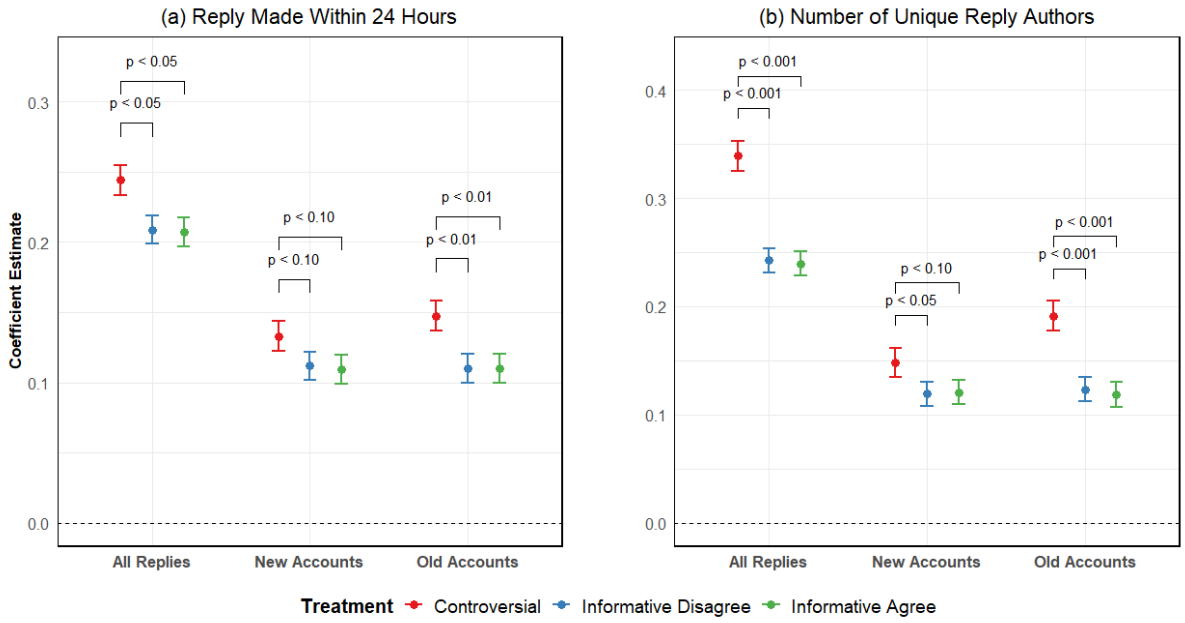
*** p<0.01, ** p<0.05, * p<0.1

Table 14: Interaction Effects of Sarcasm and Treatments

	Immediate reply (24 hours)		*Total contributions (7 days)*	
	(1)	(2)	(3)	(4)
Controversial	0.225***	0.225***	0.012	0.012
	(0.014)	(0.014)	(0.014)	(0.014)
Informative agree	0.189***	0.189***	0.027**	0.027**
	(0.014)	(0.014)	(0.013)	(0.013)
Informative disagree	0.208***	0.208***	0.029**	0.029**
	(0.014)	(0.014)	(0.013)	(0.013)
Controversial × Sarcasm		0.108**		-0.014
		(0.043)		(0.040)
Sarcasm	-0.018	-0.018	-0.014	-0.014
	(0.022)	(0.022)	(0.025)	(0.025)
Pre-treatment controls	Yes	Yes	Yes	Yes
Community and Week FEs	Yes	Yes	Yes	Yes
Observations	8,509	8,509	8,509	8,509

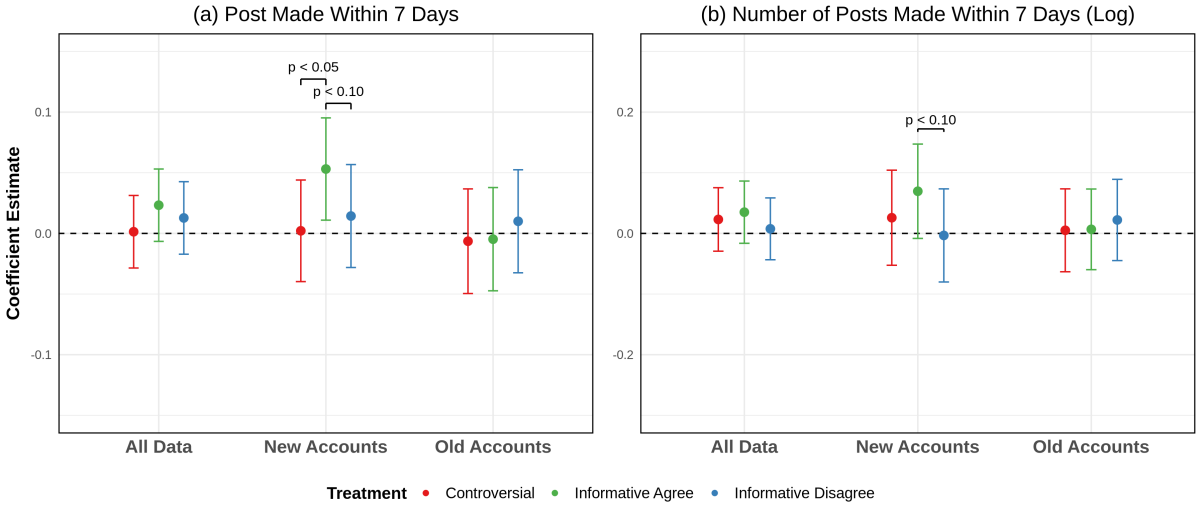
*** p<0.01, ** p<0.05, * p<0.1

Figure 17: Heterogeneity in Short-term Content Production by Treatment Arm



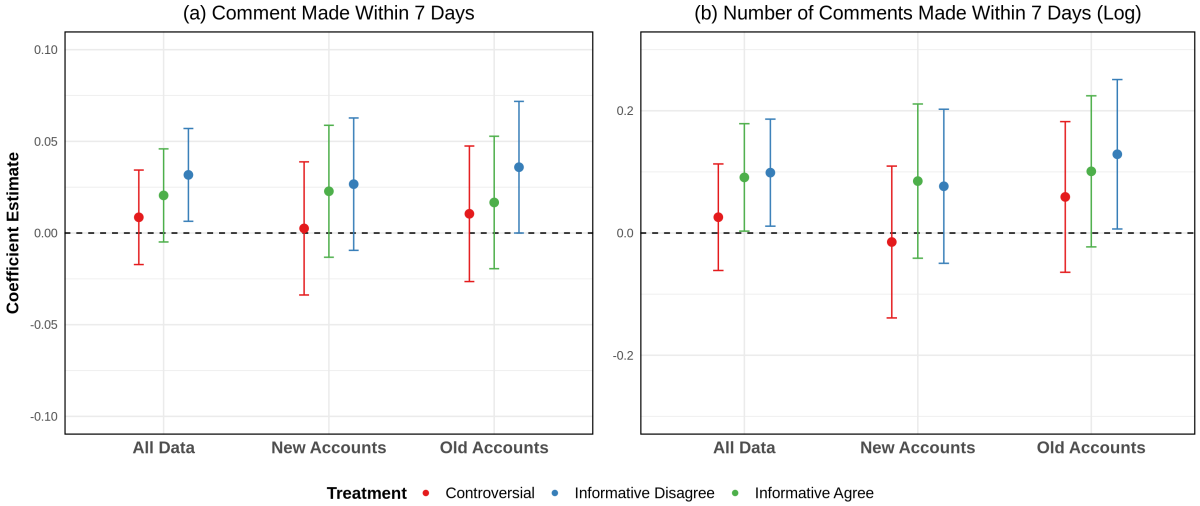
Note: These figures plots regression coefficients from OLS regressions by treatment arm relative to the *uninformative* arm. In Figure(a), the dependent variable is whether the treatment comment receives a reply from any user, new users (i.e. users who replied and were below the median age of all users who replied) and old users (i.e., users who replied and were of or above the median age of all those who replied). In Figure (b), the dependent variable is the number of replies received on the treatment comment in the 24-hour period following the treatment from all users, new users and old users respectively. Results are shown for the full sample ($n = 8,509$) under the primary model specification, which controls for pre-treatment characteristics, i.e., the number of comments made by the treatment agent prior to the treatment comment, the number of comments on the treated post prior to receiving the treatment comment and the number of words in the treated post. Data are presented as coefficients with the vertical bars representing 95% confidence intervals derived from robust standard errors.

Figure 18: Weekly Original Posts by Treatment Arm



Note: These figures plots regression coefficients from OLS regressions on (a) whether the treated author makes an original post again on the platform, and (b) the log number of original posts they make in the 7-day period following the treatment, i.e. log (x+1), where x is the number of original posts made. Results are shown for the full sample ($n = 8,509$), new accounts, i.e., treated authors whose accounts were below the median age of the entire sample ($n = 4,247$), and old accounts, i.e., treated authors whose accounts were of or above the median age of the entire sample ($n = 4,262$). Results are shown for the primary model specification, which controls for pre-treatment characteristics (i.e., the number of comments made by the treatment agent prior to the treatment comment, the number of comments on the treated post prior to receiving the treatment comment and the number of words in the treated post) and includes week and community fixed effects. Data are presented as coefficients with the vertical bars representing 95% confidence intervals derived from robust standard errors.

Figure 19: Weekly Comments by Treatment Arm



Note: These figures plots regression coefficients from OLS regressions on (a) whether the treated author makes a comment again on the platform, and (b) the log number of comments they make in the 7-day period following the treatment, i.e. log (x+1), where x is the number of comments made. The analysis corresponds to those reported in Figure 18.

Table 15: Robustness: Weekly Original Posts by Treatment Arm

	Post made			*Number of posts made*		
	(1)	(2)	(3)	(4)	(5)	(6)
Controversial	0.00	0.00	-0.01	0.02	0.03	0.01
	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)	(0.03)
Informative Agree	0.02	0.05**	-0.00	0.04	0.07	0.01
	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)	(0.03)
Informative Disagree	0.01	0.01	0.01	0.01	-0.00	0.02
	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)	(0.03)
Control group mean	0.47	0.49	0.45	0.65	0.73	0.57
Pre-treatment controls	No	Yes	Yes	No	Yes	Yes
Week fixed effects	No	No	Yes	No	No	Yes
Community fixed effects	No	No	Yes	No	No	Yes
Observations	8509	4247	4262	8509	4247	4262

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Note: This table shows results from OLS regressions relative to the *uninformative* treatment arm. In columns (1)-(3), the dependent variable is a binary variable that takes value 1 if the treated user makes an original post again on the platform in the 7 day period following the treatment and zero otherwise. In columns (4)-(6), the dependent variable is a continuous variable given by the log number of original posts made by the treated user in the 7-day period following the treatment, i.e., $\log(x + 1)$, where x is the total number of posts and comments made. Results are shown for the full sample ($n = 8,509$) in Columns (1) and (4), new accounts, i.e., treated authors whose accounts were below the median age of the entire sample ($n = 4,247$) in Columns (2) and (5), and old accounts, i.e., treated authors whose accounts were of or above the median age of the entire sample ($n = 4,262$) in Columns (3) and (6). In Columns (2-3) and (5-6), the model specification controls for pre-treatment characteristics, i.e., the number of comments made by the treatment agent prior to the treatment comment, the number of comments on the treated post prior to receiving the treatment comment and the number of words in the treated post. In Columns (3) and (6), I additionally include week fixed effects and community fixed effects. Robust standard errors in parentheses. Stars are based on Romano–Wolf step-down adjusted p-values, controlling the family-wise error rate within each column (List et al., 2019; Romano and Wolf, 2010).

Table 16: Robustness: Weekly Comments by Treatment Arm

	Comment made			*Number of comments made*		
	(1)	(2)	(3)	(4)	(5)	(6)
Controversial	0.01	0.00	0.01	0.03	-0.01	0.06
	(0.01)	(0.02)	(0.02)	(0.04)	(0.06)	(0.06)
Informative Agree	0.02	0.02	0.02	0.09*	0.08	0.10
	(0.01)	(0.02)	(0.02)	(0.04)	(0.06)	(0.06)
Informative Disagree	0.03**	0.03	0.04	0.10*	0.08	0.13*
	(0.01)	(0.02)	(0.02)	(0.04)	(0.06)	(0.06)
Control group mean	0.74	0.74	0.73	1.82	1.87	1.77
Pre-treatment controls	No	Yes	Yes	No	Yes	Yes
Week fixed effects	No	No	Yes	No	No	Yes
Community fixed effects	No	No	Yes	No	No	Yes
Observations	8509	4247	4262	8509	4247	4262

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Note: This table shows results from OLS regressions by treatment arm relative to the *uninformative* treatment arm. In columns (1)-(3), the dependent variable is a binary variable that takes value 1 if the treated user makes an original post again on the platform in the 7 day period following the treatment and zero otherwise. In columns (4)-(6), the dependent variable is a continuous variable given by the log number of original posts made by the treated user in the 7-day period following the treatment, i.e., $\log(x + 1)$, where x is the total number of posts and comments made. The columns and analysis correspond to those reported in Table 15.

Table 17: Robustness of Weekly and Monthly Contributor Retention to Fixed Effects

	Post made within 7 days			*Post made within 30 days*		
	All	New	Old	All	New	Old
	(1)	(2)	(3)	(4)	(5)	(6)
Uninformative	0.002	0.001	0.010	0.017	0.021	0.016
	(0.015)	(0.021)	(0.022)	(0.014)	(0.019)	(0.020)
Informative agree	0.020	0.045**	0.005	0.019	0.045**	-0.002
	(0.015)	(0.021)	(0.021)	(0.014)	(0.019)	(0.020)
Informative disagree	0.020	0.022	0.023	0.033**	0.037*	0.033*
	(0.015)	(0.021)	(0.021)	(0.014)	(0.019)	(0.020)
Pre-treatment controls	Yes	Yes	Yes	Yes	Yes	Yes
Week fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Community fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	8509	4247	4262	8509	4247	4262

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Note: This table shows results from OLS regressions by treatment arm relative to the *controversial* treatment arm. In columns (1)-(3), the dependent variable is a binary variable that takes value 1 if the treated user posts again on the platform in the 7 day period following the treatment and zero otherwise. In columns (4)-(6), the dependent variable is a binary variable that takes value 1 if the treated user posts again on the platform in the 30 day period following the treatment and zero otherwise. Results are shown for the full sample ($n = 8,509$) in Columns (1) and (4), new accounts, i.e., treated authors whose accounts were below the median age of the entire sample ($n = 4,247$) in Columns (2) and (5), and old accounts, i.e., treated authors whose accounts were of or above the median age of the entire sample ($n = 4,262$) in Columns (3) and (6). Results are shown for the model specification with community and week fixed effects alongside controls for pre-treatment characteristics, i.e., the number of comments made by the treatment agent prior to the treatment comment, the number of comments on the treated post prior to receiving the treatment comment and the number of words in the treated post. Robust standard errors in parentheses.

Table 18: Longer-run Content Production by Treatment Arm (Monthly Retention)

	Post made			*Number of posts (log)*		
	All	New	Old	All	New	Old
	(1)	(2)	(3)	(4)	(5)	(6)
PANEL A: RELATIVE TO THE "UNINFORMATIVE" GROUP						
Controversial	−0.02	−0.02	−0.02	0.04	0.04	0.03
	(0.01)	(0.02)	(0.02)	(0.04)	(0.05)	(0.05)
Informative agree	0.00	0.04*	−0.03	0.06	0.11**	0.01
	(0.01)	(0.02)	(0.02)	(0.04)	(0.05)	(0.05)
Informative disagree	0.01	0.01	0.02	0.06	0.03	0.10**
	(0.01)	(0.02)	(0.02)	(0.04)	(0.05)	(0.05)
PANEL B: RELATIVE TO THE "CONTROVERSIAL" GROUP						
Uninformative	0.02	0.02	0.02	−0.04	−0.04	−0.03
	(0.01)	(0.02)	(0.02)	(0.04)	(0.05)	(0.05)
Informative agree	0.02	0.06***	−0.01	0.02	0.06	−0.02
	(0.01)	(0.02)	(0.02)	(0.04)	(0.05)	(0.05)
Informative disagree	0.03**	0.03	0.03*	0.02	−0.02	0.07
	(0.01)	(0.02)	(0.02)	(0.04)	(0.05)	(0.05)
Pre-treatment controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	8509	4247	4262	8509	4247	4262

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Note: This table shows results from OLS regressions by treatment arm. Panel A shows results relative to the *uninformative* treatment arm whereas Panel B shows results relative to the *controversial* treatment arm. In columns (1)-(3), the dependent variable is a binary variable that takes value 1 if the treated user posts again on the platform in the 30 day period following the treatment and zero otherwise. In columns (4)-(6), the dependent variable is a continuous variable given by the log number of posts made by the treated user in the 30-day period following the treatment, i.e., $\log(x + 1)$, where x is the number of posts made. Results are shown for the full sample ($n = 8,509$) in Columns (1) and (4), new accounts, i.e., treated authors whose accounts were below the median age of the entire sample ($n = 4,247$) in Columns (2) and (5), and old accounts, i.e., treated authors whose accounts were of or above the median age of the entire sample ($n = 4,262$) in Columns (3) and (6). Results are shown for the primary model specification, which controls for pre-treatment characteristics, i.e., the number of comments made by the treatment agent prior to the treatment comment, the number of comments on the treated post prior to receiving the treatment comment and the number of words in the treated post. Robust standard errors in parentheses.

Table 19: Interaction Effects of Agreement and Treatments for New and Old Users

	Total contributions (7 days)					
	New users			Old users		
	(1)	(2)	(3)	(4)	(5)	(6)
Controversial	-0.027	-0.027	-0.027	0.039	0.039	0.039
	(0.030)	(0.030)	(0.030)	(0.031)	(0.031)	(0.031)
Informative agree	-0.030	-0.030	-0.030	0.086***	0.086***	0.086***
	(0.031)	(0.031)	(0.031)	(0.030)	(0.030)	(0.030)
Informative disagree	-0.005	-0.005	-0.005	0.064**	0.064**	0.064**
	(0.031)	(0.031)	(0.031)	(0.030)	(0.030)	(0.030)
Controversial	0.004			-0.003		
× Community agreement	(0.003)			(0.003)		
Informative agree		0.008***			-0.007**	
× Community agreement		(0.003)			(0.003)	
Informative disagree			0.003			-0.003
× Community agreement			(0.003)			(0.003)
Community Agreement	0.001	0.001	0.001	0.005**	0.005**	0.005**
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Pre-treatment controls	Yes	Yes	Yes	Yes	Yes	Yes
Community & Week FEs	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4,254	4,254	4,254	4,255	4,255	4,255

*** p<0.01, ** p<0.05, * p<0.1