

Algorithms and Parasitic Content

Hannah Solheim

Nancy Wang
Job Market Paper

October 13, 2026

([Click here for most recent version](#))

Abstract

To shed light on how recommendation algorithms are optimized to maintain user attention with content that makes users worse off, we introduce the concept of *parasitic content* and show its relevance in two experiments. We first define parasitic content in a theoretical model of a user with dual selves interacting with an algorithm. Parasitic content is a user-specific temptation good, targeted to a user tagged as susceptible based on her consumption of other, non-tempting content. We then study this phenomenon on TikTok. In an eight-week online field experiment, 424 young women were incentivized to use randomly assigned TikTok accounts seeded with varying levels of body-positive engagement. Leveraging second-by-second watch histories spanning 1.38 million videos, we find that body-positive seeding caused the algorithm to serve not only more body-positive *but also* more weight-loss content, and worsened body image by 0.27 standard deviations. To disentangle the roles of body-positive and weight-loss content in this decline, we conducted a mechanism experiment with 1,000 young women. Weight-loss content worsened body image and mood, while body-positive content improved them. Consistent with temptation, nearly half of participants preferred *ex ante* to avoid weight-loss content—37% of whom were willing to pay to do so—yet consumed it when served. Thus, for many young women, weight-loss content is parasitic to body-positive content on TikTok. Watch histories from the field experiment further demonstrate that weight-loss content served in response to body-positive seeding increased ad revenue and per-video watch time. By learning how consumption of one good predicts consumption of others, algorithms can identify and deliver personalized temptations, increasing revenue at the expense of user well-being.

Solheim: Columbia University, h.solheim@columbia.edu; Wang: Massachusetts Institute of Technology, wangn@mit.edu. We thank David Autor, Esther Duflo, Daron Acemoglu, Frank Schilbach, Ebonya Washington, Elizabeth Ananat, and Suresh Naidu for incredible support. We are also grateful to Abhijit Banerjee, Drew Fudenberg, Sendhil Mullainathan, Colin Camerer, Mark Dean, Charles Angelucci, Alexey Makarin, Ro’ee Levy, and seminar participants at MIT, Columbia, BEAM, the Caltech Behavioral and Experimental Workshop, the CEPR Workshop on Media, Technology, Politics, and Society, and NEUGD for helpful feedback, and to Rafael Jiménez-Durán for a productive discussion. Halie Boyer, Slawomir Frankowski, Armaan Gill, Annie Li, Angela Shi, Weixiao Sun, Diya Tekriwal, and Maggie Zhang provided excellent research assistance. We acknowledge generous financial support from J-PAL North America’s U.S. HCDI, the MIT George and Obie Shultz Fund, and the Institute for Humane Studies (IHS018684). Experiments were approved by IRB (MIT: E-5573, 2408001393A001, E-6079, E-7374; Columbia: AAV1224, AAV3856, AAV4058, ACYY0989) and pre-registered prior to treatment (AEARCTR-0014958, AEARCTR-0017457).

1 Introduction

Online content providers, such as social media sites, streaming services, and news sites, frequently employ recommendation algorithms that determine what users see. Powered by machine learning and behavioral data on billions of individuals, recommendation algorithms predict what a given user is most likely to consume, and they serve content accordingly.¹ Prima facie, one may think that these algorithms benefit users by lowering search costs and surfacing desired content, while unwanted recommendations can be easily skipped. Yet algorithms may have adverse effects—for instance, recent lawsuits allege that social media algorithms fuel the youth mental health crisis by promoting content that encourages social comparison and disordered eating (Kang, 2026). Consistent with these concerns, there is evidence that social media reduces well-being and increases depression (Mosquera et al., 2020; Allcott et al., 2020; Braghieri et al., 2022); however, evidence on the mechanisms (Allcott et al., 2022; Bursztyn et al., 2025) and specifically the role of algorithms is scarce (Aridor et al., 2024). If algorithms help users find content they want, why do they often appear to make users worse off?

This paper investigates how algorithms learn to systematically serve personalized temptations—what we call *parasitic content*—that increase user engagement while reducing user well-being. Recent research suggests that users engage with content that reduces their well-being (Kleinberg et al., 2023, 2024; Beknazar-Yuzbashev et al., 2024); while this work helps to explain why platforms may serve such content, there is little direct evidence on the algorithmic mechanisms that generate these recommendations. We propose parasitic content as one such mechanism.

Parasitic content has two defining properties. First, it is a *temptation good* (Gul and Pesendorfer, 2001; Fudenberg and Levine, 2006; Banerjee and Mullainathan, 2010) because a user derives positive utility from it at the moment of consumption, even though it lowers her long-run payoff. We use the long-run payoff as our welfare benchmark (O’Donoghue and Rabin, 1999) and refer to it as well-being. Second, it is *personalized* because the algorithm serves it in response to her past engagement with other, non-tempting content. The content is parasitic in the sense that the user does not seek it, yet it specifically enters her feed following past engagement that tags the user as susceptible to the tempting content.

To formalize these properties and characterize when parasitic content arises, we present a model of a sequential game between a user with dual selves (Fudenberg and Levine, 2006) and a platform recommending content from several genres. The user’s myopic short-run self chooses a consumption level for each recommendation based on her private genre-specific

¹We use consumption, engagement, and watch time interchangeably throughout this paper.

tastes and temptation for a focal, potentially parasitic, genre. Her long-run self receives a payoff from taste- but not temptation-driven consumption (Banerjee and Mullainathan, 2010). After observing the user’s consumption of content from a potential host genre, the platform’s algorithm updates its beliefs about her engagement with *other* genres and ranks the remaining content in a feed maximizing expected engagement. Our key assumption is that users with greater taste for the host genre also tend to consume more focal-genre content. For a given user, the focal genre is *parasitic* to the host genre if the algorithm supplies more focal-genre content in response to the user’s consumption of host-genre content (the *host-parasite criterion*), and the user’s long-run well-being would be higher if she could block the focal genre (the *well-being criterion*). Both criteria are met when the user’s taste for the host genre is sufficiently high and her temptation for the focal genre sufficiently strong.

Beyond these defining criteria, our model yields implications for *average well-being*, *consumption*, and *revenue*. First, if parasitic content is sufficiently prevalent and detrimental, algorithmic personalization based on engagement with the host genre reduces average user well-being; second, sophisticated users, who anticipate their temptation, prefer ex ante to avoid parasitic content but consume it if served in the future; third, parasitic content increases engagement and ad revenue.

We examine these criteria and implications in two experiments: a field experiment on TikTok—the largest manipulation of TikTok’s algorithm to date—and a mechanism experiment. In our application, body positivity is the potential host genre and weight loss is the focal, potentially parasitic, genre. We choose this setting because TikTok is an extremely popular yet understudied social media platform (Aridor et al., 2024), and there are allegations that it promotes content related to weight loss and disordered eating to girls and young women (Greene et al., 2022). The field experiment tests the host-parasite criterion and the implications for average well-being and revenue; the mechanism experiment tests the well-being criterion and the consumption implication.

Our field experiment on TikTok enrolled 424 young women in the U.S., UK, and Canada and lasted eight weeks for each participant. Testing the host-parasite criterion requires exogenous variation in the level of body-positive consumption observed by the algorithm; testing the well-being and revenue implications requires exposing participants to the resulting feeds and measuring their well-being and engagement. Because TikTok relies heavily on watch time to rank content (TikTok, 2020), the challenge is to vary the watch histories observed by the algorithm independently of participants’ preferences.²

²Levy (2021), which was not studying parasitic content, met both methodological requirements by encouraging users to follow new political pages on Facebook. This approach is less suited to TikTok, which relies heavily on watch time to rank content.

We overcome this challenge by introducing a new methodology. We created TikTok accounts, followed pre-selected influencers on these accounts, and scrolled through their content to generate watch histories, a process we call *seeding*. Participants were randomly assigned to accounts that differed in their seeded histories and were incentivized to scroll for an initial 30-minute scrolling period, allowing us to capture the algorithm’s response to seeding.³ Participants randomized into the *body-positive treatment* received accounts seeded with content from body-positive influencers, while those randomized into the *control* received accounts without seeding. To assess whether weight-loss recommendations specifically follow body-positive engagement—or whether body-positive engagement instead plays a protective role by reducing weight-loss content relative to other appearance-related engagement—some participants were randomized into a *fashion treatment* and received accounts seeded with content from fashion influencers. Participants were free to engage with content as they liked, allowing us to examine the model’s revenue implication.

To test how the feeds affected well-being, body image was measured via surveys two weeks before randomization and immediately after the initial scrolling period. Participants were then incentivized to use their assigned accounts for two more weeks before completing a third survey. They completed an endline survey two weeks after incentives ended, allowing us to examine whether the effects persisted beyond the treatment period. Because well-being is multifaceted and difficult to capture (Benjamin et al., 2014), we focus on body image, a particularly relevant dimension in our setting, and note that worse body image predicts poorer subsequent mental and physical health, even after accounting for factors such as BMI (Stice and Bearman, 2001; Neumark-Sztainer et al., 2006; Vannucci and Ohannessian, 2018).

Content supply and consumption were measured using the assigned accounts’ watch histories, which provide a second-by-second record of all content served and cover 1.38 million videos in total.⁴ Metadata were scraped for 99.8% and MP4 files retrieved for 89.5% of unique videos; content was then classified by genre, and consumption was inferred from the elapsed time between consecutive timestamps. Watch histories were also collected from 262 participants’ personal TikTok accounts, covering 11 million videos. To our knowledge, we are the first in the economics literature to collect and analyze TikTok watch histories.

To test whether body-positive engagement increases the algorithm’s supply of weight-loss content (the host-parasite criterion), we examine weight-loss content supply across treatments. Immediately after seeding, participants in the body-positive treatment received 3.5 (2.3pp) *more* weight-loss videos during the initial scrolling period than the control

³All participants used TikTok prior to the experiment and retained access to their own TikTok accounts throughout the experiment.

⁴TikTok content comprises primarily videos, but users may also encounter posts containing collections of still images. For brevity, we refer to all content as “videos” throughout.

($p < 0.001$). As a manipulation check, body-positive seeding increased body-positive content by 15.8 (10.7pp) videos relative to the control ($p < 0.001$). Control participants received no body-positive videos and fewer than one weight-loss video on average. These differences persisted over two weeks, attenuating slightly as the algorithm responded to participant behavior, such as whether a participant watched a video fully or quickly scrolled past it. Even compared to participants in the fashion treatment, who received 15.1 (10.5pp) more fashion videos ($p < 0.001$) and 0.7 (0.5pp) more weight-loss videos than the control ($p = 0.048$), the body-positive treatment group received 2.8 more weight-loss videos ($p < 0.001$). Thus, body-positive engagement instigates weight-loss content.

We further support these findings by exploiting a source of quasi-random variation: during seeding, the browsers we used crashed haphazardly, truncating some watch histories at a few hundred videos while others reached thousands. Doubling the number of body-positive videos seeded caused the algorithm to serve 0.44 *additional* weight-loss videos ($p = 0.043$), whereas an analogous increase in fashion videos caused the algorithm to serve 0.56 *fewer* weight-loss videos ($p = 0.165$); the difference between estimates is significant ($p = 0.022$). Together, our results establish the host-parasite criterion: body-positive engagement increases the algorithm’s supply of weight-loss content—a seemingly antithetical genre.

Multiple robustness checks confirm our finding. Excluding weight-loss videos surfaced from seeded body-positive influencers yields similar estimates. Inadvertent weight-loss seeding also cannot explain the result. During seeding, we avoided influencers who promoted weight loss, and weight-loss content represented less than 5% of seeded material for all accounts; controlling for the amount or presence of seeded weight-loss content yields larger estimates. Finally, the finding does not reflect TikTok’s conflating weight-loss and body-positive content. If this were the case, engagement with either should predict the subsequent supply of both. Instead, over two weeks of account use, weight-loss (body-positive) supply was predicted specifically by the previous day’s weight-loss (body-positive) watch time.

We next test the model’s contention that, by serving parasitic content in response to engagement with body-positive content, recommendation algorithms may diminish average user well-being. We estimate treatment effects on a pre-registered body image index (Anderson, 2008) comprising validated body image scales and self-reported behaviors related to body image. Immediately after the initial scrolling period, participants in the body-positive treatment reported worse body image than the control (-0.23 standard deviations, $p = 0.008$). The effect remained similar in magnitude after two weeks of treatment (-0.27 SD, $p = 0.010$) and at endline, two weeks after treatment ended (-0.21 SD, $p = 0.041$). These results support the model’s implication for average well-being.

At this point, one possibility for the above result is that body-positive, rather than

weight-loss, content caused the decline in body image. We disentangle the effects of body-positive and weight-loss content—and thereby test the well-being criterion—in a mechanism experiment that randomly assigned 1,000 young women to controlled feeds varying in genre composition, some of which featured body-positive or weight-loss content drawn from the field experiment. Afterward, participants completed a survey measuring body image and mood, which were combined into a pre-registered index. Body-positive content alone *raised* the index by 0.27 standard deviations ($p = 0.010$); by contrast, weight-loss content alone *reduced* it by 0.23 SD ($p = 0.043$), supporting the well-being criterion. These findings also show that body-positive content alone cannot explain the decline observed in the field, whereas weight-loss content contributed to it.

In the same mechanism experiment, we also test the consumption implication: if weight-loss content is a temptation good, sophisticated users prefer to avoid it *ex ante* yet consume it when served. Before assignment to a feed, participants chose whether to include or exclude various genres from their feed using an incentive-compatible Becker-DeGroot-Marschak mechanism, with choices implemented for a small random subsample. Interpreting these choices as preferences requires that the act of choosing not directly affect welfare (Bernheim et al., 2026); under this assumption, we compare participants’ choices with their consumption, as they were free to watch or skip videos. Approximately half preferred to exclude weight-loss content—37% of whom were willing to pay to do so, suggesting a self-control problem—yet watched weight-loss videos for 4.2 seconds (52%) longer per video than cryptocurrency content ($p < 0.001$), which was included in some feeds to benchmark how young women behave when served undesired and non-tempting content.⁵ Because the feeds omitted comments and other potential sources of engagement (Beknazar-Yuzbashev et al., 2025b; Donati and Song, 2026), this finding is consistent with weight-loss videos themselves operating as temptation goods. Further bolstering this interpretation, participants who preferred to exclude weight-loss content experienced larger declines in body image and mood after exposure to it (-0.32 SD, $p = 0.049$), consistent with their anticipation of its effects.

Open-ended responses clarify the psychology and consequences of temptation. One participant who chose to exclude weight-loss content explained, *“Even with the strongest will, it’s hard to not listen to society telling you [to] be thinner no matter your weight.”* Importantly, over half of those who chose to exclude weight-loss content described it as detrimental to their well-being. One participant willing to pay said, *“I...have a history of eating disorders, restricting myself, and compulsive exercising. I know my triggers, and my well-being, mood, and liking myself are more valuable than money.”* These responses illustrate how weight-loss content can be damaging in the long run yet difficult to resist in the moment.

⁵In open-ended responses, 96% of participants who chose to exclude cryptocurrency cited disinterest.

Together, these results show that for many young women, weight-loss content is parasitic to body-positive content on TikTok.

Why does TikTok serve parasitic content? Using consumption data from the field experiment, we find that weight-loss content served in the body-positive treatment was highly engaging. Consistent with our model’s revenue implication, these videos generated more watch time per video than content served in the control and fashion treatments—two proxies for content that the algorithm might otherwise have served—and similar watch time per video to other content served in the body-positive treatment. Using scraped metadata identifying advertisements (ads), we further find that weight-loss ads served in the body-positive treatment were over 70% more likely to generate a billable view than ads in any other category (p -values between 0.024 and 0.063). Under an illustrative set of assumptions, we estimate that serving weight-loss content to young U.S. women who want body-positive content increases annual platform revenue by \$96–124 million, depending on the comparison. These findings connect to a broader literature showing that advertising incentives shape media content supply (Cagé et al., 2024; Beknazar-Yuzbashev et al., 2025a).

Parasitic content demonstrates how big data and machine learning offer firms a new way to learn about and exploit their consumers. Previous work in behavioral industrial organization shows how firms exploit broadly held consumer biases, including self-control problems (DellaVigna and Malmendier, 2004, 2006; Allcott et al., 2022), inattention (Gabaix and Laibson, 2006; Loewenstein and Wojtowicz, 2024; Allcott et al., 2025), and susceptibility to cues (Laibson, 2001; Shleifer et al., 2025); see Heidhues and Köszegi (2018) for a review. Behavioral data and machine learning instead allow firms to infer *each* consumer’s vulnerabilities and deliver *personalized* temptations. In our application, women may seek body-positive content to improve their body image, but doing so reveals the body-image concerns that make weight-loss content tempting in the first place.⁶

Weight-loss content on TikTok is one example of a much broader class of parasitic content. Other parasitic pairings on social media may include sports betting ads served in response to engagement with video game content and pornography surfaced in response to engagement with male fitness content. As technologies advance, such personalization may become more consequential (Rambachan et al., 2020; Agan et al., 2023; Kleinberg et al., 2024; Acemoglu et al., 2025; Campante et al., 2025). For instance, firms that use generative AI tools to interact with consumers gain rich information about their desires, doubts, and states of mind. This information may be more revealing than social media consumption and allow firms to identify and exploit consumers’ vulnerabilities with much greater precision.

⁶We are agnostic about whether firms knowingly supply parasitic content or simply cannot distinguish temptation-driven from genuine engagement. Regardless, there is a financial incentive to supply it.

2 Setting: TikTok and Body Image

With nearly two billion users worldwide, TikTok is one of the world’s most popular social media platforms, and usage is particularly high among American teenage girls, who average nearly two hours per day on the platform (Rothwell, 2023). Despite its popularity, TikTok has received relatively little attention in the economics literature compared to platforms such as Facebook and X, perhaps because TikTok was not available in the U.S. until 2018. Its introduction marked a shift in the social media landscape from feeds shaped largely by users’ choices, such as whom to follow, toward feeds curated by algorithms that infer users’ interests from their behavior (Aridor et al., 2024).

TikTok’s main feed, the *For You Page*, is an infinite feed of videos algorithmically ranked using implicit signals of interest such as watch time. Consequently, these videos are drawn almost entirely from accounts the user does not follow or know offline. The algorithm has generated fascination due to its uncanny ability to predict what users want to watch (Herrman, 2019; Merrill et al., 2025), but its workings remain opaque; while the TikTok algorithm is understood to optimize for engagement and profit, little is known about the recommendation policy used to achieve these objectives (Smith, 2021).

Social media recommendations may affect well-being and, of particular relevance to this paper, body image. Conceptualized by Schilder (1935), body image is “a person’s perceptions, thoughts, and feelings about his or her body” (Grogan, 2002). Because body image is subjective, it can be shaped by social influences such as media and differ substantially from outside observers’ perceptions; for instance, many young women with anorexia nervosa believe themselves to be much heavier than they are (Smeets et al., 1999). Substantial academic and clinical effort has gone into developing validated survey instruments of body image. Using these instruments, scholars have documented that negative body image affects half of adolescent girls (Quick et al., 2013) and predicts subsequent disordered eating, depression, anxiety, and lower levels of exercise, even after accounting for body mass index (BMI) (Stice and Bearman, 2001; Neumark-Sztainer et al., 2006; Rohde et al., 2015; Vannucci and Ohannessian, 2018).

TikTok was once heralded as a force for body positivity (Bala, 2021), a popular social movement celebrating diverse body types, but its algorithm has since been accused of promoting content related to weight loss and disordered eating (Greene et al., 2022). Such content may be particularly consequential for adolescent girls, among whom the negative association between social media use and mental health is especially pronounced, with body image proposed as a potential mechanism (Haidt, 2024).

3 A Model of Algorithmic Delivery of Parasitic Content

To define parasitic content and frame our experiments, we build on models of temptation goods (Gul and Pesendorfer, 2001; Fudenberg and Levine, 2006; Banerjee and Mullainathan, 2010): goods whose appeal at the moment of consumption leads users to consume more than is optimal in the long run. We situate this temptation framework within an engagement-maximizing platform that uses an algorithm to tailor content based on what it learns about the user’s preferences from her past consumption. In particular, past consumption of content from a potential “host” genre predicts consumption of a different, potentially tempting focal genre. Distinct from the prior literature, we show that this combination of temptation, algorithmic learning, and engagement maximization can lead the algorithm to serve more content from the focal genre in response to consumption of content from the host genre, regardless of whether the focal genre improves or reduces the user’s long-run well-being. In the latter case, the focal genre is *parasitic* to the host genre for that user.

3.1 Environment and Payoffs

The platform recommends content from three genres indexed by $k \in \{H, P, O\}$, where H denotes the potential host genre, P the focal, potentially parasitic genre, and O all other content genres. Within a genre, the platform has an infinite supply of homogeneous content items. A user’s session consists of a sequence of slots indexed by $n = 1, 2, \dots$, and the platform fills each slot with one item. The resulting feed is denoted by $r = (r_1, r_2, \dots)$, where $r_n \in \{H, P, O\}$ denotes the genre of the item in slot n . The user stays on the platform for N slots, after which she exits. For $n \leq N$, her short-run self chooses consumption $x_n \geq 0$; consumption is zero thereafter, $x_{n>N} = 0$. In our application, items are videos, and x_n measures the number of times the user watches the video in slot n : $x_n < 1$ represents watching part of the video and $x_n \geq 1$ watching it once or more.

User Preferences and Payoffs. We assume that the user’s session length N is exogenously drawn from a geometric distribution with continuation probability $\kappa \in (0, 1)$. The user’s type is independent of N and given by vector $\theta = (u_H, u_P, u_O, \tau) \in \mathbb{R}^3 \times \mathbb{R}_{\geq 0}$, where $u_k \in \mathbb{R}$ is her taste for genre $k \in \{H, P, O\}$ and $\tau \geq 0$ is her temptation for genre P . Denote the joint density of the random elements of the user’s type vector by $f(u_H, u_P, u_O, \tau)$.

The user’s short-run self chooses consumption of each item. We assume that the short-run self is myopic, valuing only the immediate payoff from the current item. This assumption simplifies the model by ruling out strategic consumption intended to influence the platform’s recommendations, while capturing the impulsive nature of much online content consumption

as discussed in [Kleinberg et al. \(2023\)](#). In slot $n \leq N$ with $r_n = k$, the short-run self's payoff from consuming $x_n \geq 0$ is

$$u_k x_n + \mathbb{1}\{k = P\} \tau x_n - \frac{\gamma}{2} c_k(n) x_n^2,$$

where $\gamma > 0$ and $c_k(n) \equiv \sum_{m \leq n} \mathbb{1}\{r_m = k\}$ is the number of genre- k videos in the first n slots. The first two terms capture taste and, for genre P , temptation ([Banerjee and Mullainathan, 2010](#)). The final term creates diminishing returns as more items from the same genre are served, which yields a closed-form solution and enables the targeting mechanism described later; this specification is also consistent with platforms' own reports that they diversify feeds to prevent user boredom ([Smith, 2021](#)). Observe that for the first video of each genre, optimal consumption is proportional to $\max\{0, u_k + \mathbb{1}\{k = P\} \tau\}$, a result we use later.

The user's long-run self makes no choice. Instead, its payoff represents the user's well-being. Following [Banerjee and Mullainathan \(2010\)](#), temptation is valued only by the short-run self and does not enter the long-run self's payoff,

$$V = \sum_{n=1}^N \left(u_{r_n} x_n - \frac{\gamma}{2} c_{r_n}(n) x_n^2 \right).$$

Platform Payoff. The platform determines the order in which items are served to the user in a session, that is, the feed r . Since content within a genre is homogeneous, the platform's problem amounts to choosing the genre of the item shown in each slot. The platform maximizes ad revenue, which we assume is proportional to total consumption. Given a feed r and consumption x_n in each slot n , the platform's payoff is

$$\sum_{n=1}^{\infty} x_n = \sum_{n=1}^N x_n.$$

Information Structure. Session length N and type θ are private to the user, while the platform knows their distributions, characterized by κ and f . Note that since its objective depends only on consumption, the platform can equivalently base its decisions on the distribution of consumption across genres conditional on observed behavior, which could be learned directly from user data.

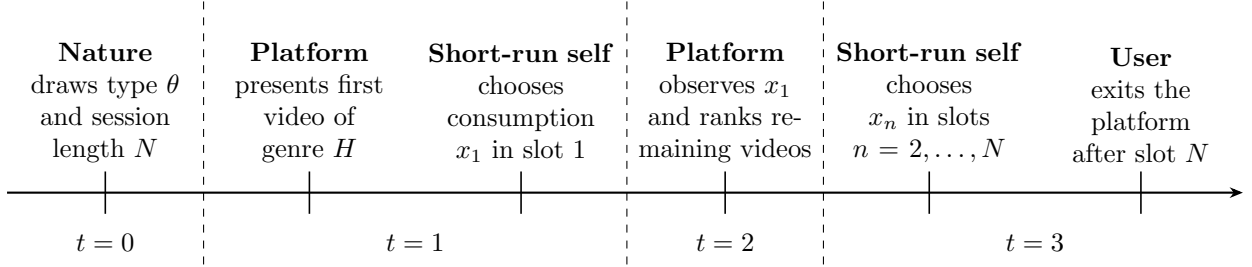
Distributional Assumptions. We impose the following restrictions on the distribution of user types. We assume that u_H, u_P, u_O , and τ have finite second moments and that each genre k generates positive expected consumption under the prior, $\mathbb{E}[\max\{0, u_k + \mathbb{1}\{k = P\} \tau\}] > 0$. More substantively, we assume:

Assumption 1: The variables u_H and u_O are independent under distribution f .

Assumption 2: The ratio $\mathbb{E}[\max\{0, u_P + \tau\} \mid u_H = u] / (\max\{0, u\} + \mathbb{E}[\max\{0, u_O\}])$ is continuous and strictly increasing in u .

Assumption 1 implies that taste for genre H is uninformative about taste for genre O . Assumption 2 captures positive dependence between taste for genre H and relative expected engagement with genre P : as u_H increases, expected engagement with genre P rises relative to genres H and O combined. This key assumption allows the platform to infer greater expected engagement with genre P , relative to the alternatives, from stronger interest in genre H .

Figure 1: Model Timing



Actions and Timing. Figure 1 summarizes the game between the platform and the user.

- (i) At $t = 0$, nature draws the user's session length N and type vector θ .⁷
- (ii) At $t = 1$, the user begins her session. The platform presents the first video, exogenously assigned to genre $r_1 = H$, and the user's short-run self chooses consumption $x_1 \geq 0$.
- (iii) At $t = 2$, the platform observes x_1 and ranks the remaining videos, determining the feed $(r_2, r_3, \dots)$.
- (iv) At $t = 3$, the user's short-run self chooses consumption $x_n \geq 0$ in each remaining slot she reaches, $n = 2, \dots, N$. The session ends after slot N , with $x_{n>N} = 0$.

⁷For models with endogenous session lengths, see Hébert and Zhong (2026); Koh and Sanguanmoo (2024); Saedi et al. (2024), which study engagement-maximizing platforms that differ from ours in two ways. First, users make an irreversible stopping decision rather than choose consumption in each slot. Second, the platform supplies information about a payoff-relevant state, which the user values for its use in decisions or for suspense (Ely et al., 2015), whereas here videos are consumption goods with direct entertainment and temptation value.

The initial video of genre $r_1 = H$ provides a simplified representation of the seeded watch histories in our field experiment. Although x_1 is also exogenous in the experiment, we model it as endogenous to capture the information conveyed by the histories. We assume that the platform does not revise its ranking after $t = 2$ to isolate its response to our seeding.

3.2 Solution Concept

We solve the game using perfect Bayesian equilibrium. Denote the short-run self's strategy by $\sigma_k(c; \theta)$, which specifies consumption of the c -th genre- k video for a user of type θ . Let $\mu(\cdot \mid x_1)$ be the platform's posterior belief about the user's type and $\rho(x_1)$ be its ranking strategy, or *algorithm*.

A pure-strategy perfect Bayesian equilibrium consists of strategies (σ^*, ρ^*) and beliefs μ^* such that the user's strategy σ^* is myopically optimal, that is,

$$\sigma_k^*(c; \theta) \in \arg \max_{x \geq 0} \left\{ u_k x + \mathbb{1}\{k = P\} \tau x - \frac{\gamma}{2} c x^2 \right\}$$

for every θ , k , and $c \geq 1$; the platform's strategy ρ^* is optimal given its beliefs μ^* , that is,

$$\rho^*(x_1) \in \arg \max_{(r_2, r_3, \dots) \in \{H, P, O\}^{\mathbb{N}}} \sum_{n=2}^{\infty} \Pr(N \geq n) \mathbb{E}_{\mu^*(\cdot \mid x_1)} [\sigma_{r_n}^*(c_{r_n}(n); \theta)];$$

and μ^* is derived from the prior f and σ^* using Bayes' rule whenever possible. If the next available videos from two genres yield the same expected engagement, the platform assigns the genre that comes first in the order H, P, O to the lower slot number.

Proposition 1. *A pure-strategy perfect Bayesian equilibrium exists.*

The equilibrium can be constructed via backward induction. Note that at both $t = 1$ and $t = 3$, given c , k , and θ , taking the first-order condition of the myopic short-run self's problem yields $\sigma_k^*(c; \theta) = \max \left\{ 0, \frac{u_k + \mathbb{1}\{k=P\}\tau}{\gamma c} \right\}$, which is the unique best response since the short-run payoff is strictly concave in x . At $t = 2$, the platform updates its beliefs using Bayes' rule given σ_k^* . Given these beliefs, it maximizes expected consumption by ranking the remaining videos in decreasing order of expected engagement, breaking ties according to the order H, P, O . This ordering is feasible because expected engagement decreases with exposure within each genre, and optimal because the probability that the user reaches a slot decreases with the slot number.

3.3 Parasitic Content in Equilibrium

To define parasitic content, we introduce two benchmarks: one for the platform's feed and another for the user's long-run payoff. Let $r^\varnothing$ be the platform's feed under the prior f , and denote the resulting number of genre- k videos in the first n slots by $c_k^\varnothing(n)$. This counterfactual is as if the platform does not observe x_1 and cannot personalize the feed at $t = 2$.

The second benchmark is used to evaluate the effect of genre P on the long-run self. For any feed r , let $V^{\text{open}}(r)$ be the long-run self's payoff under the user's equilibrium consumption and $V^{\text{block}}(r)$ be her payoff when genre P is blocked, so that $x_n = 0$ for n such that $r_n = P$.

Definition 1. *For a given user with consumption x_1 of genre H , genre P is parasitic to genre H if:*

1. **(Host-Parasite Criterion)** *Relative to $r^\varnothing$, the feed r contains more genre- P videos, that is, $c_P(N) > c_P^\varnothing(N)$; and*
2. **(Well-Being Criterion)** *The user's long-run payoff would be higher if she could block genre P , that is, $V^{\text{block}}(r) > V^{\text{open}}(r)$.*

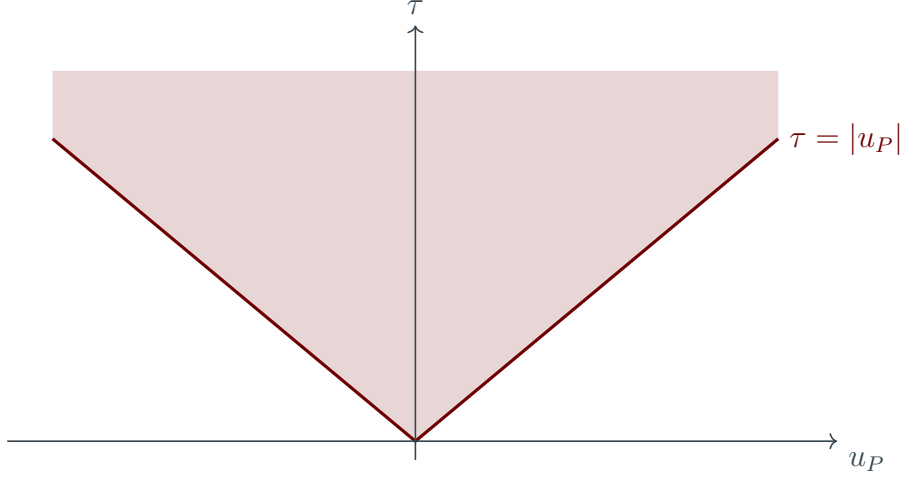
The first criterion requires genre P to expand in response to the algorithm's knowledge of genre- H consumption. The second criterion is necessary because some users benefit from consuming genre- P content, so that $V^{\text{block}}(r) < V^{\text{open}}(r)$. Both criteria hold in equilibrium under the conditions described below.

Proposition 2. *For a given user, genre P is parasitic to genre H for all sufficiently large N if and only if:*

1. *(High taste for genre H) u_H is sufficiently high; and*
2. *(Strong temptation for genre P) $\tau > |u_P|$.*

All proofs for this section are provided in Appendix C. Intuitively, high taste for genre H establishes the **host-parasite criterion**. Consider a user with taste u_H for genre H . At $t = 1$, she chooses her consumption x_1 (for a video from genre H) given her taste u_H . At $t = 2$, the platform updates its beliefs using her consumption and the distribution of user types. Under **Assumption 1**, expected engagement with genre O is unchanged, while, under **Assumption 2**, expected engagement with genre P relative to expected engagement for the other genres is increasing in her consumption of genre H . Thus, there is a cutoff rule: above a cutoff in x_1 and thus in u_H , the algorithm's updating raises expected engagement with genre P relative to the alternatives enough to increase its supply in sufficiently long sessions. If the user's consumption x_1 (or equivalently u_H) is below the cutoff, then the amount of genre P

Figure 2: Parameter Region for Parasitic Content



NOTE.—This figure fixes host taste u_H sufficiently high so the host-parasite criterion holds for sufficiently large N . The shaded region satisfies $\tau > |u_P|$, so blocking genre P increases the user's long-run well-being. Genre P is therefore parasitic to genre H in the shaded region.

content decreases for all sufficiently long feeds. Note that the feed must be long enough for the additional genre- P videos to be served; for example, if $N = 1$, the user sees only the initial genre- H video.

Strong temptation ensures that, at $t = 3$, the user consumes genre- P content to her long-run self's detriment, satisfying the **well-being criterion**: $V^{block}(r) > V^{open}(r)$. Figure 2 illustrates this region. If her taste for genre P is negative, her long-run self prefers zero consumption. When temptation exceeds her distaste, however, she consumes a positive amount. If her taste for genre P is positive, her long-run self prefers some consumption, and temptation induces her to consume more than is optimal. When temptation exceeds her taste, the costs of this excessive consumption outweigh its benefits, leaving her worse off than if she could block it. Conversely, if $\tau \leq |u_P|$, the user either consumes no genre- P content or obtains a weakly positive long-run payoff from consuming it.

3.4 Empirical Tests and Implications

Consider sessions sufficiently long for Proposition 2 to apply. To determine whether genre P is parasitic to genre H for some users, we experimentally test both criteria in Definition 1. In our setting, genre H corresponds to body positivity, which, as discussed in Section 2 and further supported by evidence from our mechanism experiment, is widely desired among young women and therefore plausibly generates high genre- H consumption for many users.

Thus, in the field experiment, we directly induce high genre- H engagement to test the **host-parasite criterion**. We then randomize exposure to genre P in the mechanism experiment to test the **well-being criterion**. Together, these tests provide evidence for the presence and operation of parasitic content.

Beyond the two defining criteria, our model yields three testable implications, for which proofs are given in Appendix C. The first is a behavioral consequence of temptation and provides evidence on mechanisms.

Implication 1 (Consumption). *If genre P is parasitic to genre H for a user sophisticated about her temptation, she prefers to block genre P but consumes a positive amount when it is served.*

The remaining implications concern broader consequences of algorithmic personalization for user well-being and platform revenue, given the presence of parasitic content.

Implication 2 (Average Effect of Personalization on Well-Being). *Let $p > 0$ be the share of users for whom genre P is parasitic to genre H , and let ℓ be their average long-run well-being loss from the additional genre- P videos induced by algorithmic personalization. If $p\ell$ is sufficiently large, algorithmic personalization lowers average long-run well-being:*

$$\mathbb{E}[V^{\text{open}}(r) - V^{\text{open}}(r^{\emptyset})] < 0.$$

Implication 3 (Revenue). *Choose x_1 such that $c_P(N) > c_P^{\emptyset}(N)$. For $c \in \{c_P^{\emptyset}(N) + 1, \dots, c_P(N)\}$, the c -th genre- P video generates weakly higher expected engagement and thus revenue than the next video from genres H or O that could be served instead:*

$$\mathbb{E}_{\mu^*(\cdot|x_1)}[\sigma_P^*(c; \theta)] \geq \mathbb{E}_{\mu^*(\cdot|x_1)}[\sigma_k^*(c_k(N) + 1; \theta)], \quad k = H, O.$$

4 Field Experiment: Design & Data

We conducted an eight-week online field experiment on TikTok to evaluate the **host-parasite criterion** and the implications for **average well-being** and **revenue**. Section 6 evaluates the **well-being criterion** and the model’s **implication for consumption**. Figure 3 summarizes the field experiment design.

4.1 Treatment Construction: Account Seeding

We exogenously varied the watch histories observed by TikTok using a novel account-seeding procedure. We created new TikTok accounts, assigned each a body-positive, fashion, or unseeded watch history, and randomly assigned participants to use them.

We followed a standardized protocol to create accounts so that they differed only in their seeded watch histories. Each account was created using a unique email address and a randomly assigned birth date in 2002, matching the average age of our participants. We created accounts on mobile devices using U.S. IP addresses from several states. Restricting accounts to U.S. IP addresses helped ensure that content served during the experiment was representative of what our primarily U.S.-based sample would see, while varying the state reduces the likelihood that our results reflect a recommendation policy specific to one state.

We then manipulated each account’s watch history to correspond to one of three treatment arms. In our main *body-positive treatment*, comprising 50% of the sample, participants received accounts seeded with body-positive watch histories. To generate these histories, we first followed approximately 30 prominent body-positive influencers, identified by searching for keywords such as “body positive” and “plus-sized,” and followed related hashtags (e.g., #bodypositive, #plussizefashion). Figure A1a provides examples. We then logged into each account from a server and auto-scrolled through TikTok’s “Following” page, which presents content exclusively from influencers followed by the account. This procedure generated a watch history comprising mostly body-positive content, with each video played in full before the automated browser advanced to the next.

In the *control*, comprising 20% of the sample, participants received newly created TikTok accounts with no seeded watch history. During the experiment, content served to the control group on TikTok’s algorithmically curated *For You Page* therefore captures what the algorithm supplies absent prior engagement data. Comparing the control and the body-positive treatment group identifies the effect of body-positive seeding on content supply.

In the *fashion treatment*, comprising the remaining 30% of the sample, participants received accounts seeded with fashion watch histories. The fashion treatment serves as an active control that helps distinguish three interpretations of the host-parasite criterion. If body-positive and fashion seeding generate similar increases in weight-loss content, the relevant host may be appearance-related content broadly. If body-positive seeding generates less weight-loss content than fashion seeding, body-positive engagement may instead protect users relative to other appearance-related engagement. If it generates more, the result is consistent with body-positive engagement revealing a particular susceptibility to weight-loss content. We constructed these accounts analogously to those used in the body-positive treatment. We followed approximately 30 prominent fashion influencers, identified by searching for “fashion,” and related hashtags, then generated watch histories by auto-scrolling on the “Following” page. Unlike the body-positive influencers, the fashion influencers were all thin (Figure A1b).

Quasi-Random Variation in Seeding Intensity.—In addition to varying the genre of seeded content, we leverage quasi-random variation in the amount of content seeded across accounts used in the body-positive and fashion treatments. This variation arose unintentionally from technological constraints: because seeding watch histories is memory-intensive, haphazard browser crashes interrupted the process. As a result, some accounts were seeded with only a few hundred videos, while others were seeded with several thousand. This variation allows us to examine how the intensity of the body-positive or fashion seeding affects weight-loss content supply, addressing the concern that the average treatment effects merely reflect heavily seeded accounts crossing an unusually high threshold. We show in Table A1 that within weekly account-assignment waves (since crashes may have occurred more frequently when the server was busier), baseline participant characteristics do not jointly predict seeding intensity.

4.2 Experimental Design

After creating and seeding the TikTok accounts, we randomly assigned each participant to use one account. Randomization ensures that seeded watch histories are independent of participant characteristics, allowing us to attribute subsequent differences in content supply and participant outcomes to the experimental variation in seeding.

Recruitment, Screening, and Baseline Survey.—Due to constraints in account creation and seeding, participants were recruited on a rolling basis in approximately weekly waves between November 2024 and January 2025, with each participant followed for eight weeks over the experiment. We recruited participants primarily through Prolific, supplemented by Instagram ads. Participants were paid for completing surveys throughout the experiment and could earn additional bonuses described below. To minimize selection on study topic, neither the Prolific recruitment post nor the Instagram ad mentioned body image (Figure A2). A total of 2,861 individuals clicked on the Prolific post. On Instagram, 26,640 individuals viewed our ad, and 2,777 clicked it.

Clicking on the recruitment post took individuals to a screening survey containing a consent form and questions about demographics; height and weight, which were used to calculate BMI; social media use; and study interest. In total, 2,415 individuals recruited through Prolific and 1,642 through Instagram completed the screening survey. Individuals who consented were eligible for the full study if they (i) were women ages 18 to 26, inclusive; (ii) resided in the U.S., UK, or Canada; (iii) used TikTok for at least 15 minutes per day, verified using screentime screenshots; (iv) expressed interest in participating in a multi-week study; and (v) responded to a follow-up message confirming that they could be recontacted.

We restrict the sample to young women because they are particularly likely to encounter body-positive and weight-loss content on social media, and to residents of the U.S., UK, and Canada because prior work finds a particularly strong association between social media use and worse mental health for girls relative to boys in these countries, with appearance-based social comparison proposed as a potential mechanism (Haidt, 2024). The third condition reduces the likelihood that participant consumption patterns are driven by unfamiliarity with TikTok, and the fourth and fifth select for participants likely to return for a multi-week study. We also excluded low-quality responses, including participants whose self-reported screentime was inconsistent with their screenshots or who failed attention checks.

Two weeks after screening, we invited 1,002 eligible participants—84% recruited through Prolific and 16% through Instagram—to complete a baseline survey measuring the pre-treatment outcomes described in Section 4.4. We implemented a two-week lag between recruitment and baseline surveys to further reduce attrition by selecting for participants likely to return for surveys spaced weeks apart (Allcott et al., 2022). In the baseline survey, we additionally elicited participants’ demand for several content genres, including body positivity, for heterogeneity analysis relevant to the model’s *revenue implication*.⁸

Midline 1 Survey: Randomization and Initial Scrolling Period.—Two weeks after the baseline survey, participants were invited to take the midline 1 survey. In the survey, participants were randomized into one of the three treatment groups: 50% to the body-positive treatment, 30% to the fashion treatment, and 20% to the control. Each participant received the username and password for her assigned account and logged in on her own mobile device. Participants retained access to their personal TikTok accounts throughout the experiment. Although TikTok can recognize multiple accounts used on the same device, it is unclear whether activity on one account influences recommendations on another. Regardless, randomization prevents such spillovers from biasing treatment effects. In total, 444 participants were assigned an account, of whom 424 successfully logged in.

Immediately following randomization, participants were instructed to scroll on their assigned account’s *For You Page* for 30 minutes in what we refer to as the initial scrolling period. Participants could scroll freely, search for their own interests, follow new influencers, and otherwise interact with the account as they liked, allowing us to test how participants engage with different types of content. We treat the initial scrolling period as the cleanest measure of the response to seeded histories because participants had had the least opportunity to

⁸Participants were asked the following incentivized question: “If this choice is randomly selected by the computer, we will offer you a \$3 bonus for watching short-form videos for 15 minutes. The videos will be selected based on what percent of your feed you want to devote to different topics.” Options included body positivity, fashion, humor, news and current events, beauty, animals, food and cooking, and fitness. Participants were told this would occur after all participants had completed the survey.

reshape their recommendations through their own behavior. As qualitative evidence, when later invited to compare their assigned and personal TikTok feeds in an open-ended response, one-fifth of participants volunteered that the assigned feed had become more aligned with their interests over the two-week period (see below). Although participants’ behavior during the initial scrolling period and throughout the experiment could influence subsequent recommendations, random assignment preserves the causal interpretation of differences across treatment groups. We verified compliance using screentime screenshots and TikTok watch histories, which are described in Section 4.3.

After the initial scrolling period, participants returned to the midline 1 survey. Outcomes are described in Section 4.4.

Two-Week Scrolling Period and Midline 2 Survey.—After completing the midline 1 survey, participants were incentivized to continue using their assigned account for two weeks. This two-week scrolling period allows us to examine whether treatment effects on content supply and participant outcomes persist and how the algorithm responds to participants’ engagement over time. We offered an \$8 bonus to participants who used their assigned account for at least five minutes on ten of the next fourteen days and completed the midline 2 survey. The bonus was designed to encourage regular use without requiring excessive engagement, to remain modest enough to avoid coercion, and to allow participants flexibility in when they used the account. Participants could otherwise interact with the account as they liked, and we sent daily reminders to encourage compliance.

After the two-week scrolling period, participants completed the midline 2 survey. The \$8 bonus was contingent on survey completion to reduce attrition. We then removed all incentives to use the assigned account.

Endline Survey and Personal Data Upload.—Two weeks after incentives for using the assigned accounts were removed, participants completed an endline survey measuring outcomes to assess the persistence of treatment effects. Participants could also upload their personal TikTok data for a \$5 bonus; 262 did so. In total, participants were followed for approximately eight weeks from recruitment through endline, with 15% attrition between successfully logging into the assigned account and completing the endline survey.

Ethics.—The treatment was not designed to expose participants to unusually extreme or niche content. Seeded content came from prominent TikTok creators, while subsequent content was selected by TikTok’s own recommendation algorithm from material already widely available on the platform. As we document in the mechanism experiment, most

young women regularly encounter both weight-loss and body-positive content in their own feeds. Additionally, participants’ incentivized use of the assigned accounts was limited. Over the full two-week treatment period, participants were incentivized to use the accounts for only 80 minutes, which is modest relative to their baseline average daily TikTok use of nearly 120 minutes. Moreover, to minimize coercion, participants were compensated at the recommended Prolific rate of \$12 per hour for surveys (Prolific, 2026).

Finally, because treatment effects are measured relative to the control group, a negative effect of the body-positive treatment on well-being does not necessarily imply that participants were worse off relative to their usual TikTok experience. In open-ended responses from the endline survey, 14% of the control group explicitly reported that receiving less body-related content made them think less about their appearance, suggesting that part of the estimated effect may reflect an improvement in the control group.

External Validity.—Our seeding procedure resembles computer-science audits that use automated, or “sock-puppet,” accounts to study recommendation algorithms (Sandvig et al., 2014). Such audits are criticized because sock puppets behave unlike real users, whose interests are varied rather than concentrated in a single genre. We address this in three ways. First, the seeded interest is not unusual in this population: 62% of participants reported at baseline that they see body-positive content in their feeds sometimes or often. Second, the seeded accounts are used by human participants, which lets us test whether the algorithm’s response persists once varied human behavior enters. Third, the quasi-random variation in seeding intensity identifies how supply responds as engagement accumulates, which is relevant for real users. Seeding nonetheless remains artificial, and we interpret the experiment as a proof of concept rather than an estimate of the algorithm’s response under ordinary use.

4.3 Data

We combine two datasets: survey data and TikTok watch histories.

Survey Data.—We collected pre-treatment measures at recruitment and baseline and post-treatment measures at midline 1, midline 2, and endline. Section 4.4 details outcomes.

TikTok Watch Histories.—After the experiment, we downloaded the watch histories for 417 of the 424 accounts used in the study. Each watch history records a link to every video viewed on the account and the time, measured to the second, at which the video was served. In total, the downloaded watch histories cover 1,383,317 videos.

Using the video links, we scraped metadata for 99.8% of unique items and retrieved MP4 files for 89.5%. When we were unable to scrape metadata, the item was either private or had been removed from TikTok. Missing MP4 files may also reflect the fact that some items were not videos at all, but collections of still images. The metadata for each video include the creator’s name, auto-generated transcript, hashtags, video description, text overlaid on the content, broad content genre, video length, and aggregate engagement statistics, such as the number of likes, shares, and comments.

We classified videos into genres, focusing on weight loss, body positivity, and fashion. Because these categories are not separately identified in TikTok’s broad content genres, we developed classifications for each. To classify weight-loss content, we provided GPT-4o with all relevant text data (i.e., auto-generated transcripts, hashtags, video descriptions, and text overlaid on the videos) using a fixed prompt and 24 few-shot examples; Appendix A1 provides details. We validated these classifications with two research assistants, each of whom independently categorized 500 videos using the relevant text data as promoting weight loss or not. The research assistants agreed with each other on 94% of videos, and their average agreement with GPT-4o was 92%.

For body-positive and fashion content, we classified a video according to whether it was posted by a body-positive or fashion influencer, respectively, followed during seeding. We use this as our main classification because it maps directly onto how these genres were defined in the seeding procedure. Because body-positive and fashion content are classified by creator while weight-loss content is classified by the content of the video itself, these genres are not mutually exclusive. For other genres, we use the broad categories provided in the metadata.

The metadata also identify whether each video was an advertisement (hereafter, ad) or organic content, that is, content not associated with paid promotion. Ads include content created or posted by firms, as well as content created by influencers who sell a product and receive a portion of the profits. Overall, 8.1% of videos served to participants were ads.

We measure consumption using the timestamps in the watch histories, inferring watch time as the elapsed time between consecutive videos. If the gap between consecutive videos exceeded ten minutes, we assumed that the participant ended her scrolling session and set consumption for the preceding video to missing. These timestamps allow us to measure how long each participant watched each video and total daily use of the assigned account.

4.4 Outcomes

Outcomes are organized into three categories.

Content Supply.—We measure the content supplied by the algorithm using TikTok watch histories. Our primary content-supply outcomes are the numbers of body-positive, fashion, and weight-loss videos served. Appendix analyses use each genre’s share of all videos viewed.

Content Consumption.—We measure participants’ consumption of organic content and ads using watch time inferred from the TikTok watch histories. For organic content, our primary measure is *watch time* for each video. For advertisements, we construct a binary variable, *billable*, equal to one if an ad was watched for at least six seconds. We use this threshold as a proxy for a qualifying ad view, since under some pricing schemes, TikTok charges the advertiser only if the ad is watched for at least six seconds (TikTok, 2025). We also aggregate watch time to the daily level to assess compliance with assigned-account use.

Body Image.—Our primary participant outcome is body image. We measure body image using nine pre-registered outcomes and combine the measures available in each survey into a pre-registered *body image index*, weighting by the inverse of the covariance between variables in the control group and signing each component so that higher values indicate better body image (Anderson, 2008). The index combines measures of body attitudes, self-objectification and appearance consciousness, appearance-related concerns, and body-image-related behaviors.

We measure attitudes using two validated scales. *Body appreciation* is measured using the 10-item Body Appreciation Scale (BAS) (Tylka and Wood-Barcalow, 2015), and *body dissatisfaction* is measured using the 9-item Body Dissatisfaction Scale (Clausen et al., 2011).

We measure self-objectification using the Self-Objectification Beliefs and Behaviors Scale (Lindner and Tantleff-Dunn, 2017). We analyze its two subscales separately: *self objectification*, which captures the extent to which individuals equate their bodies with their self-worth, and *observer objectification*, which captures the frequency with which individuals view their bodies from an external perspective. We also elicit participants’ incentivized willingness to accept compensation to make a social media post about body positivity (*WTA to post*), which captures reluctance to publicly associate oneself with body-positive content.

To measure the salience of appearance-related concerns, we ask participants at the beginning of each survey, before any body-image-related questions, to list their three most pressing concerns. A research assistant categorized each concern as appearance-related, health-related, or neither. We construct two binary measures, *concerned with appearance* and *concerned with health*, equal to one if any of a participant’s listed concerns falls into the corresponding category. We also measure *perceived weight* on a five-point scale ranging from very underweight to very overweight. When coding *perceived weight*, we take higher values

to indicate greater body dissatisfaction since previous research has shown that young women often idealize body types that correspond to being underweight (MacNeill and Best, 2015).

Finally, we measure body-image-related behaviors using five questions about actions participants took during the previous two weeks, such as avoiding activities requiring certain clothing or refraining from eating around others, *because* they were afraid of what others would think or say about their body. We combine these measures into a single *index of poor body image behaviors* (Anderson, 2008).

4.5 Sample Characteristics and Balance Checks

Table 1 reports descriptive statistics by treatment group, and Table A2 compares our sample to U.S. women ages 18–26, U.S. TikTok users, and U.S. adults, where comparable statistics are available. All participants were women, with an average age of 22.5 years; 51% were from the U.S., 34% from the UK, and 15% from Canada. While the racial composition of our sample differed from that of the U.S. population, it was similar to that of U.S. TikTok users. Participants used TikTok regularly, averaging 119 minutes per day on the platform, close to the 114-minute daily average reported for U.S. teenage girls (Rothwell, 2023). Given our focus on body image, we also asked participants for their height and weight to compute BMI, which averaged 25.0. This value is just above the healthy range (18.5–24.9) and below the 27.9 average among U.S. women ages 18–26.

The top two panels of Table 1 show that baseline participant characteristics were balanced across treatment groups. Three variables differ significantly at the 5% level, which is expected given the number of tests; the baseline body image index, our primary outcome, is not among them. Pairwise joint F -tests find no evidence of systematic differences in baseline observable characteristics across treatment groups. Nonetheless, all analyses control for the baseline value of each outcome when available, along with other baseline characteristics.

Attrition from login to endline was modest at 15%, and compliance was high: among participants who logged in, all completed the initial 30-minute scrolling period, and nearly 80% used their accounts enough over the two-week scrolling period to qualify for the \$8 bonus. Neither attrition nor compliance differed significantly across treatment groups, as shown in the third panel of Table 1.

Figure A3 shows account usage during the initial and two-week scrolling periods, and Figure A4 provides a day-by-day breakdown. Participants spent 41 minutes and scrolled through 225 videos in the initial period on average, and averaged 30 minutes and 100 videos per day over the two weeks, conditional on account use. Time spent was similar across treatment groups, but the number of videos was not: relative to the control, participants in the body-positive treatment scrolled through 38 fewer videos in the initial period ($p = 0.042$),

while those in the fashion treatment scrolled through 29 more per day over the two weeks ($p = 0.080$). For interpretability, the main analyses use raw video counts to quantify content supply. Because differences in scrolling rate could mechanically affect counts, we also report estimates using shares and find similar results.

5 Field Experiment: Host-Parasite and Average Well-Being

This section tests the **host-parasite criterion** and the model’s **average well-being implication**, with robustness checks presented at the end. Section 7 examines the **revenue implication**.

5.1 The Host-Parasite Criterion

We estimate the effect of seeding on the content received by participants, exploiting variation on two margins: the genre seeded (the extensive margin) and, within genres, the number of videos seeded (the intensive margin). The first margin comes from randomized assignment to treatment arms, and the second from quasi-random variation in seeding intensity.

Extensive Margin.—We estimate the effect of treatment assignment on the amount of body-positive, fashion, and weight-loss content served. For each genre, we estimate the following intent-to-treat regression:

$$Y_i = \beta_0 + \beta_1 T_i^B + \beta_2 T_i^F + X_i' \gamma + \varepsilon_i, \quad (1)$$

where Y_i is the number of videos within the genre served to participant i ; T_i^B and T_i^F are indicators for assignment to the body-positive and fashion treatments, respectively; and X_i is a vector of the baseline participant characteristics reported in Table 1. Standard errors are heteroskedasticity-robust. Mapping to the model, the **host-parasite criterion** requires $\beta_1 > 0$ when Y_i is the number of weight-loss videos served. We restrict our main analysis to content served during the initial scrolling period to reduce noise from the algorithm’s adaptation to participants’ behavior.

Figure 4 plots the average number of body-positive, fashion, and weight-loss videos served to participants in each treatment group during the initial scrolling period, with error bars indicating 95% confidence intervals; the corresponding coefficients are reported in Table A3. As a manipulation check, participants in the body-positive treatment received 15.8 more body-positive videos relative to the control group ($p < 0.001$; control mean = 0). Moreover, the body-positive treatment group *also* received 3.5 more weight-loss videos than the control ($p < 0.001$; control mean = 0.8), establishing the host-parasite criterion.

To put the magnitude of these effects in context, we use the fashion treatment as a

benchmark. Assignment to the fashion treatment increased the amount of weight-loss content received by 0.7 videos relative to the control ($p = 0.048$) but resulted in 2.8 *fewer* weight-loss videos than assignment to the body-positive treatment ($p < 0.001$). This latter difference does not reflect a weaker seeding manipulation in the fashion treatment than in the body-positive treatment, as participants assigned to the fashion treatment received 15.1 more fashion videos than the control ($p < 0.001$), which is statistically indistinguishable from the analogous 15.8-video increase in body-positive content for the body-positive treatment group ($p = 0.626$). These results are inconsistent with the interpretation that body-positive engagement protects users from weight-loss content relative to engagement with fashion content. Instead, body-positive engagement induces a larger increase in the algorithm’s supply of weight-loss content.

While the initial scrolling period provides the cleanest test of the algorithm’s response to seeding, the economic significance of this response also depends on its persistence. We therefore reestimate Equation (1) at the participant-day level using the number of weight-loss videos served on each day that a participant used the account over the two-week scrolling period. We pool observations separately over days 1–7 and days 8–14, include day fixed effects, and cluster standard errors at the participant level. Columns 1 and 2 of Table A4 report the pooled estimates; Figure A5 displays the corresponding covariate-adjusted estimates by study day. Participants in the body-positive treatment received 1.8 more weight-loss videos per day than the control over the first week ($p < 0.001$; control mean = 0.58) and 1 more per day over the second ($p < 0.001$; control mean = 0.71). The fashion treatment group also received persistent, although smaller, increases of 0.4 videos ($p = 0.039$) and 0.7 videos ($p = 0.051$) over the first and second weeks, respectively. The algorithm therefore generated sustained exposure to weight-loss content, with some attenuation likely due to the algorithm adjusting to the participants’ own engagement patterns.

Intensive Margin.—To complement the above result, we test whether weight-loss supply was affected by seeding intensity by exploiting quasi-random variation in seeding intensity within the body-positive and fashion treatments, as described in Section 4.1. Restricting the sample to participants assigned to these two treatments, we estimate

$$Y_i = \alpha + \delta T_i^F + \beta_B S_i + \theta (T_i^F \times S_i) + X_i' \gamma + \lambda_{w(i)} + \varepsilon_i, \quad (2)$$

where Y_i is the number of weight-loss videos served to participant i during the initial scrolling period; T_i^F indicates assignment to the fashion treatment, with the body-positive treatment as the omitted category; and S_i is the demeaned log of the count of seeded videos for

participant i 's assigned account, with the mean calculated over the pooled body-positive and fashion treatments. The vector X_i contains the baseline participant characteristics reported in Table 1, and $\lambda_{w(i)}$ denotes fixed effects for the week of account assignment. We restrict the analysis to content served during the initial scrolling period to increase precision. Heteroskedasticity-robust standard errors are reported.

Figure 5 plots mean covariate-adjusted weight-loss supply by binned seeding intensity, along with the fitted relationships from Equation (2). The slopes of the body-positive and fashion treatment lines are β_B and $\beta_B + \theta$, respectively; the estimates are also reported in Table A5. Greater body-positive seeding intensity *increased* the algorithm's supply of weight-loss content: doubling the number of body-positive videos seeded caused the algorithm to serve an additional 0.44 weight-loss videos during the initial scrolling period ($p = 0.043$). The binned means also exhibit an upward trend across the observed range of body-positive seeding intensity, suggesting a gradient in weight-loss supply rather than a discrete jump above a particular level of seeding. By contrast, the estimate on the intensity of fashion seeding implies that doubling the number of seeded fashion videos reduced the number of weight-loss videos served by 0.56, although this estimate is imprecise ($p = 0.165$). The difference between the body-positive and fashion slopes is statistically significant ($p = 0.022$). Thus, the intensive-margin results complement the extensive-margin findings: weight-loss supply increases not only when accounts are seeded with body-positive content but also as the amount of body-positive content seeded increases.

Finding 1 (Host-Parasite Criterion): *Body-positive engagement increases the algorithm's supply of weight-loss content. The increase in weight-loss content is larger than that following fashion engagement.*

5.2 Average Well-Being

We test the model's **average well-being implication**—that algorithmic personalization can reduce average user well-being—using body-image measures collected at midline 1, midline 2, and endline. We estimate Equation (1), where Y_i is participant i 's body image index in the relevant survey and X_i contains baseline participant characteristics, including baseline outcome measures, reported in Table 1. Here, β_1 estimates the effect of assignment to the body-positive treatment on well-being at that point in time, corresponding to $V^{open}(r) - V^{open}(r^{\emptyset})$ in the model. The implication holds if $\beta_1 < 0$.

Figure 6 plots treatment effects on the body image index; the corresponding estimates are reported in Table A6. Immediately after the initial scrolling period, participants assigned to the body-positive treatment reported *worse* body image relative to the control (-0.23

standard deviations, $p = 0.008$). The effect remained negative and similar in magnitude at midline 2 (-0.27 SD, $p = 0.010$) and endline (-0.21 SD, $p = 0.041$), providing no evidence that the initial short-run decline was offset by subsequent improvements over the period we observe. The corresponding estimates for the fashion treatment were smaller and statistically insignificant at midline 1 (-0.10 SD, $p = 0.301$), midline 2 (-0.10 SD, $p = 0.376$), and endline (-0.10 SD, $p = 0.417$). Although the differences between the body-positive and fashion treatment effects are imprecise (e.g., $p = 0.141$ at midline 1), their direction is consistent with **Finding 1** showing that body-positive seeding increased weight-loss supply substantially more than fashion seeding. Together, these results support the contention that in our context, algorithmic personalization based on body-positive engagement reduces average user well-being.

Our estimates cannot rule out offsetting benefits over longer horizons or for health outcomes not measured in our experiment. However, a large literature examines the relationship between body image and mental and physical health; see [Nolen and Panisch \(2022\)](#) and [Wilfred et al. \(2026\)](#) for reviews. Worse body image predicts subsequent depression and anxiety, even after accounting for BMI and other baseline characteristics ([Stice and Bearman, 2001](#); [Vannucci and Ohannessian, 2018](#)). Moreover, longitudinal studies do not find evidence that lower body satisfaction motivates healthy weight-management behavior; instead, lower body satisfaction predicts behaviors associated with weight gain and poorer health, including binge eating, smoking, and less physical activity ([Neumark-Sztainer et al., 2006](#)).

To examine which outcomes contribute to the above effect, we disaggregate the body image index in Appendix Figures [A6](#) and [A7](#). At midline 1, the negative estimates for the body-positive treatment are concentrated in concern with appearance, concern with health, and the willingness-to-accept measure for posting about body positivity. By midline 2, the estimates are negative across all components. The largest declines are in concern with appearance (-0.28 standard deviations), willingness to accept to post about body positivity (-0.23 standard deviations), and the body-image behaviors index (-0.21 standard deviations). None of the component estimates for the fashion treatment are significant.

Finding 2 (Average Well-Being): *Participants assigned to the body-positive treatment reported worse body image than those in the control group at midline 1, midline 2, and endline. Effects of the fashion treatment were also negative but smaller in magnitude and statistically insignificant.*

5.3 Robustness

Robustness of Finding 1.—We assess four alternative explanations for the increase in weight-loss content served to the body-positive treatment group.

1. *Influencers followed during body-positive seeding.* Weight-loss videos posted by body-positive influencers followed during seeding do not drive the result. As shown in column 3 of Table A4, the body-positive treatment group received 3.1 more weight-loss videos than the control even after excluding these videos ($p < 0.001$). Similarly, as shown in column 2 of Table A5, when these videos are excluded from the intensive-margin analysis, doubling body-positive seeding caused the algorithm to serve 0.4 additional weight-loss videos ($p = 0.050$).

2. *Weight-loss content during seeding.* We avoided influencers who promoted weight loss, but accounts could nevertheless encounter weight-loss content during seeding, which may have led TikTok to infer an interest in weight-loss content. Although weight-loss videos comprised less than 5% of the seeded content for every account, body-positive accounts encountered more of this content during seeding than fashion accounts (Figure A8). We test whether this imbalance explains our results by controlling directly for incidental weight-loss exposure. On the extensive margin, the estimated effect is 3.9 additional weight-loss videos when controlling for the share of seeded videos classified as weight-loss content and 3.8 additional videos when controlling for any weight-loss exposure (both $p < 0.001$; columns 4 and 5 of Table A4). On the intensive margin, the estimated effect per doubling of body-positive seeding is 1.3 additional weight-loss videos when controlling for the number of seeded videos classified as weight-loss content and 0.5 additional videos when controlling for any weight-loss exposure ($p < 0.001$ and $p = 0.091$, respectively; columns 3 and 4 of Table A5). All four point estimates are larger than their corresponding main estimates, *strengthening* the finding, although the last intensive-margin estimate is less precise.

3. *Representation of content.* TikTok may not distinguish between body-positive and weight-loss content. If this is the case, engagement with one genre should predict the subsequent supply of the other. To assess this explanation, we estimate

$$Y_{it} = \alpha + \sum_{g \in \{W, B\}} \beta_g E_{i,t-1}^g + \sum_{g \in \{W, B\}} \gamma_g Y_{i,t-1}^g + \varepsilon_{it}, \quad (3)$$

where Y_{it} is the number of weight-loss or body-positive videos supplied to participant i on day t , $E_{i,t-1}^g$ is watch time for genre g for participant i on day $t - 1$, and $Y_{i,t-1}^g$ is the number of genre- g videos supplied to participant i on day $t - 1$. Here, $g \in \{W, B\}$, with W denoting weight-loss content and B body-positive content. We estimate the equation separately for the two outcomes. Standard errors are clustered at the participant level.

Column 1 of Table 2 shows that an additional 100 seconds of weight-loss watch time predicts 0.69 more weight-loss videos on the next observed study day ($p = 0.001$), controlling for lagged weight-loss supply. Column 2 includes only lagged body-positive watch time and supply; here, body-positive watch time also positively predicts weight-loss supply, although weakly (0.25 videos per 100 seconds, $p = 0.088$). Once lagged watch time and supply for both genres are included in column 3, however, the coefficient on body-positive watch time shrinks to 0.08 and becomes statistically insignificant ($p = 0.409$), while the coefficient on weight-loss watch time remains virtually unchanged from column 1 (0.69, $p = 0.001$). Columns 4–6 show a similar genre-specific pattern for the supply of body-positive content. These results suggest that the algorithm can distinguish the two genres.

4. *Video counts versus shares.* The result is not an artifact of participants in the body-positive treatment encountering more videos overall: using the share of weight-loss videos yields similar extensive- and intensive-margin results (column 6 of Table A4 and column 5 of Table A5, respectively).

Robustness of Finding 2.—We assess several explanations for the decline in body image.

1. *Content consumption from personal accounts.* A natural concern is that participants offset their experience on the study account by adjusting behavior on their personal accounts, leaving their overall social-media diet unchanged. Three observations address this. First, the significant decline in body image rules out complete compensation: had participants fully neutralized the treatment elsewhere, we would observe no effect. Second, any compensating substitution, such as seeking body-positive content on personal accounts, would attenuate our estimate, so the measured effect is a lower bound on the decline in body image. Third, the midline 1 results show that body image declines immediately after the initial scrolling period, before such cross-account adjustment could occur.

2. *Total TikTok use.* One may wonder how much of the effect is driven by the effect of social media exposure rather than content genre. The midline 1 effect is measured immediately after all participants completed the same incentivized initial scrolling period, limiting the role of differential exposure time. Over the subsequent two weeks, however, assignment could change participants’ total TikTok use rather than only the composition of content they consumed. Column 3 of Table A6 adds TikTok use measured at midline 2 to the corresponding specification. The estimated decline remains similar and, if anything, becomes larger (-0.30 standard deviations, $p = 0.006$).

3. *Effect of body-positive vs. weight-loss content.* One possible explanation is that body-positive content, rather than the weight-loss content served alongside it, worsened body image. The mechanism experiment described in Section 6 separately randomizes exposure

to each genre and provides evidence that body-positive content alone cannot explain the decline, whereas weight-loss content contributed to it. Other genres amplified in response to body-positive seeding (Figure A9) may have also contributed. This possibility does not undermine the conclusion that weight-loss content is parasitic; in fact, some of these other genres may also be parasitic.

6 Mechanism Experiment: Well-Being and Consumption

Finding 2 is consistent with our model but admits alternative explanations. For example, body-positive content itself may worsen body image by making insecurities more salient. Our mechanism experiment disentangles the roles of body-positive and weight-loss content, thereby testing the **well-being criterion**. We also test our model more precisely by examining the **consumption implication**: sophisticated users who are sufficiently tempted by weight-loss content should prefer to block it but consume it if served.

6.1 Mechanism Experiment Design

We recruited 1,000 young women and other non-male participants through Prolific between December 2025 and January 2026. We used the field experiment’s demographic eligibility criteria but admitted Instagram as well as TikTok users to increase sample size. The samples are broadly similar, with some small but statistically significant differences: for instance, mechanism experiment participants are 0.6 years older on average, and 81% use TikTok (Table B1). Our main analysis uses the full sample to maximize statistical power. As a robustness check, we reweight the sample to match the field-experiment sample and report the results in the appendix.

Participants were randomly assigned to one of four video treatments; Figure B1 summarizes the design, and Table B2 reports balanced baseline characteristics across treatments. Each treatment consisted of a controlled 10-minute TikTok feed. Participants could freely skip or replay videos, and we recorded watch time for each video.

The experiment has two conceptual components. The first component evaluates the effects of body-positive and weight-loss content on body image using three treatment groups. The *control* received only neutral content, the *body-positive treatment* received 50% body-positive and 50% neutral content, and the *weight-loss treatment* received 50% weight-loss and 50% neutral content. Neutral videos came from the *For You Page* of a new TikTok account; body-positive and weight-loss videos came from content served in the field experiment. After viewing, participants completed the BAS, adapted to ask how they felt “right now, at this very moment” to increase sensitivity to changes caused by content exposure, and the Positive

and Negative Affect Schedule (Watson et al., 1988). We combined the BAS, positive mood, and negative mood (with its sign reversed) into a pre-registered index (Anderson, 2008).

The second component tests the **consumption implication** by comparing stated preferences with consumption behavior. Before randomization, participants separately chose whether to include or exclude body-positive, weight-loss, and cryptocurrency content from their feed and reported their willingness to pay (WTP) to implement each choice using a multiple price list in \$1 increments. Cryptocurrency provided a benchmark as a plausibly undesired but non-tempting genre for young women, a choice supported ex post by participants’ explanations for their choices (Appendix B1). Truthful reporting was incentivized using the Becker–DeGroot–Marschak mechanism (Becker et al., 1964), with 0.2% of participants selected to have their choices implemented. Before choosing, participants practiced with another genre and completed comprehension checks. Some participants were then randomized into the *mixed treatment*, containing equal shares of body-positive, weight-loss, cryptocurrency, and neutral content. Among those who chose to exclude both weight-loss and cryptocurrency content, we compare watch time across these genres to distinguish weight-loss consumption from the time needed to skip unwanted, non-tempting content.

6.2 Well-Being

We estimate the effect of body-positive and weight-loss content using the regression:

$$Y_i = \beta_0 + \beta_1 T_i^B + \beta_2 T_i^W + \gamma X_i + \epsilon_i, \quad (4)$$

where Y_i is the post-treatment body image and mood index for participant i , T_i^B (T_i^W) is an indicator for randomization into the body-positive (weight-loss) treatment, and X_i contains age, BMI, and indicators for race, country, gender, and LGBTQ identity. As specified in our pre-analysis plan, we exclude the mixed treatment; the version including all arms is in the appendix. Since we use body image to proxy for long-run well-being, the **well-being criterion** is satisfied if $\beta_2 < 0$.⁹

Figure 7 shows the results, which are also reported in Table B3. Column 1 shows that body-positive content alone *increased* the body image and mood index by 0.27 standard deviations ($p = 0.010$), whereas weight-loss content alone *reduced* it by 0.23 standard deviations ($p = 0.043$); the effects differ by 0.49 SD ($p < 0.001$). These results do not support body-positive content itself as the explanation for the decline in body image in the field.

⁹Comparing weight-loss content with neutral content parallels the field experiment, whose control group received content from unseeded accounts rather than no content. The model’s blocked benchmark instead leaves genre- P slots unfilled; the two comparisons coincide if replacing those slots with neutral content has no effect on well-being, as we assume here.

Instead, they support the **well-being criterion** and show that weight-loss content contributed to the field-experiment decline.

As shown in column 2 of Table B3, the decline was larger among participants who had preferred to exclude weight-loss content (0.32 standard deviations, $p = 0.049$). Column 3 shows that among those who had preferred to include it, the effect was small and statistically insignificant (-0.10 SD, $p = 0.554$). This pattern is consistent with sophisticated participants anticipating harm from weight-loss content and choosing to avoid it. Open-ended responses shown in Appendix B1 support this interpretation: among participants who excluded weight-loss content, 54% cited psychological harm, including past eating disorders, body-image concerns, and self-control problems.

Finding 3 (Well-Being): *Weight-loss content worsens the body image and mood index. Body-positive content improves it.*

6.3 Consumption

To test the **consumption implication**, we restrict the sample to participants assigned to the mixed treatment who chose to exclude both weight-loss and cryptocurrency content. Using weight-loss and cryptocurrency videos for these participants, we estimate

$$Y_{ij} = \alpha_i + \beta W_j + \varepsilon_{ij}, \quad (5)$$

where Y_{ij} is participant i 's watch time for video j , measured in seconds; W_j equals one for weight-loss videos and zero for cryptocurrency videos; and α_i denotes participant fixed effects. Thus, β measures the within-participant difference in per-video watch time between weight-loss and cryptocurrency content, in seconds; $\beta > 0$ indicates longer watch time for weight-loss content. Standard errors are clustered at the participant level.

Table 3 reports the results. Column 1 shows that among participants who chose to exclude both genres, participants watched each weight-loss video for 4.2 seconds (52%) *longer* than each cryptocurrency video ($p < 0.001$). Columns 2–4 assess the robustness of this result to additional sample restrictions. Column 2 restricts the sample to participants with strictly positive WTP to exclude weight-loss content (37% of those who chose to exclude it in the full sample), addressing the possibility that the result is driven by weak preferences to avoid it. Column 3 restricts to participants whose WTP to exclude weight-loss content was at least as high as their WTP to exclude cryptocurrency, addressing the possibility that stronger distaste for cryptocurrency drives the result. Finally, although weight-loss videos displayed explicit genre cues in their titles, such as “weight-loss transformation” and “diet

tips,” participants may still have confused weight-loss with body-positive content and thus watched weight-loss content for longer than intended. Column 4 addresses this concern by restricting to participants who also chose to exclude body-positive content, making such confusion less likely to drive the result. Estimates are similar and significant across all restrictions.

To examine whether the above results are true on average but driven by a small subset of participants, Figure B2 plots the distribution of within-participant differences in mean watch time between weight-loss and cryptocurrency content. Among participants in the main estimation sample, 81% watched weight-loss videos longer on average than cryptocurrency videos, despite having chosen to exclude both genres. Altogether, these results establish Finding 4 below.

Finding 4 (Consumption): *Among participants who preferred to exclude both weight-loss and a benchmark non-tempting genre, each weight-loss video was consumed for 52% longer than each benchmark video.*

Since the field experiment establishes the **host-parasite criterion** and the mechanism experiment establishes the **well-being criterion**, weight-loss content is parasitic to body-positive content for some young women. This group may be substantial: as shown in Table B2, over 80% of participants desired body-positive content, meaning that many may enter the first step of the parasitic pathway to weight-loss content by engaging with the host genre. Furthermore, nearly half preferred to avoid weight-loss content, suggesting that it may have entered their feeds parasitically. For brevity, we hereafter refer to weight-loss content served in the body-positive treatment as *parasitic content*, with the understanding that this designation applies to users for whom the well-being criterion holds.

7 Platform Incentives: Engagement and Revenue

Why does the platform supply parasitic content? Returning to the watch history data from our field experiment, we examine the **model’s revenue implication** and compare the engagement and billable ad views of parasitic content with alternative content the platform could have served.

We estimate the following regression separately for non-ads and ads:

$$Y_{ij} = \beta_0 + \beta_1 P_{ij} + \beta_2 B_{ij} + \beta_3 F_i + X_i' \gamma + Z_j' \delta + \varepsilon_{ij}, \quad (6)$$

where Y_{ij} is participant i ’s watch time, in seconds, for non-ad video j , or the *billable* indicator

for ad j , our proxy for advertising revenue; P_{ij} and B_{ij} indicate, respectively, whether video j is weight-loss or non-weight-loss content served in the body-positive treatment, F_i indicates assignment to the fashion treatment, and the control group is the omitted category; X_i is a vector of baseline participant controls; and Z_j contains video-level characteristics, including likes, comments, and other measures of popularity. Standard errors are clustered at the participant level. We test the revenue implication by comparing engagement with parasitic content to content in the control, non-weight-loss content in the body-positive treatment, and content in the fashion treatment, corresponding to β_1 , $\beta_1 - \beta_2$, and $\beta_1 - \beta_3$, respectively.

Table 4 reports the coefficients and corresponding differences. On a per-video basis, parasitic videos were consumed for 3.3 seconds (39%) longer than control videos ($p < 0.001$), 1.6 seconds (15.5%) longer than fashion-treatment videos ($p = 0.034$), and 0.3 seconds longer than other non-weight-loss videos in the body-positive treatment, although the last difference is not significant ($p = 0.679$). On a per-ad basis, parasitic ads were 13.9 percentage points more likely to generate a billable view than control ads, a difference equivalent to 73% of the control mean ($p = 0.024$; control mean = 19.1%). Relative to other ads in the body-positive and fashion treatments, parasitic ads were 12.4 ($p = 0.024$) and 10.7 percentage points ($p = 0.063$) more likely to generate a billable view, respectively.

To determine the economic impact of these differences, we perform a back-of-the-envelope calculation estimating the increase in annual ad revenue from parasitic weight-loss advertising. Approximately 63% of U.S. women ages 18–26, or 12.8 million, use TikTok (McClain and Eddy, 2026); based on our mechanism experiment, we assume that 81% of them desire body-positive content (Table B2), and, using estimates from the initial scrolling period, that these users encounter an average of 0.3 weight-loss ads per 30 minutes (Figure A11). We further assume daily TikTok use of 114 minutes (Rothwell, 2023), that weight-loss advertisers use TikTok’s pay-per-view pricing scheme rather than alternatives such as pay per click, that they pay TikTok’s suggested rate of \$0.22 per billable view, and that exposure rates, prices, and billable-view differences persist throughout the year. Under these assumptions, the estimated effects imply approximately \$124 million in additional annual revenue relative to control ads, \$111 million relative to other ads in the body-positive treatment, and \$96 million relative to fashion-treatment ads.

Finding 5 (Revenue): *Parasitic content receives greater engagement than content served in the control and fashion treatments. Parasitic ads are also more likely to generate billable views than ads served in the control and other ads served in the body-positive treatment, and marginally more likely relative to ads in the fashion-treatment.*

Appendix Table A7 reports two robustness checks. Panel A restricts the sample to the most engaging control and fashion-treatment videos and the last few non-weight-loss videos in the body-positive treatment, approximating the marginal content displaced by parasitic content in the model. Panel B restricts to participants with positive baseline demand for body-positive content, who may naturally generate the host engagement that triggers the delivery of parasitic content in our context. Estimates are similar, with some loss of precision due to a smaller sample size.

8 Conclusion

This paper introduces *parasitic content*: algorithmically personalized temptation goods that are served in response to engagement with the host genre. We propose a simple model to define and characterize parasitic content and its implications for user behavior, well-being, and firm revenue. We document the presence and operation of parasitic content using two complementary experiments. In the first, we conducted a novel eight-week field experiment on TikTok in which we seeded TikTok accounts with exogenous watch histories and randomly assigned participants to use them. Using TikTok watch histories spanning 1.38 million videos generated during the experiment, we document that, within the first 30 minutes of participant use, participants assigned accounts seeded with body-positive histories received 15.8 more body-positive and 3.5 more weight-loss videos than the control. Immediately after this initial scrolling period, their body image declined by 0.23 standard deviations. Over the following two weeks, the algorithm continued to serve these participants weight-loss content, and the decline in body image grew to 0.27 SD and remained significant two weeks after treatment ended. In a mechanism experiment with 1,000 young women, we document that these effects cannot be explained by body-positive content alone. Instead, weight-loss content worsened both short-run body image and mood. To further support our interpretation, we show that nearly half of young women prefer to avoid weight-loss content, over a third of whom were willing to pay to do so, yet consume it when served, consistent with sophistication about self-control problems. Finally, we show that parasitic content generates both engagement and advertising revenue, providing the platform with a financial incentive to serve it.

This paper establishes that social media algorithms serve parasitic content in one narrowly scoped domain: weight loss videos directed at young women with body image concerns. It does not provide representative measures of the prevalence of parasitic content in this domain or across platform-served content more generally. Future research may consider how frequently algorithms generate other parasitic pairings, which users are most exposed to them, and their longer-run welfare consequences. More broadly, our findings show that

learning more accurately what users will consume need not lead platforms to serve what users would choose to receive. As firms increasingly use behavioral data and machine learning to personalize content, understanding when consumption diverges from user well-being will become increasingly important.

References

- Acemoglu, Daron, Ali Makhdoumi, Azarakhsh Malekian, and Asuman Ozdaglar**, “When Big Data Enables Behavioral Manipulation,” *American Economic Review: Insights*, March 2025, 7 (1), 19–38.
- Agan, Amanda, Diag Davenport, Jens Ludwig, and Sendhil Mullainathan**, “Automating Automaticity: How the Context of Human Choice Affects the Extent of Algorithmic Bias,” February 2023.
- Allcott, Hunt, Juan Camilo Castillo, Matthew Gentzkow, Leon Musolf, and Tobias Salz**, “Sources of Market Power in Web Search: Evidence from a Field Experiment,” January 2025.
- , **Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow**, “The Welfare Effects of Social Media,” *American Economic Review*, March 2020, 110 (3), 629–676.
- , **Matthew Gentzkow, and Lena Song**, “Digital Addiction,” *American Economic Review*, July 2022, 112 (7), 2424–2463.
- Anderson, Michael L.**, “Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects,” *Journal of the American Statistical Association*, December 2008, 103 (484), 1481–1495.
- Aridor, Guy, Rafael Jiménez-Durán, Ro’ee Levy, and Lena Song**, “The Economics of Social Media,” *Journal of Economic Literature*, 2024.
- Auxier, Brooke and Monica Anderson**, “Social Media Use in 2021,” April 2021.
- Bala, Divya**, “How TikTok put ‘mid-size’ at the frontier of body inclusivity,” May 2021.
- Banerjee, Abhijit and Sendhil Mullainathan**, “The Shape of Temptation: Implications for the Economic Lives of the Poor,” 2010.
- Becker, Gordon M., Morris H. Degroot, and Jacob Marschak**, “Measuring utility by a single-response sequential method,” *Behavioral Science*, 1964, 9 (3), 226–232. _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/bs.3830090304>.

- Beknazar-Yuzbashev, George, Rafael Jiménez Durán, and Mateusz Stalinski**, “A Model of Harmful yet Engaging Content on Social Media,” *SSRN Electronic Journal*, 2024.
- , **Rafael Jiménez-Durán, Andrey Simonov, and Mateusz Stalinski**, “Social media advertising loads as prices,” December 2025.
- , **Rafael Jiménez Durán, Jesse McCrosky, and Mateusz Stalinski**, “Toxic Content and User Engagement on Social Media: Evidence from a Field Experiment,” *SSRN Electronic Journal*, 2025.
- Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Nichole Szembrot**, “Beyond Happiness and Satisfaction: Toward Well-Being Indices Based on Stated Preference,” *American Economic Review*, September 2014, *104* (9), 2698–2735.
- Bernheim, B Douglas, Kristy Kim, and Dmitry Taubinsky**, “Welfare and the Act of Choosing,” *Journal of Political Economy*, 2026, *Forthcoming*.
- Braghieri, Luca, Ro’ee Levy, and Alexey Makarin**, “Social Media and Mental Health,” *American Economic Review*, November 2022, *112* (11), 3660–3693.
- Bursztyn, Leonardo, Benjamin Handel, Rafael Jiménez-Durán, and Christopher Roth**, “When Product Markets Become Collective Traps: The Case of Social Media,” *American Economic Review*, December 2025, *115* (12), 4105–4136.
- Cagé, Julia, Nicolas Herve, and Beatrice Mazoyer**, “Social Media Influence Mainstream Media: Evidence from Two Billion Tweets,” February 2024.
- Campante, Filipe, Ruben Durante, Felix Hagemeister, and Ananya Sen**, “GenAI Misinformation, Trust, and News Consumption: Evidence from a Field Experiment,” September 2025.
- Clausen, Loa, Jan H. Rosenvinge, Oddgeir Friborg, and Kristian Rokkedal**, “Validating the Eating Disorder Inventory-3 (EDI-3): A Comparison Between 561 Female Eating Disorders Patients and 878 Females from the General Population,” *Journal of Psychopathology and Behavioral Assessment*, 2011, *33* (1), 101–110.
- DellaVigna, Stefano and Ulrike Malmendier**, “Contract Design and Self-Control: Theory and Evidence*,” *The Quarterly Journal of Economics*, May 2004, *119* (2), 353–402.
- and —, “Paying Not to Go to the Gym,” *American Economic Review*, June 2006, *96* (3), 694–719.
- Donati, Dante and Lena Song**, “The Influence of the Vocal Minority: Evidence from Social Media Comments,” April 2026.

- Ely, Jeffrey, Alexander Frankel, and Emir Kamenica**, “Suspense and Surprise,” *Journal of Political Economy*, 2015, 123 (1).
- EMARKETER**, “TikTok leads time spent on social for most US adults,” April 2024.
- Fryar, Cheryl, Qiuping Gu, Joseph Afful, Margaret Carroll, and Cynthia Ogden**, “Anthropometric Reference Data for Children and Adults: United States, August 2021–August 2023,” Technical Report, National Center for Health Statistics (U.S.) June 2025.
- Fudenberg, Drew and David K. Levine**, “A Dual-Self Model of Impulse Control,” *American Economic Review*, December 2006, 96 (5), 1449–1476.
- Gabaix, Xavier and David Laibson**, “Shrouded Attributes, Consumer Myopia, and Information Suppression in Competitive Markets*,” *The Quarterly Journal of Economics*, May 2006, 121 (2), 505–540.
- Gottfried, Jeffrey and Eugenie Park**, “Americans’ Social Media Use 2025,” November 2025. Section: Social Media.
- Greene, Amanda K., Elana Maloul, Devin A. Kelly, Hannah N. Norling, and Lisa M. Brownstone**, ““An Immaculate Keeper of My Social Media Feed”: Social Media Usage in Body Justice Communities During the COVID-19 Pandemic,” *Social Media + Society*, January 2022, 8 (1), 20563051221077024.
- Grogan, Sarah**, *Body Image: Understanding Body Dissatisfaction in Men, Women and Children*, Routledge, January 2002.
- Gul, Faruk and Wolfgang Pesendorfer**, “Temptation and Self-Control,” *Econometrica*, 2001, 69 (6), 1403–1435. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0262.00252>.
- Haidt, Jonathan**, *The Anxious Generation: How the Great Rewiring of Childhood Is Causing an Epidemic of Mental Illness*, Penguin, March 2024. Google-Books-ID: n9fDEAAAQBAJ.
- Heidhues, Paul and Botond Köszegi**, “Behavioral Industrial Organization,” in “Handbook of Behavioral Economics: Applications and Foundations 1,” Vol. 1, Elsevier, 2018, pp. 517–612.
- Herrman, John**, “How TikTok Is Rewriting the World,” *The New York Times*, March 2019.
- Hébert, Benjamin and Weijie Zhong**, “Engagement maximization,” *Theoretical Economics*, 2026, 21 (3), 849–887.
- Kang, Cecilia**, “What to Know About the Social Media Addiction Trials,” *The New York Times*, January 2026.

- Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Manish Raghavan**, “The Inversion Problem: Why Algorithms Should Infer Mental State and Not Just Predict Behavior,” *Perspectives on Psychological Science*, September 2024, *19* (5), 827–838.
- , **Sendhil Mullainathan, and Manish Raghavan**, “The Challenge of Understanding What Users Want: Inconsistent Preferences and Engagement Optimization,” October 2023. arXiv:2202.11776 [cs].
- Koh, Andrew and Sivakorn Sanguanmoo**, “Attention Capture,” 2024.
- Laibson, David**, “A Cue-Theory of Consumption*,” *The Quarterly Journal of Economics*, February 2001, *116* (1), 81–119.
- Levy, Ro’ee**, “Social Media, News Consumption, and Polarization: Evidence from a Field Experiment,” *American Economic Review*, March 2021, *111* (3), 831–870.
- Lindner, Danielle and Stacey Tantleff-Dunn**, “The development and psychometric evaluation of the self-objectification beliefs and behaviors scale,” *Psychology of Women Quarterly*, 2017, *41* (2), 254–272. Place: US.
- Loewenstein, George and Zachary Wojtowicz**, “The Economics of Attention,” October 2024.
- MacNeill, Lillian P. and Lisa A. Best**, “Perceived current and ideal body size in female undergraduates,” *Eating Behaviors*, August 2015, *18*, 71–75.
- McClain, Colleen and Kirsten Eddy**, “8 facts about Americans and TikTok,” March 2026. Section: TikTok.
- Merrill, Jeremy B., Caitlin Gilbert, and Clara Ence Morse**, “We teamed up with our audience to measure TikTok’s unmistakable draw,” *The Washington Post*, October 2025.
- Mosquera, Roberto, Mofioluwasademi Odunowo, Trent McNamara, Xiongfei Guo, and Ragan Petrie**, “The economic effects of Facebook,” *Experimental Economics*, June 2020, *23* (2), 575–602.
- Neumark-Sztainer, Dianne, Susan J. Paxton, Peter J. Hannan, Jess Haines, and Mary Story**, “Does body satisfaction matter? Five-year longitudinal associations between body satisfaction and health behaviors in adolescent females and males,” *The Journal of Adolescent Health: Official Publication of the Society for Adolescent Medicine*, August 2006, *39* (2), 244–251.
- Nolen, Erin and Lisa S. Panisch**, “The Relationship between Body Appreciation and Health Behaviors among Women and Adolescent Girls: A Scoping Review,” *Health & Social Work*, April 2022, *47* (2), 113–122.

- O'Donoghue, Ted and Matthew Rabin**, “Doing It Now or Later,” *American Economic Review*, March 1999, 89 (1), 103–124.
- Prolific**, “How much should I pay participants?,” April 2026.
- Quick, Virginia, Marla E. Eisenberg, Michaela M. Bucchianeri, and Dianne Neumark-Sztainer**, “Prospective Predictors of Body Dissatisfaction in Young Adults: 10-year Longitudinal Findings,” *Emerging adulthood (Print)*, December 2013, 1 (4), 271–282.
- Rambachan, Ashesh, Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan**, “An Economic Perspective on Algorithmic Fairness,” *AEA Papers and Proceedings*, May 2020, 110, 91–95.
- Rohde, Paul, Eric Stice, and C. Nathan Marti**, “Development and Predictive Effects of Eating Disorder Risk Factors during Adolescence: Implications for Prevention Efforts,” *The International journal of eating disorders*, March 2015, 48 (2), 187–198.
- Rothwell, Jonathan**, “Teens Spend Average of 4.8 Hours on Social Media Per Day,” October 2023. Section: Wellbeing.
- Saeedi, Maryam, Yikang Shen, and Ali Shourideh**, “Getting the Agent to Wait,” July 2024. arXiv:2407.19127 [econ.TH].
- Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort**, “Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms,” *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*, 2014.
- Schilder, Paul**, *The image and appearance of the human body*, London: K. Paul, Trench, Trubner & Co. Ltd., 1935.
- Shleifer, Andrei, Pedro Bordalo, Nicola Gennaioli, Giovanni Burro, and Gad Nacamura**, “Ads As Cues,” October 2025.
- Smeets, Monique A. M, J. David Ingleby, Hans W Hoek, and Geert E. M Panhuysen**, “Body size perception in anorexia nervosa: A signal detection approach,” *Journal of Psychosomatic Research*, May 1999, 46 (5), 465–477.
- Smith, Ben**, “How TikTok Reads Your Mind,” *The New York Times*, December 2021.
- Stice, E. and S. K. Bearman**, “Body-image and eating disturbances prospectively predict increases in depressive symptoms in adolescent girls: a growth curve analysis,” *Developmental Psychology*, September 2001, 37 (5), 597–607.
- TikTok**, “How TikTok recommends videos #ForYou,” June 2020.

— , “Available bidding methods,” August 2025.

Tylka, Tracy L. and Nichole L. Wood-Barcalow, “The Body Appreciation Scale-2: Item refinement and psychometric evaluation,” *Body Image*, January 2015, *12*, 53–67.

U.S. Census Bureau, “Annual Estimates of the Resident Population by Single Year of Age, Sex, Race, and Hispanic Origin for the United States: April 1, 2020 to July 1, 2024 (NC-EST2024-ALLDATA),” 2025.

Vannucci, Anna and Christine McCauley Ohannessian, “Body Image Dissatisfaction and Anxiety Trajectories during Adolescence,” *Journal of clinical child and adolescent psychology*, 2018, *47* (5), 785–795.

Watson, D., L. A. Clark, and A. Tellegen, “Development and validation of brief measures of positive and negative affect: the PANAS scales,” *Journal of Personality and Social Psychology*, June 1988, *54* (6), 1063–1070.

Wilfred, Salomé Adelia, Michelle Okyere, Shumyla Lopez, Lisa Smith Kilpela, and Carolyn Black Becker, “State of the Science: Body Image and Mental Health,” *Behavior Therapy*, September 2026, *57* (5), 789–810.

Figures and Tables

Figure 3: Field Experiment Design

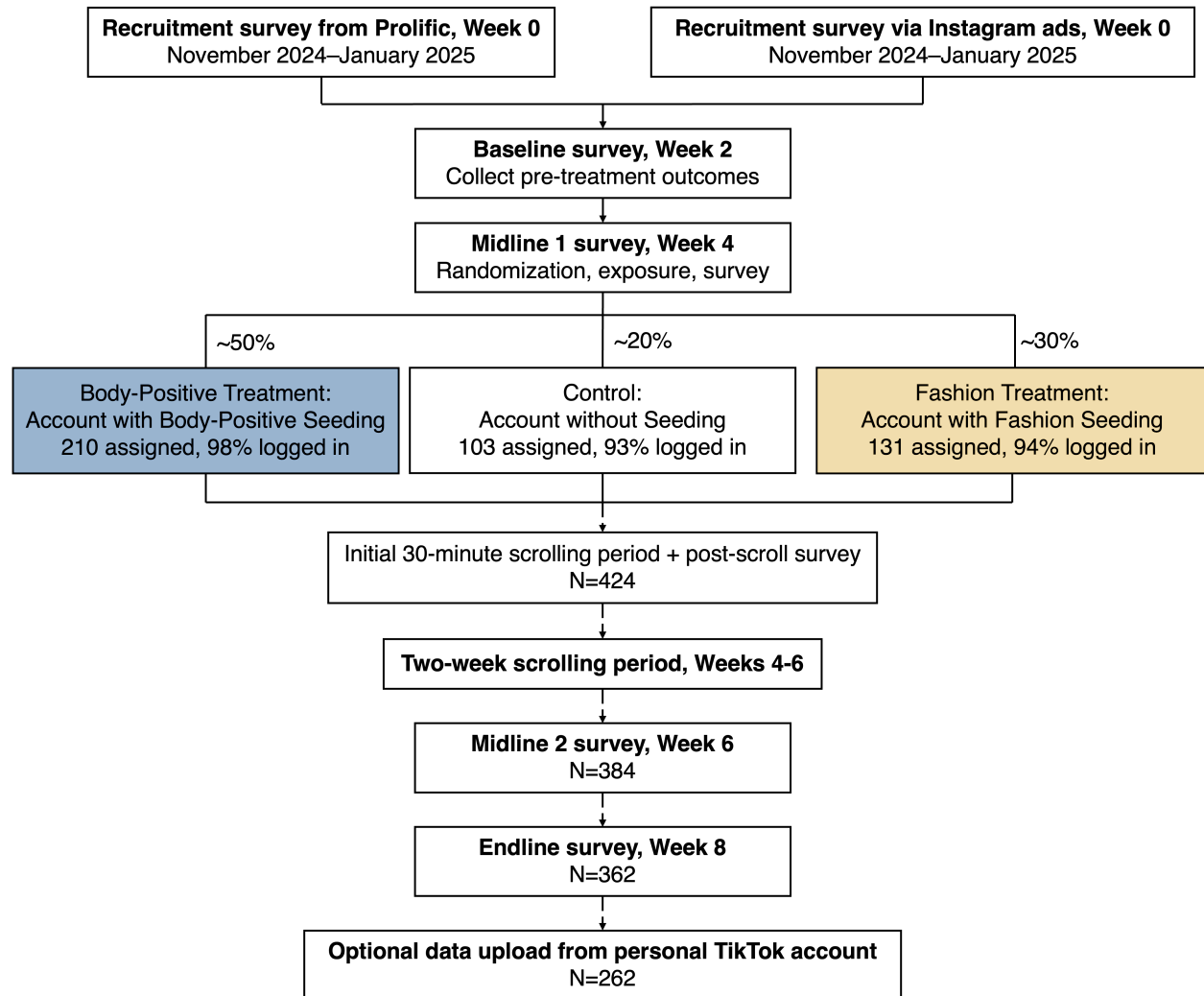
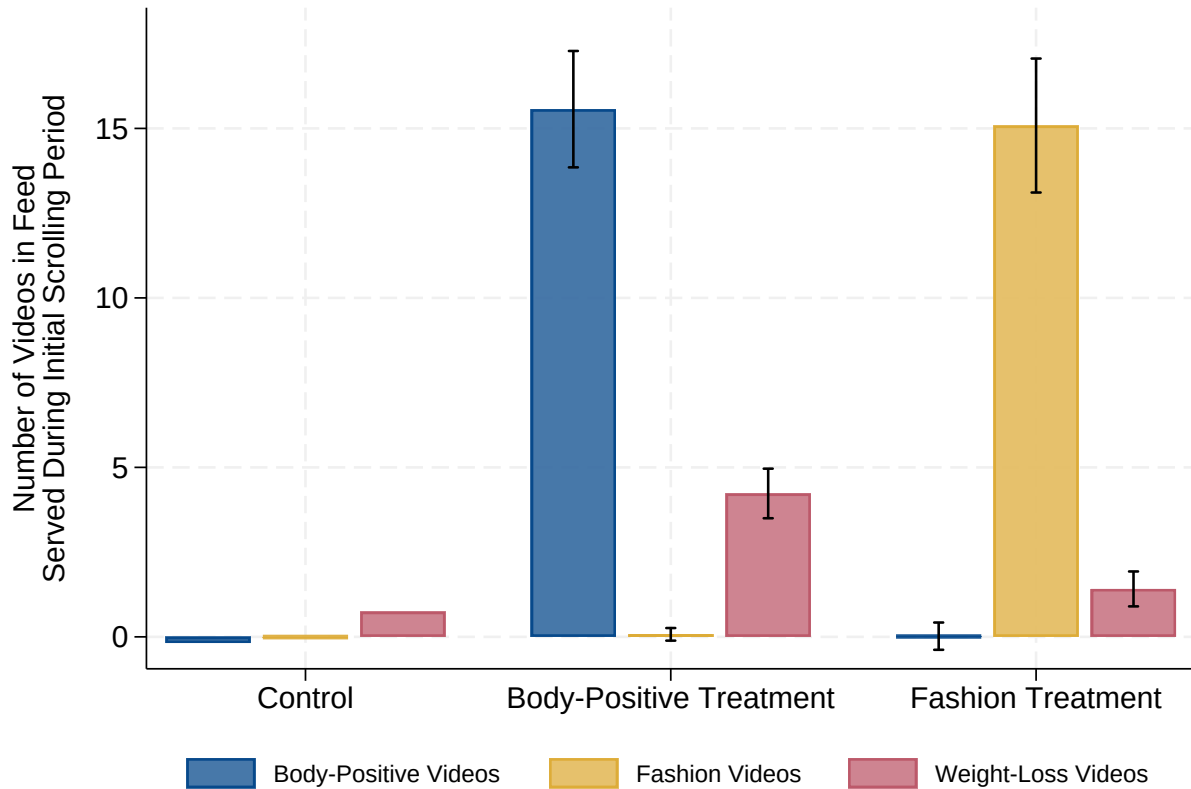
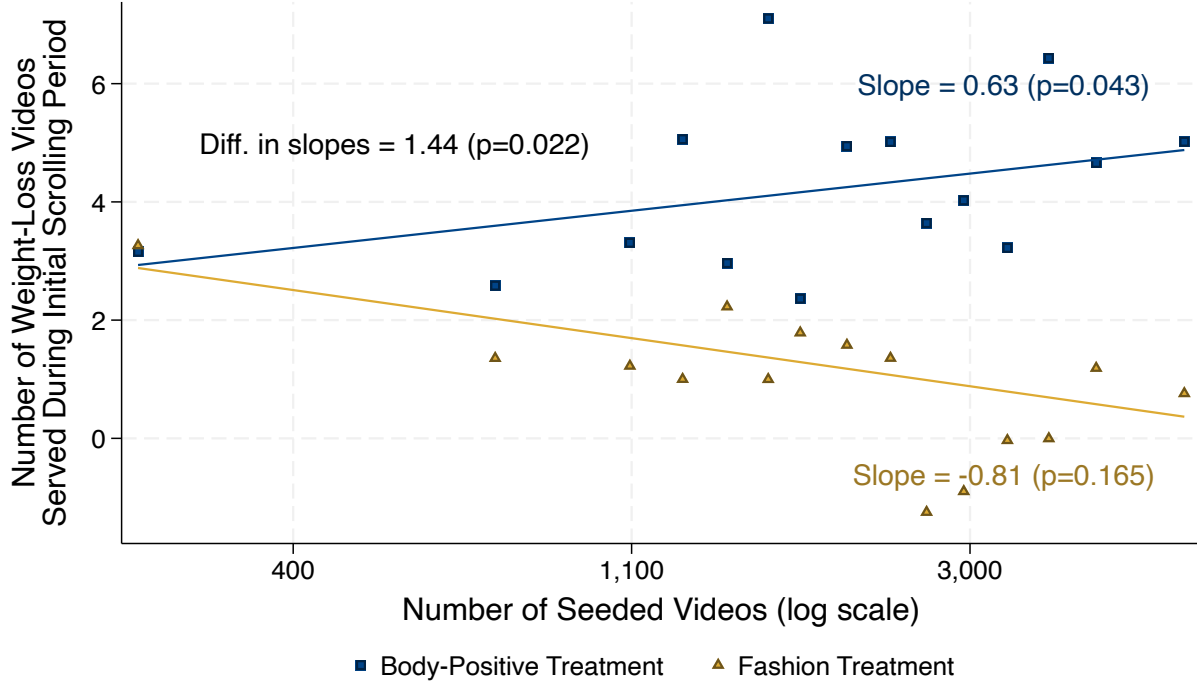


Figure 4: Treatment Effects on Content Supply



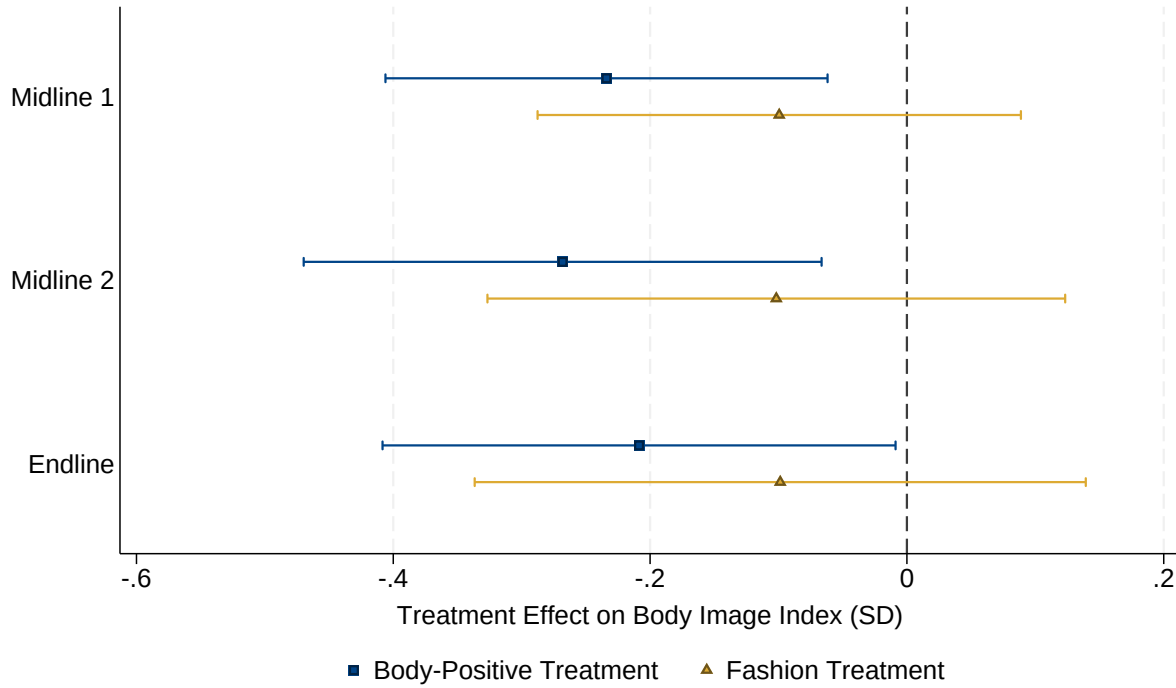
NOTE.—This figure shows the treatment effects on content supply during the initial scrolling period. We plot the average number of body-positive, fashion, and weight-loss videos served by treatment arm, controlling for participants' age, daily TikTok use in minutes, BMI, the body image index, country of residence, LGBTQ identity, recruitment platform, race, and ethnicity, all measured at baseline; missing baseline values were replaced with zero, and an indicator for any such missing value was included. (The plotted counts cover only these three genres, not all videos served.) We construct these by adding the treatment effect to the covariate-adjusted control group mean. The coefficients and confidence intervals are based on the estimates and standard errors in Table A3. Error bars indicate 95% confidence intervals.

Figure 5: Seeding Intensity and Weight-Loss Supply



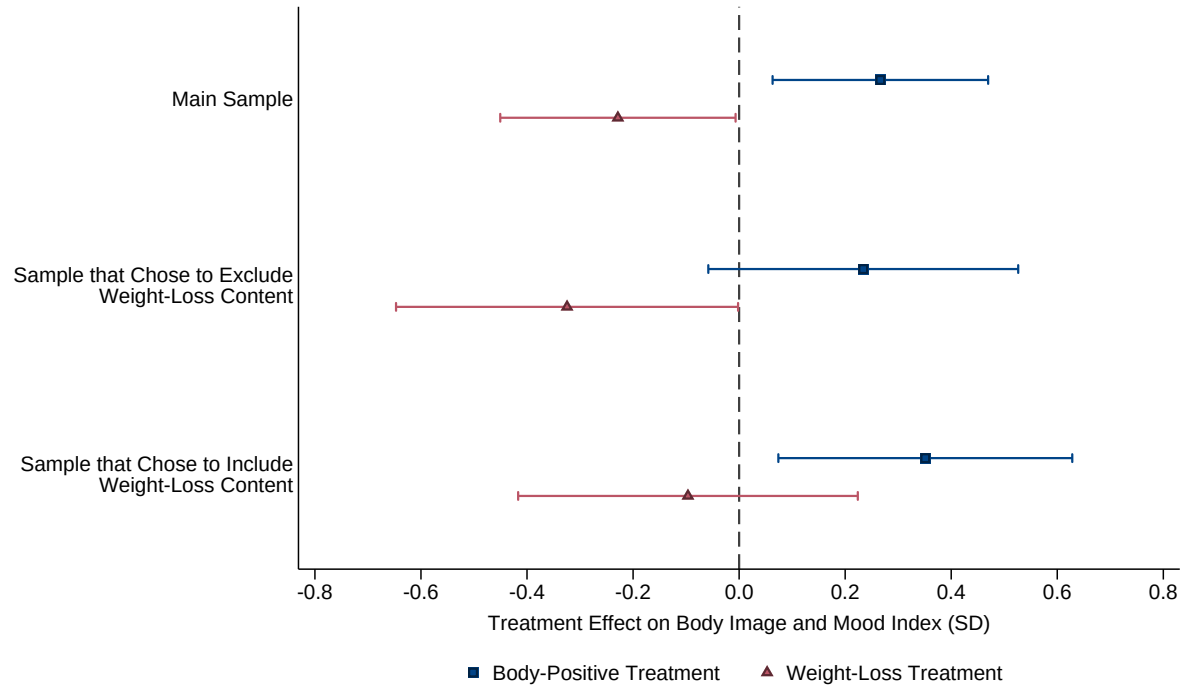
NOTE.—This figure plots covariate-adjusted binned means of the number of weight-loss videos served during the initial scrolling period against the number of videos viewed during account seeding, separately for the body-positive and fashion treatments. Each marker reports the treatment-specific mean within a pooled quantile bin after subtracting the component predicted by baseline characteristics and assignment-week fixed effects and adding back the sample average of that component. The lines plot fitted values from column 1 of Table A5, with the predicted baseline-control and assignment-week component held at its sample average. The body-positive slope is the coefficient on $\ln(\text{Seeding Intensity})$; the fashion slope is the sum of that coefficient and the interaction, with its p -value calculated by testing whether this sum equals zero. The reported difference is the body-positive slope minus the fashion slope. The displayed range is restricted to the intersection of the treatment-specific 1st–99th percentile ranges of seeding intensity. The horizontal axis is displayed on a log scale and labeled in the original number of videos. Reported p -values are based on heteroskedasticity-robust standard errors.

Figure 6: Treatment Effects on the Body Image Index



NOTE.—This figure shows treatment effects on the body image index at midline 1, midline 2, and endline, relative to the control group, denoted by the vertical line at zero, and measured in standard deviations. The plotted coefficients and confidence intervals are based on the estimates and heteroskedasticity-robust standard errors reported in columns 1, 2, and 4 of Table A6. Regressions are estimated using participants who completed each corresponding survey. In all surveys, the index combines the Body Appreciation Scale; the observer objectification and self objectification subscales of the Self-Objectification Beliefs and Behaviors Scale; indicators for reporting concern about appearance and health; self-perceived weight; and willingness to accept to make a body-positive post. The midline 2 and endline indices additionally include the Body Dissatisfaction Scale and an index of behaviors associated with poor body image, which were not measured at midline 1. Components are signed so that higher values indicate better body image and combined following [Anderson \(2008\)](#). Each regression controls for the body image index, age, daily TikTok use in minutes, BMI, indicators for LGBTQ identity, race, and ethnicity, and country-of-residence and recruitment-platform fixed effects, all measured at baseline. Missing values for age, TikTok use, race, ethnicity, and LGBTQ identity are replaced with zero, and an indicator for any such missing value is included. Error bars indicate 95% confidence intervals.

Figure 7: Treatment Effects on Body Image and Mood



NOTE.—This figure shows treatment effects on the body image and mood index in the mechanism experiment, measured in standard deviations, relative to the control group, denoted by the vertical line at zero. The plotted coefficients and confidence intervals are based on the estimates and heteroskedasticity-robust standard errors reported in columns 1–3 of Table B3. The index combines the Body Appreciation Scale, positive affect, and negative affect, with negative affect reversed so that higher values indicate better body image and mood. The mixed treatment arm is excluded, as pre-registered. Estimates are shown for the full sample and separately for participants who chose before treatment to exclude or include weight-loss content. All regressions control for age and BMI and include indicators for race, country, gender, and LGBTQ identity. Error bars indicate 95% confidence intervals.

Table 1: Field Experiment—Participant Characteristics and Balance

	Mean by treatment arm			<i>t</i> -test <i>p</i> -value		
	(1) Body Pos.	(2) Fashion	(3) Control	(1)-(2)	(2)-(3)	(1)-(3)
<i>Demographics</i>						
Female	1.00	1.00	1.00	.	.	.
Age	22.53	22.67	22.48	0.61	0.55	0.88
Based in US	0.55	0.48	0.46	0.21	0.75	0.13
Based in UK	0.33	0.32	0.41	0.79	0.17	0.21
Based in Canada	0.12	0.20	0.14	0.03	0.19	0.65
White	0.47	0.60	0.48	0.03	0.11	0.83
Black	0.16	0.08	0.16	0.05	0.09	0.98
Asian	0.12	0.15	0.12	0.42	0.48	0.96
Mixed Race	0.14	0.12	0.15	0.50	0.49	0.91
Hispanic	0.14	0.10	0.11	0.31	0.88	0.44
LGBTQ	0.29	0.38	0.26	0.11	0.07	0.58
<i>Other Baseline Characteristics</i>						
TikTok Use (mins/day)	122.94	114.83	119.15	0.44	0.71	0.75
BMI	24.93	25.40	24.61	0.55	0.38	0.69
Body Image Index	−0.08	−0.13	0.00	0.66	0.34	0.55
Recruited on Instagram	0.11	0.10	0.08	0.68	0.72	0.44
Recruited on Prolific	0.89	0.90	0.92	0.68	0.72	0.44
<i>Attrition</i>						
Initial Scrolling	1.00	1.00	1.00	.	.	.
Midline 1 Survey	1.00	1.00	1.00	.	.	.
Two-Week Scrolling	0.80	0.82	0.74	0.65	0.17	0.26
Midline 2 Survey	0.91	0.89	0.91	0.59	0.77	0.87
Endline Survey	0.88	0.82	0.84	0.16	0.66	0.42
Observations	205	123	96	.	.	.
<i>F</i> -test				0.11	0.32	0.99

NOTE.— This table reports baseline participant characteristics by treatment group and *p*-values of *t*-tests done pairwise between treatment groups. White, Black, Asian, and Mixed Race indicators are mutually exclusive. The body image index was constructed by combining the Body Appreciation Scale; the observer objectification and self objectification subscales of the Self-Objectification Beliefs and Behaviors Scale; indicators for reporting concern about appearance and health; self-perceived weight; willingness to accept to make a body-positive post; and an index of behaviors associated with poor body image following [Anderson \(2008\)](#). The attrition panel includes indicators of whether the participant indicated they completed the scrolling periods and whether they completed each survey. The *F*-test row reports *p*-values from pairwise joint orthogonality tests, calculated by restricting to the relevant treatment groups and regressing a treatment indicator on the pre-treatment variables, with missing values replaced with a constant and an indicator for a missing value.

Table 2: Cross-Genre Prediction of Content Supply

	Number of Weight-Loss Videos at t			Number of Body-Positive Videos at t		
	(1)	(2)	(3)	(4)	(5)	(6)
Weight-Loss Watch Time at $t - 1$ (100 sec.)	0.69*** (0.21)		0.69*** (0.21)	0.03 (0.15)		0.07 (0.13)
Number of Weight-Loss Videos at $t - 1$	0.23*** (0.04)		0.22*** (0.05)	0.17*** (0.05)		0.00 (0.03)
Body-Positive Watch Time at $t - 1$ (100 sec.)		0.25* (0.14)	0.08 (0.10)		0.57*** (0.19)	0.56*** (0.20)
Number of Body-Positive Videos at $t - 1$		0.07*** (0.02)	0.02 (0.02)		0.24*** (0.04)	0.24*** (0.04)
Observations	3,681	3,681	3,681	3,681	3,681	3,681
R^2	0.15	0.03	0.15	0.03	0.19	0.19
Unique Participants	369	369	369	369	369	369
Control Mean of Outcome	0.65	0.65	0.65	0.05	0.05	0.05
Control SD of Outcome	1.61	1.61	1.61	0.40	0.40	0.40

NOTE.—This table reports participant-day OLS regressions of the number of weight-loss or body-positive videos served on an observed study day on watch time and content supply on the preceding observed study day, where t is in days 1–14. When an account was not used on consecutive calendar days, $t - 1$ denotes the account's preceding observed study day rather than the preceding calendar day. Watch time is the total time spent viewing videos of the indicated genre on day $t - 1$, measured in units of 100 seconds. Columns 1–3 (4–6) use the number of weight-loss (body-positive) videos on day t as the outcome. Standard errors are clustered at the participant level and reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Weight-Loss versus Cryptocurrency Watch Time

	Watch Time (Seconds)			
	(1)	(2)	(3)	(4)
Weight-Loss Video	4.24*** (0.51)	3.60*** (0.85)	3.99*** (0.57)	3.85*** (0.81)
Observations	5,169	1,964	3,889	1,637
R^2	0.31	0.35	0.28	0.24
Participant FE	Yes	Yes	Yes	Yes
Sample	Main	WTP > 0	WTP ≥ Crypto	Exclude Body-Positive
Crypto Mean	8.10	7.03	8.33	6.96
Participants	240	90	180	71

NOTE.—This table reports regressions of watch time per video, measured in seconds, on an indicator equal to one if the video is about weight loss. Observations are restricted to weight-loss and cryptocurrency videos shown to participants assigned to the mixed treatment who chose to exclude both weight-loss and cryptocurrency content. Column 2 additionally restricts to participants with positive willingness to pay (WTP) to exclude weight-loss content. Column 3 restricts to participants whose WTP to exclude weight-loss content is at least as high as WTP to exclude cryptocurrency. Column 4 restricts to participants who also preferred to exclude body-positive content. All regressions include participant fixed effects. Standard errors are clustered at the participant level and reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: Engagement and Billable Ad Views

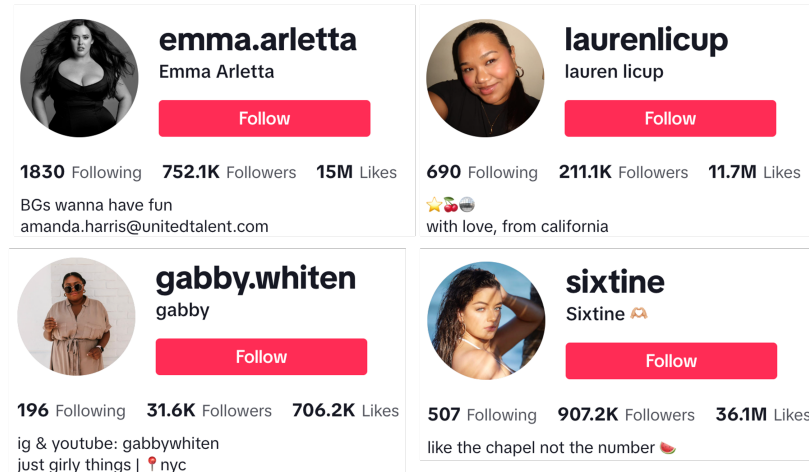
	Engagement (Seconds)	Difference	Billable View	Difference
Weight-Loss Content Served in Body-Positive Treatment	3.34*** (0.89)		0.14** (0.06)	
Non-Weight-Loss Content Served in Body-Positive Treatment	3.08*** (0.59)	0.26 (0.63)	0.02 (0.03)	0.12** (0.05)
Content Served in Fashion Treatment	1.71*** (0.49)	1.63** (0.76)	0.03 (0.03)	0.11* (0.06)
Observations	66,601		6,259	
Participants	416		348	
Control Group Mean	8.49		0.19	
Control Group SD	17.16		0.39	

NOTE.—This table reports participant-video OLS regressions of engagement and billable views during the initial scrolling period. The Engagement (Seconds) and Billable View columns report coefficients on indicators for the content category named in the row, with the control treatment omitted. Each adjacent Difference column reports the coefficient on weight-loss content served in the body-positive treatment minus the coefficient for the category named in the row; the entry is left blank in the weight-loss row because its outcome-column coefficient already reports the difference relative to the control treatment. Engagement is watch time per non-ad video in seconds. Billable View is an indicator equal to one when an ad was watched for at least six seconds, so its coefficients are proportions. The sample includes non-ad videos or ads with watch time no greater than 180 seconds from participants who completed the first midline survey. Control Group Mean and Control Group SD are the unadjusted outcome mean and standard deviation among control-treatment observations in the corresponding estimation sample. All regressions control for baseline BMI, age, and demand for fashion, humor, news, beauty, animal, and food content; race, country, LGBTQ identity, and recruitment-platform fixed effects; the numbers of ads and videos served during the session; and video order, duration, comments, plays, likes, and shares. Missing baseline and video-control values are replaced with zero; regressions include an indicator for any missing baseline control and separate missing-value indicators for each video control. Standard errors are clustered by participant and reported in parentheses. Stars in the outcome columns correspond to tests that the coefficient equals zero; stars in the Difference columns correspond to tests that the displayed difference equals zero. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Appendix

A Appendix: Field Experiment

Figure A1: Influencers Followed During Seeding

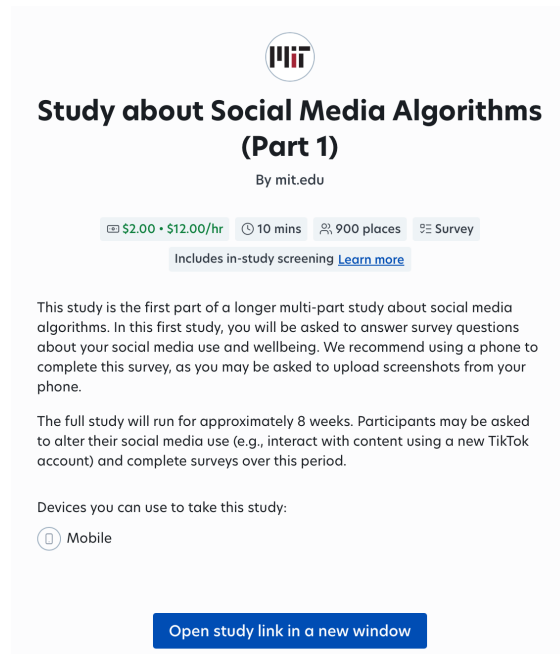


(a) Body-Positive Influencers



(b) Fashion Influencers

Figure A2: Recruitment Materials



The image shows a recruitment page for a study titled "Study about Social Media Algorithms (Part 1)" by mit.edu. The page features the MIT logo at the top. Below the title, it states "By mit.edu". A green bar indicates a payment of "\$2.00 • \$12.00/hr", a duration of "10 mins", and "900 places" available. A "Survey" button is also present. A link "Includes in-study screening [Learn more](#)" is provided. The text describes the study as the first part of a longer multi-part study about social media algorithms, where participants will answer survey questions about their social media use and wellbeing. It recommends using a phone to complete the survey, as participants may be asked to upload screenshots. The full study will run for approximately 8 weeks, and participants may be asked to alter their social media use (e.g., interact with content using a new TikTok account) and complete surveys over this period. A section titled "Devices you can use to take this study:" shows a "Mobile" option with a mobile phone icon. At the bottom, there is a blue button that says "Open study link in a new window".

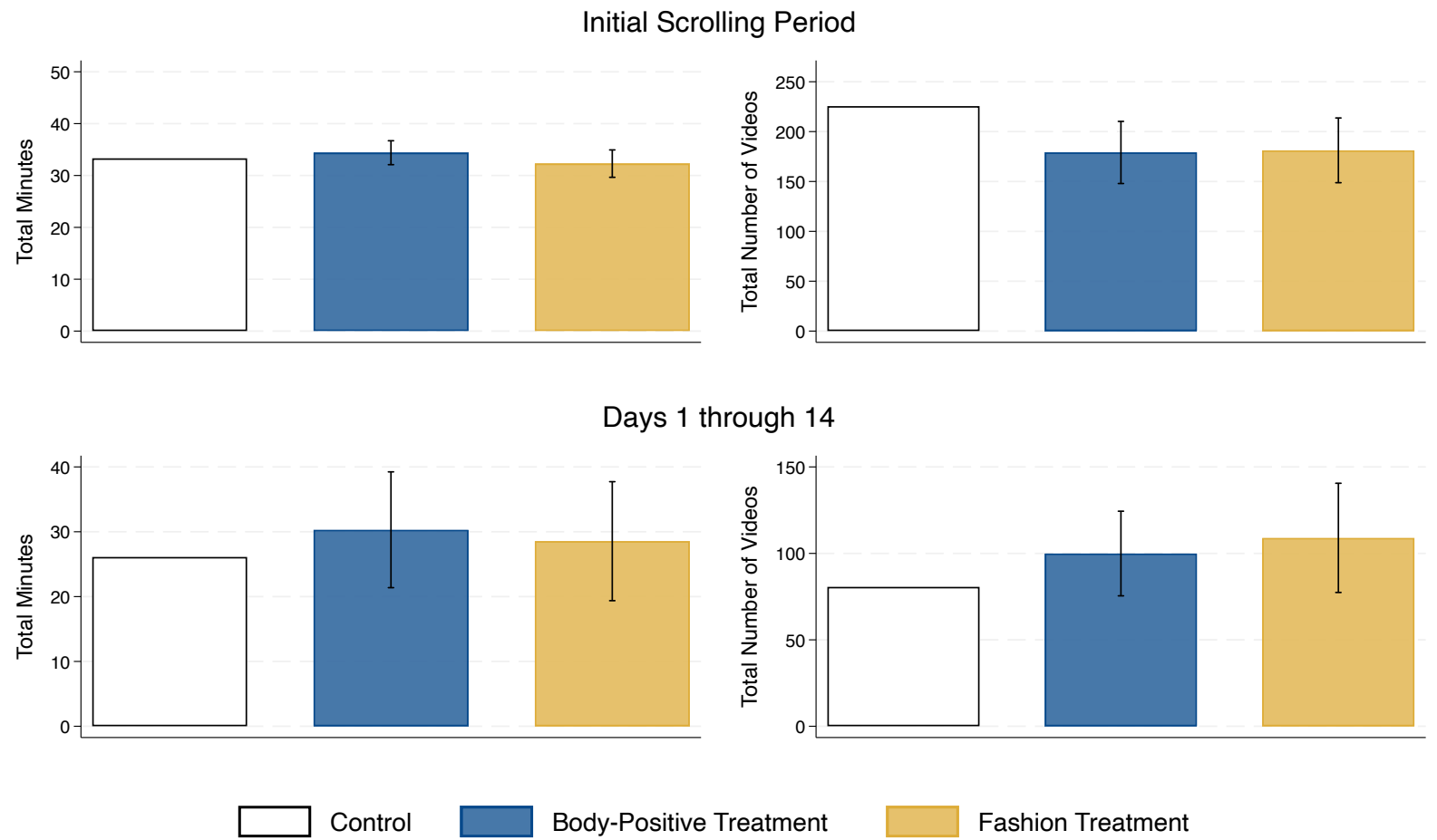
(a) Recruitment from Prolific



The image shows a recruitment advertisement for a study titled "Earn an easy \$30! Participate in an online MIT study about social media!". The advertisement is displayed on a mobile phone screen. At the top, it says "mit_columbia_research sponsored". Below this, the text "Earn an easy \$30!" is prominently displayed. Underneath, it says "Participate in an online MIT study about social media!". A row of social media icons (YouTube, Instagram, Facebook, TikTok, and Snapchat) is shown, with hands holding up the phones displaying these icons. At the bottom, there is a blue button with a link icon and the text "LEARN MORE".

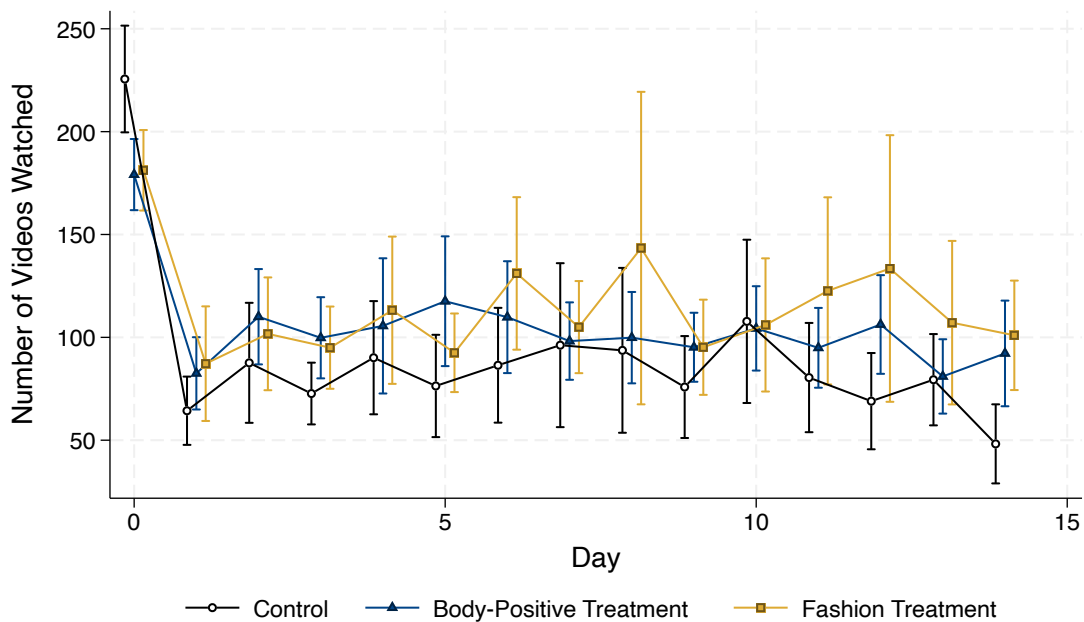
(b) Recruitment from Instagram

Figure A3: Study Account Use by Treatment

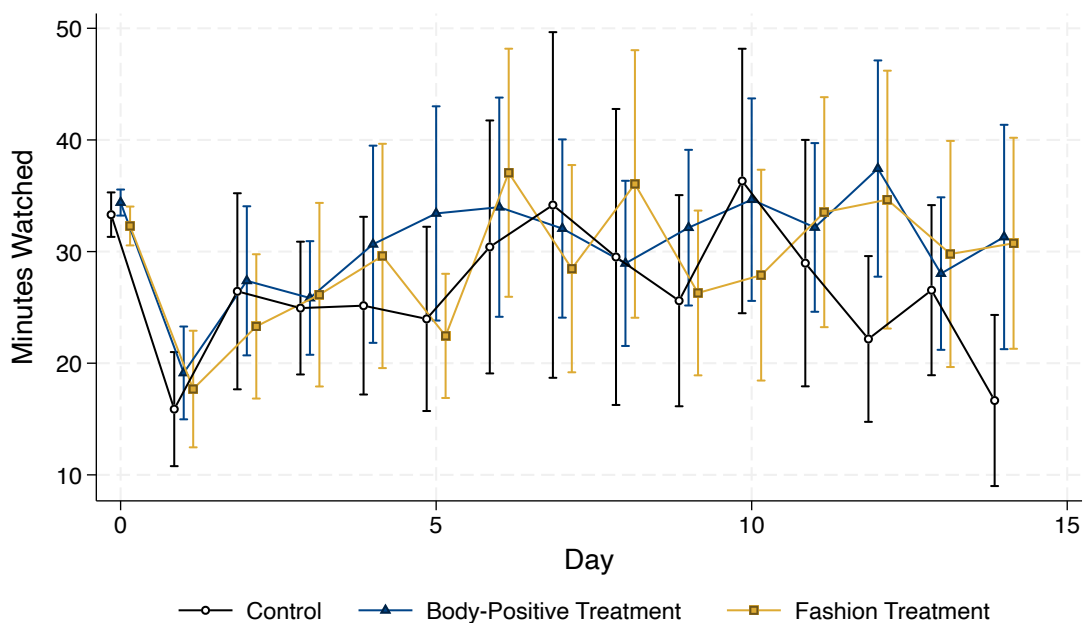


NOTE.—Difference in means in total number of videos consumed and total minutes spent by treatment assignment, for day 0 and days 1-14. Error bars indicate 95% confidence intervals. Plots for days 1-14 additionally residualize outcomes on day fixed effects. Standard errors are clustered at the participant level.

Figure A4: Study Account Use by Day



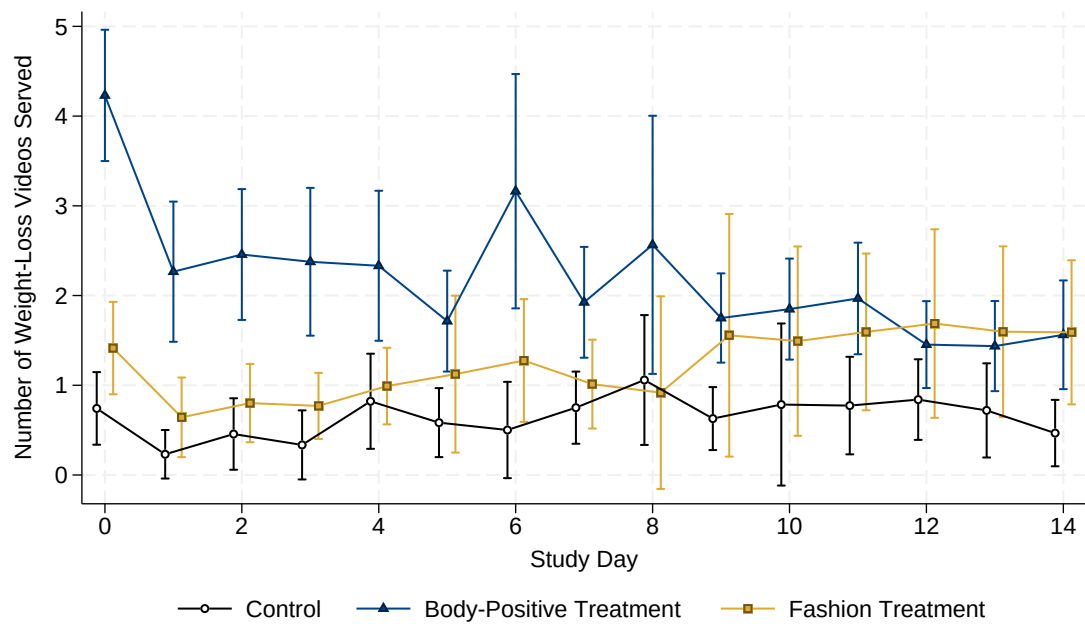
(a) Videos Viewed



(b) Minutes Viewed

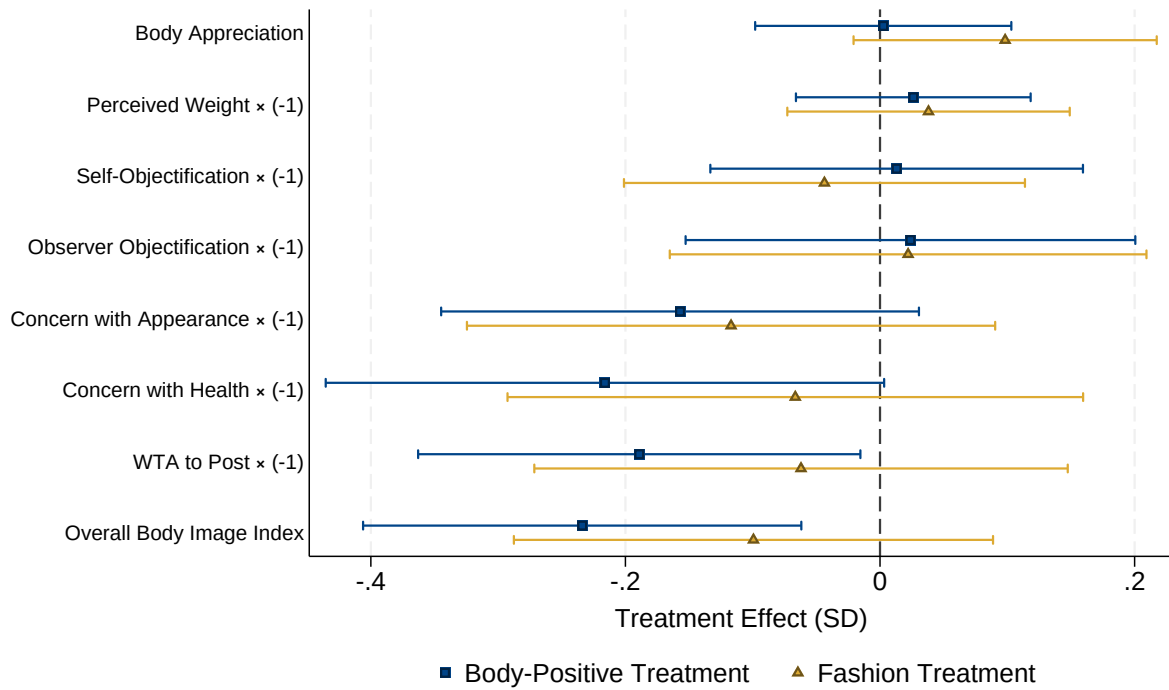
NOTE.— Day-by-day account usage during the experiment, conditional on a participant using the account on a given day. Error bars indicate 95% confidence intervals. Standard errors are robust.

Figure A5: Weight-Loss Content Supply by Day



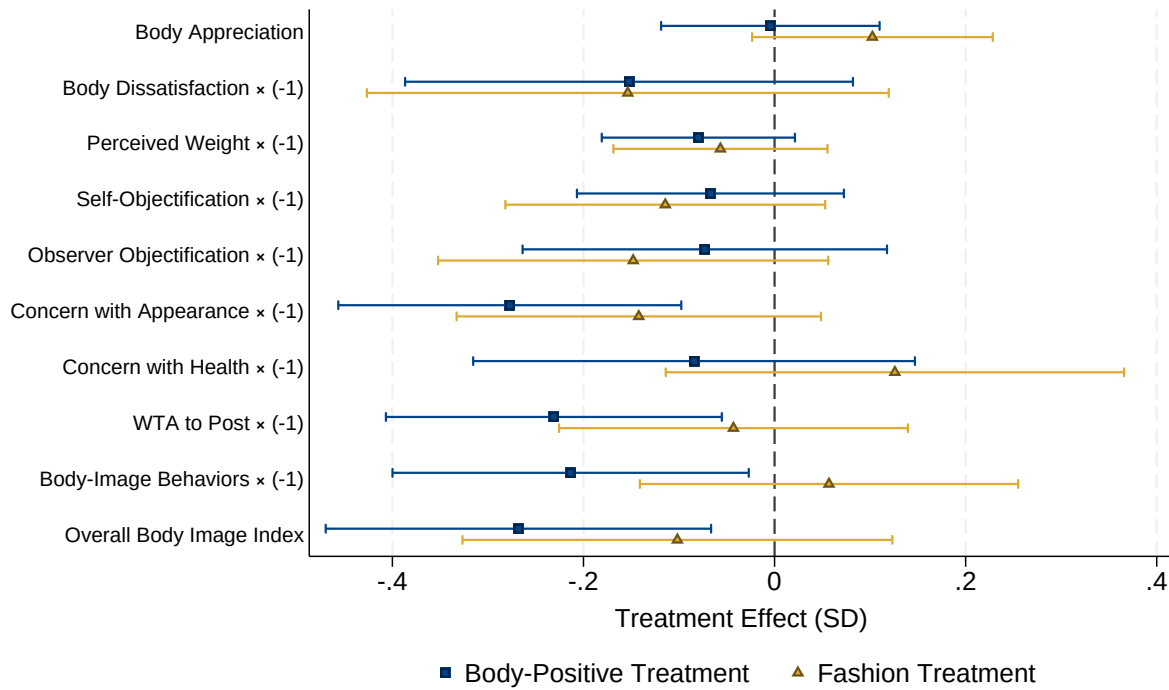
NOTE.—Weight-loss content supplied by the algorithm over the course of the experiment. Error bars indicate 95% confidence intervals.

Figure A6: Treatment Effects on Body Image Components—Midline 1



NOTE.—This figure reports OLS estimates of the effects of assignment to the body-positive and fashion treatments on body-image outcomes at midline 1, relative to the control, among participants who completed the midline 1 survey. Midline 1 was administered immediately after the initial scrolling period. Outcomes are standardized and signed so that higher values indicate better body image. Each coefficient comes from a separate regression that controls for the outcome's baseline measure and the other baseline participant characteristics reported in Table 1. Error bars indicate 95% confidence intervals based on heteroskedasticity-robust standard errors.

Figure A7: Treatment Effects on Body Image Components—Midline 2



NOTE.—This figure reports OLS estimates of the effects of assignment to the body-positive and fashion treatments on body-image outcomes at midline 2, relative to the control, among participants who completed the midline 2 survey. Outcomes are standardized and signed so that higher values indicate better body image. Each coefficient comes from a separate regression that controls for the outcome's baseline measure, when available, and the other baseline participant characteristics reported in Table 1. Body dissatisfaction was not measured at baseline. Error bars indicate 95% confidence intervals based on heteroskedasticity-robust standard errors.

Figure A8: Weight-Loss Content During Seeding

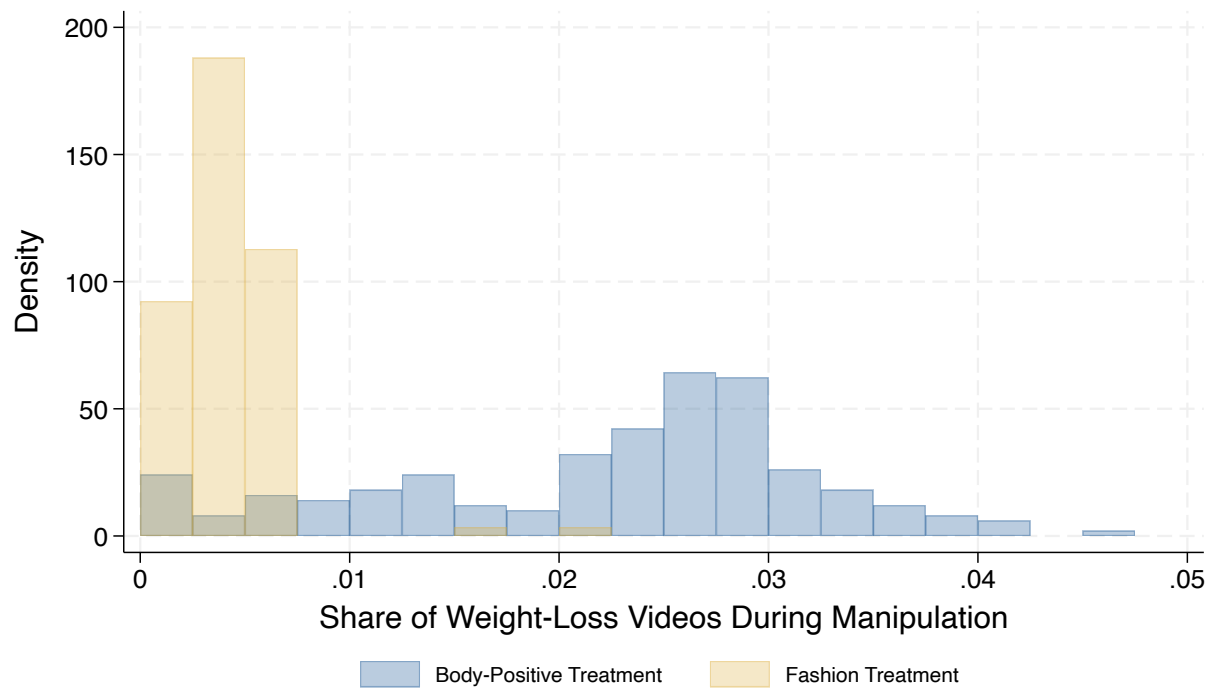
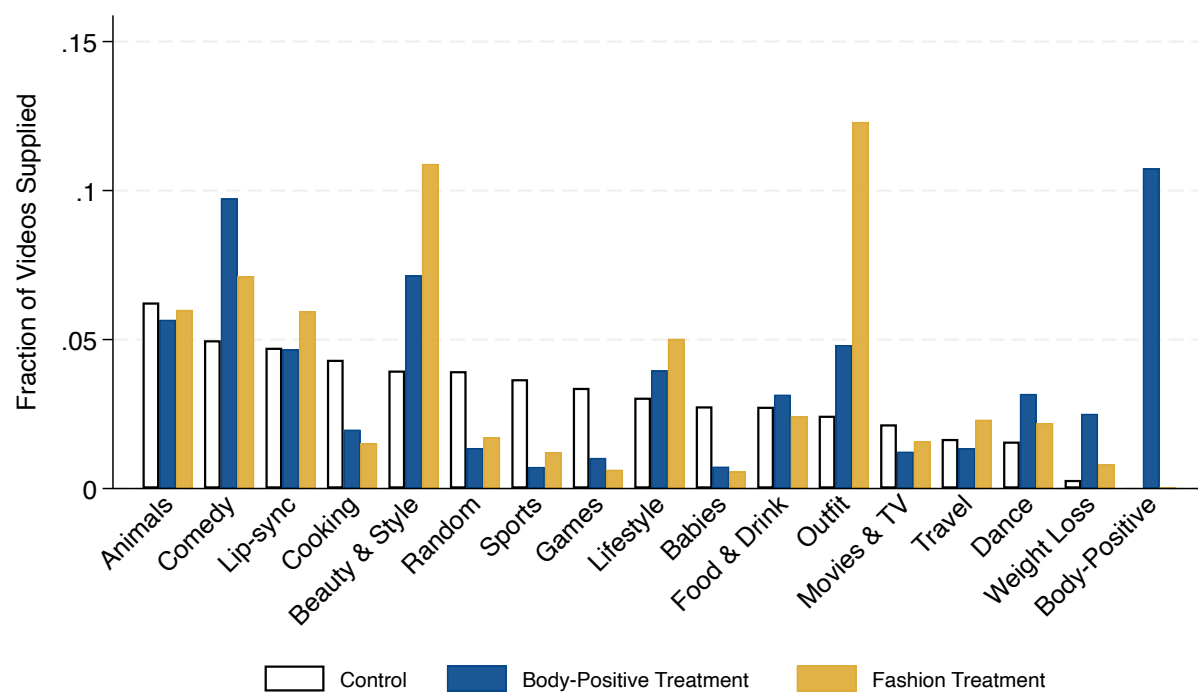
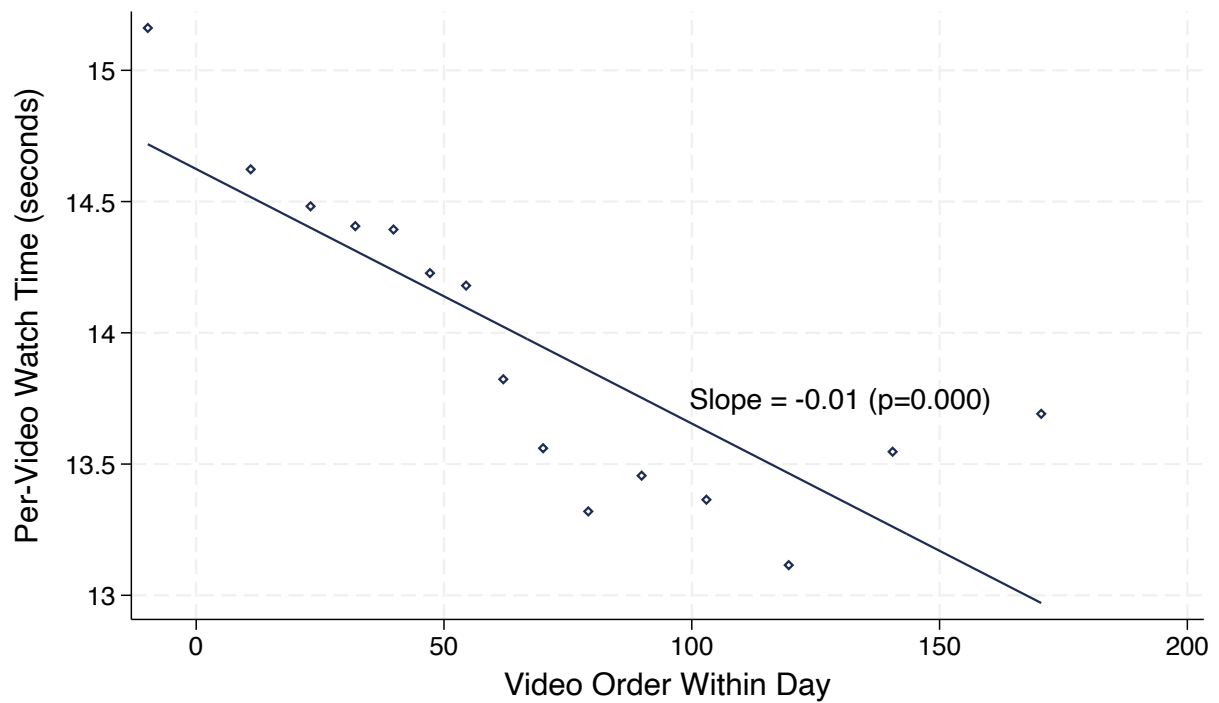


Figure A9: Content Shares by Genre



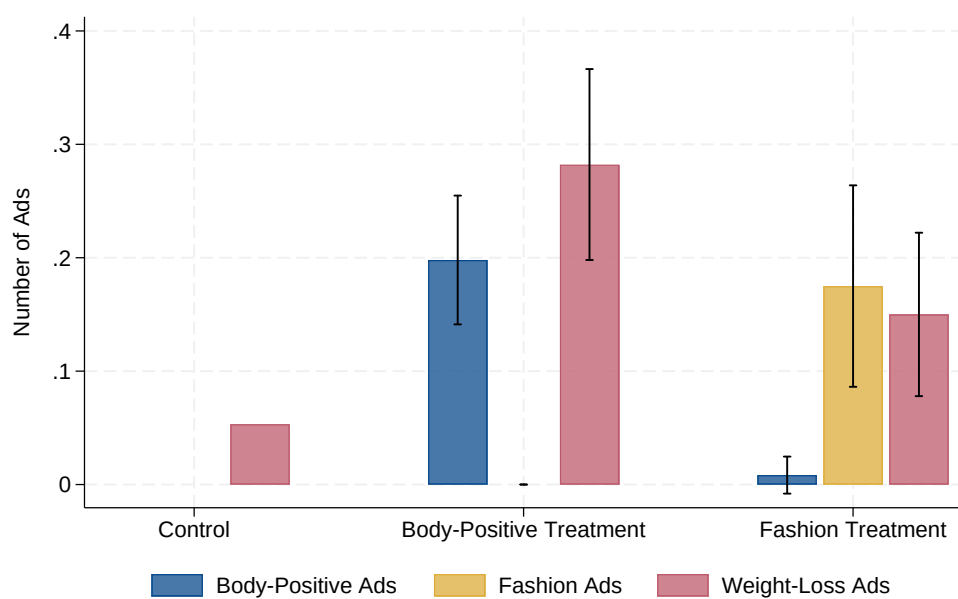
NOTE.—Top genres shown to each account type during the initial scrolling period. Genres were excluded if below the median genre popularity.

Figure A10: Watch Time by Video Order



NOTE.—Binscatter of seconds watched per video against the video's order within the day. Sample restricted to videos watched for fewer than 180 seconds and to videos with within-day ordinal position below 200. The figure reports the regression coefficient and participant-clustered standard error. Both the binscatter and regression include participant and day fixed effects.

Figure A11: Advertising Content Supply



NOTE.—This figure plots the mean numbers of body-positive, fashion, and weight-loss ads served during the initial scrolling period by treatment group. Error bars indicate 95% confidence intervals for the treatment groups; consistent with the paper’s treatment-effect figures, no error bars are shown for the control.

Table A1: Seeding Intensity Balance

	Seeding Intensity (Thousands of Videos)	
	(1) Body-Positive	(2) Fashion
<i>Demographics</i>		
Age	−0.07 (0.11)	0.04 (0.14)
Based in US	−0.03** (0.01)	−0.04 (0.02)
Based in UK	0.02 (0.02)	0.10*** (0.02)
Based in Canada	0.01 (0.02)	−0.06** (0.03)
White	−0.01 (0.02)	0.03 (0.03)
Black	−0.01 (0.02)	0.00 (0.02)
Asian	0.00 (0.02)	−0.01 (0.02)
Mixed Race	0.00 (0.03)	−0.01 (0.02)
Hispanic	−0.01 (0.01)	−0.02* (0.01)
LGBTQ	0.02 (0.03)	−0.01 (0.03)
<i>Other Baseline Characteristics</i>		
TikTok Use (mins/day)	5.00 (4.65)	−5.00 (6.56)
BMI	−0.43 (0.30)	−0.50 (0.39)
Body Image Index	0.00 (0.06)	−0.08 (0.06)
Recruited on Instagram	0.00 (0.01)	0.00 (0.02)
Recruited on Prolific	0.00 (0.01)	0.00 (0.02)
Observations	202	120
Week FE	Y	Y
<i>F</i> -test	0.42	0.35

NOTE.—This table reports OLS estimates from separate regressions of each baseline participant characteristic on seeding intensity, measured in thousands of videos, and initial-scrolling-week fixed effects, using all nonmissing observations for the corresponding characteristic. Column 1 restricts the sample to participants in the body-positive treatment, and column 2 restricts the sample to participants in the fashion treatment. Heteroskedasticity-robust standard errors are reported in parentheses. The *F*-test row reports the *p*-value from a joint orthogonality test obtained by regressing seeding intensity on the baseline characteristics and week fixed effects and testing whether the coefficients on the baseline characteristics are jointly zero. For this joint test, missing recruitment characteristics are replaced with zero and an indicator for missing recruitment characteristics is included. The observations row reports the number of participants included in the joint test. **p* < 0.10, ***p* < 0.05, ****p* < 0.01.

Table A2: Field Experiment Sample and Relevant Populations

	Experimental sample (1)	U.S. women ages 18–26 (2)	U.S. TikTok users (3)	U.S. adults (4)
Age	22.60	22.00	39.80	48.20
White	0.51	0.72	0.46	0.76
Black	0.14	0.15	0.18	0.13
Asian	0.13	0.07	0.06	0.07
Mixed race	0.14	0.05	.	0.02
Hispanic	0.12	0.25	0.28	0.18
BMI	25.00	27.90	.	29.70
Uses TikTok	1.00	.	1.00	0.37
Daily minutes on TikTok	119.70	114.00	90.00	34.00

NOTE.—Column 1 reports means for the 424 participants randomized in the field experiment. Columns 2 and 4 report characteristics of U.S. women ages 18–26 and U.S. adults, respectively. Demographics use July 2024 Census population estimates ([U.S. Census Bureau, 2025](#)); single-race categories include Hispanic individuals. BMI in column 2 uses 2020 NHANES data for women ages 18–26; column 4 uses the sex-weighted mean for adults ages 20 and older from [Fryar et al. \(2025\)](#).

Column 3 estimates characteristics of U.S. adult TikTok users. Each demographic share equals the group’s TikTok-use rate ([Gottfried and Park, 2025](#)) times its Census population share, divided by the overall use rate of 37%. Mean age weights each single year of age by its population and the TikTok-use rate for its Pew age group (18–29, 30–49, 50–64, or 65+). Race categories are non-Hispanic single-race, following Pew. The adult TikTok-use rate in column 4 is from [Gottfried and Park \(2025\)](#).

Daily TikTok minutes in column 2 refer to U.S. girls ages 13–19 in 2023 ([Rothwell, 2023](#)), rather than women ages 18–26. For columns 3 and 4, we calculate total adult TikTok minutes by multiplying age-specific minutes among social media users ([EMARKETER, 2024](#)) by age-specific social-media-use shares ([Auxier and Anderson, 2021](#)) and Census population counts, then summing across age groups. We divide this total by the estimated number of adult TikTok users for column 3 and by the adult population for column 4. Dots indicate that no comparable estimate is available.

Table A3: Treatment Effects on Content Supply

VARIABLES	(1) Body-Positive Videos	(2) Fashion Videos	(3) Weight-Loss Videos
Body-Positive Treatment	15.75*** (0.92)	0.09 (0.17)	3.49*** (0.43)
Fashion Treatment	0.20 (0.30)	15.10*** (1.02)	0.67** (0.34)
Observations	416	416	416
R^2	0.47	0.58	0.18
Baseline controls	Yes	Yes	Yes
Control mean	0.00	0.06	0.80

NOTE.—This table reports OLS estimates corresponding to the content categories shown in Figure 4. The outcomes are the numbers of body-positive videos, fashion videos, and weight-loss videos supplied during the initial scrolling period. The coefficients report the effects of assignment to the body-positive and fashion treatments relative to the control group. All specifications control for the baseline participant characteristics reported in Table 1. Missing recruitment characteristics are replaced with zero, and an indicator for missing recruitment characteristics is included. The control mean is the unadjusted mean of the dependent variable for the control group in the estimation sample. Heteroskedasticity-robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Treatment Effects on Content Supply—Robustness Checks

	Weight-Loss Videos					
	Number					Share
	(1) Days 1–7	(2) Days 8–14	(3) Excluding Influencers Followed	(4) WL Seeding Share Control	(5) Any WL Seeding Control	(6) Initial scrolling period
Body-Positive Treatment	1.78*** (0.30)	1.03*** (0.23)	3.14*** (0.42)	3.89*** (0.62)	3.85*** (1.05)	0.02*** (0.00)
Fashion Treatment	0.42** (0.20)	0.69* (0.35)	0.63* (0.34)	0.69** (0.34)	1.02 (1.06)	0.01*** (0.00)
Observations	2,012	1,758	416	416	416	416
R^2	0.06	0.04	0.16	0.18	0.18	0.22
Baseline controls	Yes	Yes	Yes	Yes	Yes	Yes
Day fixed effects	Yes	Yes	No	No	No	No
Weight-loss seeding share control	No	No	No	Yes	No	No
Any weight-loss seeding control	No	No	No	No	Yes	No
Control mean	0.58	0.71	0.80	0.80	0.80	0.00

NOTE.—This table reports OLS estimates of the effects of assignment to the body-positive and fashion treatments on the algorithm’s supply of weight-loss content, relative to the control group. Columns 1 and 2 use the number of weight-loss videos supplied per active participant-day during days 1–7 and days 8–14, respectively, and include day fixed effects. Column 3 uses the number supplied during the initial scrolling period after excluding videos posted by influencers followed during account seeding. Columns 4 and 5 use the number supplied during the initial scrolling period and control, respectively, for the share of weight-loss videos and an indicator for any weight-loss video viewed during account seeding. Column 6 uses the share of videos classified as weight-loss content during the initial scrolling period. All specifications control for the baseline participant characteristics reported in Table 1. Missing recruitment characteristics are replaced with zero, and an indicator for missing recruitment characteristics is included. The control mean is the unadjusted mean of the dependent variable for the control group in the corresponding estimation sample. Standard errors are clustered at the participant level in columns 1 and 2 and are heteroskedasticity robust in columns 3–6; all standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: Seeding Intensity and Weight-Loss Supply—Robustness Checks

	Weight-Loss Content				
	Number				Share
	(1)	(2)	(3)	(4)	(5)
Fashion Treatment	−2.87*** (0.51)	−2.55*** (0.50)	−4.40*** (0.90)	−2.86*** (0.51)	−0.02*** (0.00)
Ln(Seeding Intensity)	0.63** (0.31)	0.57* (0.29)	1.94*** (0.53)	0.76* (0.45)	0.01*** (0.00)
Fashion Treatment X Ln(Seeding Intensity)	−1.44** (0.63)	−1.29** (0.60)	−2.48*** (0.71)	−1.43** (0.61)	−0.01*** (0.00)
Observations	322	322	322	322	322
R^2	0.17	0.16	0.19	0.17	0.20
Assignment-Week FEs	Y	Y	Y	Y	Y
Baseline Controls	Y	Y	Y	Y	Y
Weight-Loss Seeding Control	N	N	Y	N	N
Any Weight-Loss Seeding Control	N	N	N	Y	N

NOTE.—This table reports OLS estimates of the effects of seeding intensity on the weight-loss content supplied during participants' initial scrolling period. The sample consists of participants assigned to the body-positive or fashion treatment. Ln(Seeding Intensity) is the log of the number of seeded videos, demeaned over the pooled estimation sample. Fashion Treatment is an indicator for assignment to the fashion rather than the body-positive treatment. Thus, the coefficient on ln(Seeding Intensity) is the slope for the body-positive treatment, the interaction coefficient is the difference between the fashion and body-positive slopes, and the coefficient on Fashion Treatment is the difference between treatment arms at mean log seeding intensity. The implied effect of doubling the number of seeded videos is obtained by multiplying the relevant treatment-specific slope by ln(2): the coefficient on ln(Seeding Intensity) for the body-positive treatment and the sum of that coefficient and the interaction coefficient for the fashion treatment. Columns 1–4 use the number of weight-loss videos supplied as the outcome. Column 1 reports the main specification. Column 2 excludes weight-loss videos supplied by influencers followed during account seeding. Column 3 controls for the number of weight-loss videos viewed during account seeding, and column 4 controls for an indicator for viewing any weight-loss video during account seeding. Column 5 uses the share of supplied videos classified as weight-loss content, measured from zero to one, as the outcome. All specifications include the baseline participant characteristics reported in Table 1 and assignment-week fixed effects. Missing recruitment covariates are set to zero, and an indicator for any missing recruitment information is included. Heteroskedasticity-robust standard errors are reported in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A6: Treatment Effects on the Body Image Index—Robustness Checks

VARIABLES	(1) Midline 1	(2) Midline 2	(3) Midline 2	(4) Endline
Body-Positive Treatment	−0.23*** (0.09)	−0.27*** (0.10)	−0.30*** (0.11)	−0.21** (0.10)
Fashion Treatment	−0.10 (0.10)	−0.10 (0.12)	−0.10 (0.12)	−0.10 (0.12)
Observations	423	383	358	361
R^2	0.49	0.48	0.49	0.45
Midline 2 TikTok Use Control	No	No	Yes	No
Control mean	0.00	0.00	0.00	0.00

NOTE.—This table reports treatment effects on the body image index, relative to the control arm. Columns 1, 2, and 4 report estimates at midline 1, midline 2, and endline, respectively. Column 3 repeats column 2 and additionally controls for TikTok use measured at midline 2. Each column restricts the sample to participants who completed the corresponding survey. The index combines the body image outcomes available in that survey, orients each component so that higher values indicate better body image, and is standardized using the control group. All regressions control for baseline body image index and the other baseline characteristics reported in Table 1. Missing recruitment characteristics are replaced with zero, with an indicator for missingness included. Heteroskedasticity-robust standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A7: Engagement and Billable Ad Views—Robustness Checks

	Engagement (Seconds)	Difference	Billable View	Difference
<i>Panel A. Marginal Sample</i>				
Weight-Loss Content Served in Body-Positive Treatment	2.19** (0.96)		0.18*** (0.07)	
Non-Weight-Loss Content Served in Body-Positive Treatment	2.78*** (0.77)	−0.60 (1.06)	0.05 (0.04)	0.12** (0.06)
Content Served in Fashion Treatment	1.54*** (0.57)	0.65 (0.88)	0.07* (0.04)	0.11* (0.06)
Observations	36,674		4,334	
Participants	416		347	
Control Group Mean	10.27		0.21	
Control Group SD	19.86		0.41	
<i>Panel B. Positive Baseline Demand for Body-Positive Content</i>				
Weight-Loss Content Served in Body-Positive Treatment	4.22*** (1.06)		0.32*** (0.07)	
Non-Weight-Loss Content Served in Body-Positive Treatment	3.33*** (0.69)	0.89 (0.92)	0.24*** (0.04)	0.08 (0.06)
Content Served in Fashion Treatment	0.90 (0.63)	3.32*** (1.03)	0.13*** (0.05)	0.19** (0.07)
Observations	31,025		1,167	
Participants	186		138	
Control Group Mean	8.45		0.20	
Control Group SD	17.23		0.40	

NOTE.—This table reports robustness tests from participant-video OLS regressions of engagement and billable views during the initial scrolling period. The Engagement (Seconds) and Billable View columns report coefficients on indicators for the content category named in the row, with the control treatment omitted. Each adjacent Difference column reports the coefficient on weight-loss content served in the body-positive treatment minus the coefficient for the category named in the row; the entry is left blank in the weight-loss row because its outcome-column coefficient already reports the difference relative to the control treatment. Panel A applies marginal-video restrictions. For engagement, it retains the first five weight-loss videos and last 100 non-weight-loss videos served in the body-positive treatment and the first 100 non-weight-loss videos served in the control and fashion treatments. For billable views, it retains all ads served in the body-positive treatment and the last five non-weight-loss ads served in the control and fashion treatments. Panel B uses the full samples but restricts to participants reporting positive demand for body-positive content at baseline. All samples exclude watch times greater than 180 seconds and include only participants who completed the first midline survey. Engagement is watch time per non-ad video in seconds. Billable View is an indicator equal to one when an ad was watched for at least six seconds, so its coefficients are proportions. Control Group Mean and Control Group SD are the unadjusted outcome mean and standard deviation among control-treatment observations in the corresponding estimation sample. All regressions use the controls, missing-value treatment, and participant-clustered standard errors described in Table 4; standard errors are reported in parentheses. Stars in the outcome columns correspond to tests that the coefficient equals zero; stars in the Difference columns correspond to tests that the displayed difference equals zero. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A1 TikTok Content: Weight-Loss Classification

We used GPT-4o to classify TikTok videos as promoting weight loss based on text associated with each video. Text was constructed by concatenating four fields: the video description, speech transcript, hashtags, and on-screen sticker text. Videos with no text were excluded from classification.

We defined a binary outcome to capture whether a video actively encouraged viewers to lose weight or reduce body size. Specifically, the binary variable for a video was coded as 1 if its text did any of the following: (1) encouraged or instructed the viewer to lose weight, burn fat, slim down, reduce waist size, flatten the stomach, tone the belly, or become skinnier; (2) promoted dieting, calorie restriction, fat-burning routines, or cosmetic or medical fat-removal procedures, framing these as desirable; (3) celebrated becoming thinner as a personal improvement in a manner implying that smaller is better (e.g., “I love being skinny now, I finally got rid of my belly”); or (4) sold or promoted a product primarily on the basis that it would make the viewer appear slimmer or reduce fat.

All other content was coded as not promoting weight loss (binary variable as 0). This included content that expressed shame, stigma, or trauma related to body size without promoting weight loss; body acceptance or anti-diet messaging; plus-size styling or shapewear content not directing viewers to become smaller; generic fitness content with no explicit slimming goal; and skincare, mental health, or unrelated content. Critically, mere mention of words such as “fat,” “skinny,” or “weight” was not sufficient to trigger a positive code; the text had to actively direct or encourage the viewer toward a smaller body.

Videos were processed in batches of 20. Each API call included the system prompt above and 24 labeled few-shot examples, a selection of which appears below, selected to illustrate the class boundary. The detailed rubric and few-shot examples were the primary mechanism for stabilizing outputs across runs.

Content Promoting Weight Loss (code as 1):

“double chin no more! #doublechin #doublechinreduction #jawline #chubbyface #weight-loss #hack before. after.”

“locking in my summer body #mealprep #caloriedeficit #iifym #macros make high protein buffalo chicken dip. prep fruit. meal prep for the week with me.”

“what 10k steps did to my face #walking10kaday #waterweight #walkingpad #fyp”

Content Not Promoting Weight Loss (code as 0):

“don’t wait to feel confident or happy. you deserve to feel good now! #plussizeconfidence #chunkychats [...] please, for your own sake, stop believing that the only way to be pretty is to be in a smaller body.”

“one thing we’re not gonna do is act like we should all have the same goal to be as small as possible. #wellness #fitness #motivation”

“the diet industry is counting on you to hate yourself. don’t listen to creators who are trying to make a buck off of your insecurities. #fuckdietculture #bodyneutrality”

B Appendix: Mechanism Experiment

B1 Open-Ended Explanations for Choices

To understand why participants chose to avoid each genre, we classify their open-ended explanations, restricting attention to the roughly half of avoiders who gave a substantive reason rather than citing only the monetary incentive. Among these participants, 54% of those avoiding weight-loss content invoked psychological harm, such as triggering content, eating disorders, body image, or insecurity, compared with only 4% of those avoiding cryptocurrency. Cryptocurrency avoidance was instead almost entirely a matter of distaste, with participants describing the genre as boring or irrelevant. This contrast supports our interpretation of weight-loss content as a temptation good that many users find decreases well-being, unlike cryptocurrency, which is merely disliked.

Participants who avoided weight-loss content most often described it as actively harmful to their mental health:

I can't put a price on avoiding toxic weight and diet culture.

Weight loss is an uncomfortable subject for me and once I watch a video I sometimes can't stop.

I struggle with an eating disorder, so I don't really want to risk triggering myself (for the most part).

A smaller share cited distaste:

Weight loss doesn't interest me at all.

I don't need to lose weight so I don't care to see it.

Participants who avoided cryptocurrency, by contrast, almost always cited distaste:

I'm not interested at all.

I don't really care for the topic. As opposed to the other topics, I view cryptocurrency as ... eye-roll-inducing.

I don't really understand what cryptocurrency is and I'm just not interested in it.

Some rare expressions of discomfort include:

I don't know anything about crypto, and it stresses me out.

Cryptocurrency feels similar to gambling. I would not want to feel pushed to invest based on a tiktok.

B2 Figures and Tables

Figure B1: Mechanism Experiment Design

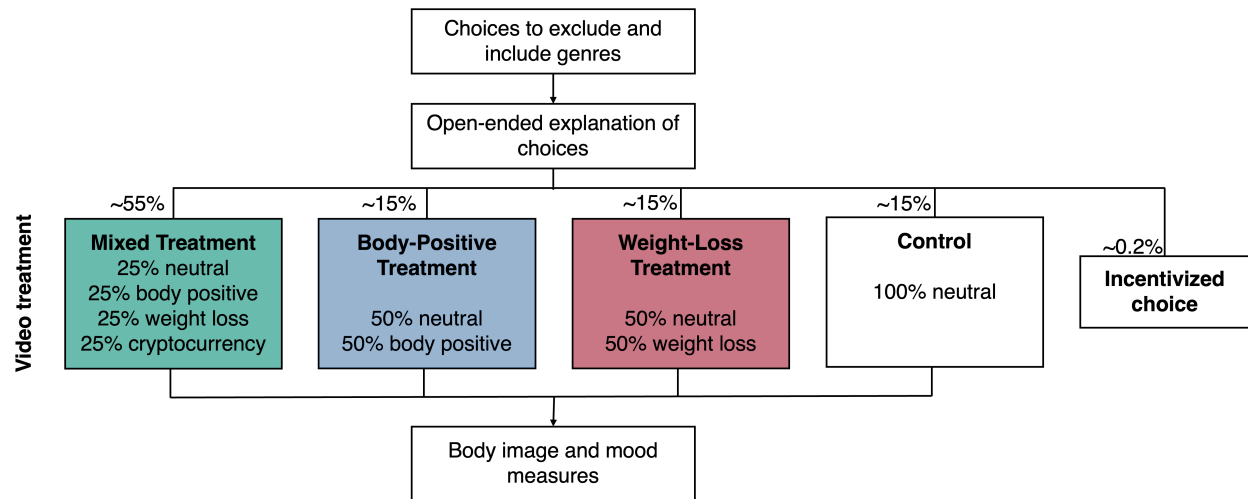
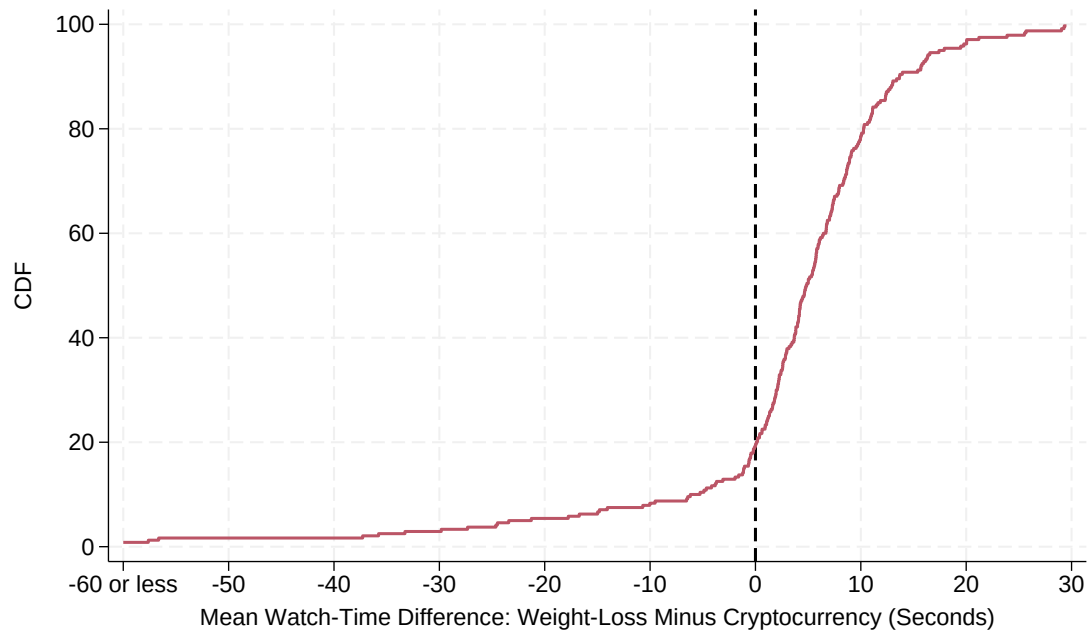


Figure B2: Weight-Loss versus Cryptocurrency Watch Time—Participant Differences



NOTE.—This figure plots the empirical cumulative distribution of each participant’s mean per-video watch time for weight-loss content minus her mean per-video watch time for cryptocurrency content. The sample comprises participants assigned to the mixed-content treatment who chose before treatment to exclude both genres. Positive values indicate longer average watch time for weight-loss content than for cryptocurrency content. Of 240 participants, 194 (80.8%) had a positive difference (two-sided exact sign test, $p < 0.001$). Two participants whose differences were below -60 seconds are grouped at the left endpoint.

Figure B3: Content Preferences and Willingness to Pay

In the video feed presented later in the study, would you like to **include or exclude videos about weight loss**? These videos promote dieting, calorie tracking, weight-loss challenges, before-and-after transformations, and other content emphasizing thinness or achieving a smaller body.

☐ I would like to **include** videos about **weight loss** in my feed

☐ I would like to **exclude** videos about **weight loss** from my feed

(a) Choice to Include or Exclude a Genre

In a previous question, you indicated:

- I would like to **exclude** videos about **weight loss** from my feed.

If your ID was selected, you will receive **\$5** at the end of the study. In each row of the table below, please indicate which option you would prefer:

- Option A:** Keep the full \$5 and **guarantee that you WILL see videos about weight loss** in your video feed.
- Option B:** Keep a smaller amount of money (as shown in that row) and **guarantee that you WILL NOT see videos about weight loss** in your video feed.

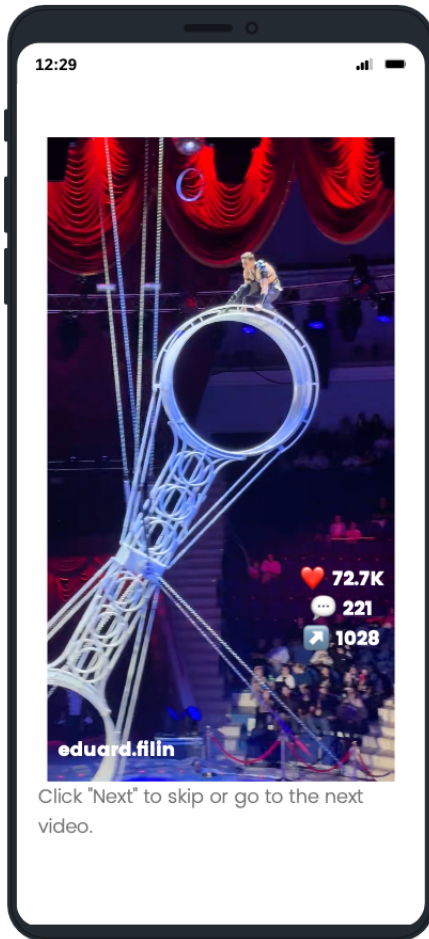
If the computer selects a particular row, **we will implement the option you chose in that row**. Therefore, you should respond based on your true preferences in each row.

	Would you rather...	
Keep \$5 and guarantee that you WILL see weight loss	☐ ☐	Keep \$0 and guarantee that you will NOT see weight loss
Keep \$5 and guarantee that you WILL see weight loss	☐ ☐	Keep \$1 and guarantee that you will NOT see weight loss
Keep \$5 and guarantee that you WILL see weight loss	☐ ☐	Keep \$2 and guarantee that you will NOT see weight loss
Keep \$5 and guarantee that you WILL see weight loss	☐ ☐	Keep \$3 and guarantee that you will NOT see weight loss
Keep \$5 and guarantee that you WILL see weight loss	☐ ☐	Keep \$4 and guarantee that you will NOT see weight loss
Keep \$5 and guarantee that you WILL see weight loss	☐ ☐	Keep \$5 and guarantee that you will NOT see weight loss

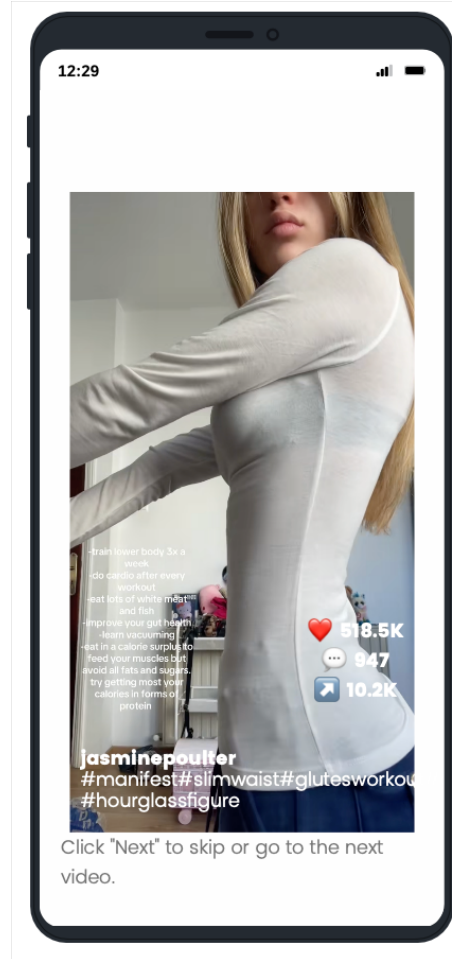
(b) Willingness to Pay to Exclude a Genre

NOTE.—Participants were first asked to make a binary choice to include or avoid a content genre, followed by an MPL to implement the BDM mechanism to elicit their willingness to pay to avoid (or include) that genre. WTP to include a genre was coded as the negative of WTP to avoid that genre.

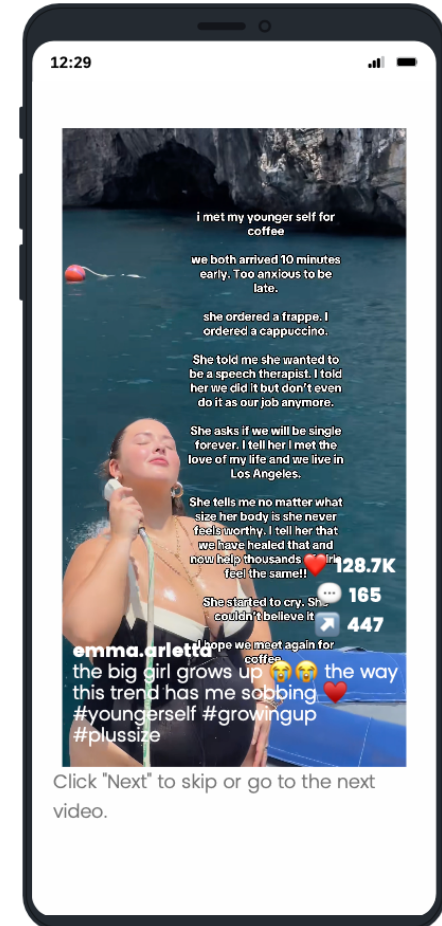
Figure B4: Mechanism Experiment—Sample Videos



(a) Neutral Video



(b) Weight-Loss Video



(c) Body-Positive Video

NOTE.—All content shown in the mechanism experiment included audio, video, and metadata (video description, influencer handle, number of likes, comments, and shares) at the time of data collection.

Table B1: Field and Mechanism Experiment Samples

	Mean		Comparison	
	Field	Mechanism	Difference	<i>p</i> -value
<i>Demographics</i>				
Female	1.00	0.96	−0.04	< 0.01
Age	22.56	23.21	0.65	< 0.01
United States	0.51	0.49	−0.02	0.46
United Kingdom	0.34	0.36	0.01	0.65
Canada	0.15	0.15	0.01	0.67
White	0.51	0.57	0.06	0.03
Black	0.14	0.10	−0.04	0.05
Asian	0.13	0.15	0.02	0.25
Mixed Race	0.14	0.12	−0.01	0.59
Hispanic	0.12	0.09	−0.03	0.14
LGBTQ	0.31	0.37	0.06	0.03
<i>Other Baseline Characteristics</i>				
Uses TikTok	1.00	0.81	−0.19	< 0.01
Daily TikTok Use Among Users (mins/day)	119.74	132.18	12.44	0.03
BMI	24.99	25.43	0.44	0.26
Observations	424	1,000		
Joint Difference <i>F</i> -test				< 0.01

NOTE.—This table compares baseline characteristics across the field- and mechanism-experiment samples. The field sample comprises participants who completed the initial scrolling period. The mechanism experiment used a new Prolific sample recruited in the United States, United Kingdom, and Canada from December 27, 2025 through January 27, 2026. Differences equal the mechanism mean minus the field mean; *p*-values are from heteroskedasticity-robust regressions. Female equals one for participants who self-identified as women. Among mechanism participants, 96.1% self-identified as women, 3.7% as nonbinary, and 0.2% as another gender. Gender is observed for 420 field participants, all of whom identified as women; four field participants have missing gender. Means use available observations; daily TikTok minutes are reported among TikTok users.

Table B2: Mechanism Experiment—Participant Characteristics and Balance

	Mean by Treatment Arm				<i>F</i> -test
	Control	Body-Positive	Weight-Loss	Mixed	<i>p</i> -value
<i>Demographics</i>					
Female	0.94	0.95	0.97	0.97	0.32
Age	23.13	23.05	23.32	23.25	0.64
United States	0.51	0.51	0.50	0.47	0.74
United Kingdom	0.32	0.33	0.35	0.38	0.49
Canada	0.17	0.16	0.15	0.15	0.95
White	0.56	0.59	0.54	0.58	0.81
Black	0.14	0.09	0.07	0.10	0.30
Asian	0.12	0.14	0.19	0.15	0.42
Mixed Race	0.13	0.09	0.14	0.13	0.47
Hispanic	0.11	0.10	0.10	0.08	0.73
LGBTQ	0.35	0.37	0.38	0.38	0.97
BMI	25.89	25.89	25.61	25.13	0.50
<i>Social Media Usage</i>					
Instagram or TikTok Use (mins/day)	109.76	121.02	113.63	109.02	0.51
Uses TikTok	0.79	0.84	0.80	0.82	0.64
Uses Instagram	0.93	0.97	0.95	0.94	0.24
Uses YouTube	0.80	0.86	0.80	0.86	0.13
Uses Facebook	0.54	0.59	0.57	0.59	0.70
Uses Snapchat	0.46	0.54	0.48	0.57	0.06
Uses Reddit	0.52	0.52	0.59	0.53	0.59
Uses X	0.35	0.36	0.33	0.38	0.70
Sees Body-Positive Content Sometimes or Often	0.59	0.55	0.57	0.60	0.72
Sees Weight-Loss Content Sometimes or Often	0.78	0.74	0.75	0.79	0.50
<i>Content Preferences</i>					
Prefers to Exclude Body-Positive Content	0.23	0.20	0.19	0.18	0.59
Prefers to Exclude Weight-Loss Content	0.40	0.46	0.44	0.48	0.37
WTP to Exclude Body-Positive Content (\$)	−0.55	−0.45	−0.66	−0.56	0.70
WTP to Exclude Weight-Loss Content (\$)	0.07	0.27	−0.13	0.12	0.22
Observations	147	164	145	544	
<i>F</i> -test	0.56	0.50	0.77	0.28	

NOTE.—This table reports participant characteristics measured before assignment to the video treatment by treatment group. The final column reports *p*-values from heteroskedasticity-robust *F*-tests of equality across the four treatment arms; for each characteristic, we regress the characteristic on treatment-arm indicators, with control omitted, and jointly test that the treatment coefficients equal zero. Female indicates self-identification as a woman; the remaining participants identified as nonbinary or another non-male gender. The *F*-test row reports *p*-values from heteroskedasticity-robust joint tests in which each treatment-arm indicator is regressed on all characteristics in the table.

Table B3: Treatment Effects on Body Image and Mood

	Body Image and Mood Index		
	(1)	(2)	(3)
Body-Positive Treatment	0.27** (0.10)	0.23 (0.15)	0.35** (0.14)
Weight-Loss Treatment	-0.23** (0.11)	-0.32** (0.16)	-0.10 (0.16)
Observations	455	198	257
R^2	0.08	0.12	0.09
Sample	Full	Exclude Weight Loss	Include Weight Loss

Notes: This table reports treatment effects on the body image and mood index in the mechanism experiment. The index was constructed by combining the Body Appreciation Scale, positive affect, and negative affect, with negative affect reversed so that higher values indicate better body image and mood. The mixed treatment arm is excluded, as pre-registered, and the neutral treatment is the omitted category. Column 1 includes the full sample. Columns 2 and 3 restrict the sample to participants who chose before treatment to exclude or include weight-loss content, respectively. All regressions control for age and BMI and include indicators for race, country, gender, and LGBTQ identity (with separate categories for Yes, No, Unsure, and Prefer not to answer). Heteroskedasticity-robust standard errors are in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table B4: Mechanism Experiment Estimates—Reweighted Sample

	Main Sample	Field-Eligible	Reweighted
<i>Panel A. Consumption: Watch Time (Seconds)</i>			
Weight-Loss Video	4.24*** (0.51)	4.35*** (0.60)	4.32*** (0.64)
Participants	240	173	173
<i>Panel B. Body Image and Mood Index: Full Sample</i>			
Body-Positive Treatment	0.27** (0.10)	0.26** (0.11)	0.30** (0.12)
Weight-Loss Treatment	−0.23** (0.11)	−0.21* (0.13)	−0.16 (0.13)
Participants	455	352	352
<i>Panel C. Body Image and Mood Index: Prefer to Exclude Weight Loss</i>			
Body-Positive Treatment	0.23 (0.15)	0.21 (0.17)	0.26 (0.18)
Weight-Loss Treatment	−0.32** (0.16)	−0.31* (0.18)	−0.28 (0.19)
Participants	198	137	137

NOTE.—Column 1 reports the main mechanism-experiment estimates. Column 2 restricts the mechanism sample to participants who identify as women and report using TikTok, matching the field experiment’s eligibility criteria. Column 3 reweights this field-eligible sample to match the field-experiment sample on age, country, race, LGBTQ identity, and BMI using normalized inverse-odds weights. Panel A compares watch time for weight-loss and cryptocurrency videos among participants in the mixed treatment who prefer to exclude both genres; specifications include participant fixed effects and cluster standard errors by participant. Panels B and C compare the body-positive and weight-loss treatments with the neutral treatment and control for age and BMI, with indicators for race, country, and the original gender and LGBTQ response categories; heteroskedasticity-robust standard errors are reported. Panel C restricts to participants who preferred to exclude weight-loss content before treatment. Standard errors treat the estimated weights as fixed.

Table B5: Treatment Effects on Body Image and Mood—Including Mixed Treatment

VARIABLES	(1) Body Image and Mood Index	(2) Body Appreciation	(3) Positive Affect	(4) Negative Affect
Body-Positive Treatment	0.27*** (0.10)	0.23** (0.10)	0.35*** (0.11)	0.01 (0.09)
Weight-Loss Treatment	−0.21* (0.11)	−0.17 (0.11)	0.26** (0.11)	0.54*** (0.12)
Mixed Treatment	−0.01 (0.08)	−0.08 (0.08)	0.31*** (0.09)	0.29*** (0.08)
Observations	999	999	999	999
R^2	0.07	0.10	0.07	0.06

NOTE.— This table reports OLS regression estimates of well-being outcomes on treatment arm, as specified in Equation (4). The body image and mood index was constructed by combining measures of body appreciation, positive affect, and negative affect as per [Anderson \(2008\)](#). All treatment arms are included. Column 1 reports the index; columns 2–4 report body appreciation, positive affect, and negative affect, respectively. All regressions control for age and BMI and include indicators for race, country, gender, and LGBTQ identity. Treatment effects are reported in standard deviations. Standard errors are robust.

Table B6: Treatment Effects on Reported Content Characteristics

VARIABLES	(1) Uplifting	(2) Upsetting	(3) Informative	(4) Interesting	(5) Funny	(6) Inspiring	(7) Inspiring
Body-Positive Treatment	0.86*** (0.10)	−0.13 (0.09)	0.22** (0.11)	0.07 (0.11)	−0.34*** (0.11)	0.85*** (0.10)	0.86*** (0.15)
Weight-Loss Treatment	0.13 (0.11)	0.73*** (0.12)	0.37*** (0.11)	−0.15 (0.12)	−0.48*** (0.12)	0.60*** (0.10)	0.31** (0.15)
Observations	455	455	455	455	455	455	198
R^2	0.21	0.16	0.08	0.09	0.11	0.20	0.26

NOTE.— This table reports OLS regression estimates of how well certain adjectives described the feed, by treatment arm, as specified in Equation (4). For each adjective X , participants were asked to rate, on a scale from 0 to 100, how strongly they agreed with the statement “The videos were X .” Regressions were estimated on the neutral, weight-loss, and body-positive treatment arms, as specified in our pre-analysis plan. Column 7 restricts to participants who preferred to exclude weight-loss content. All regressions control for age and BMI and include indicators for race, country, gender, and LGBTQ identity. Treatment effects are reported in standard deviations. Standard errors are robust.

Table B7: Treatment Effects on Reported Feed Shares

VARIABLES	(1) Body Pos.	(2) Body Neg.	(3) Thin	(4) Plus-Sized	(5) White	(6) Black	(7) Asian	(8) Hispanic	(9) LGBTQ	(10) Friend
Body-Positive Treatment	1.69*** (0.07)	-0.02 (0.05)	-1.36*** (0.09)	1.69*** (0.07)	0.06 (0.12)	-0.27** (0.12)	-0.09 (0.12)	0.24** (0.12)	0.43*** (0.12)	0.13 (0.10)
Weight-Loss Treatment	0.66*** (0.10)	1.38*** (0.10)	-0.06 (0.10)	0.85*** (0.09)	0.33*** (0.11)	-0.27** (0.12)	-0.30** (0.12)	-0.23** (0.11)	-0.22** (0.10)	0.12 (0.11)
Observations	455	455	455	455	455	455	455	455	455	455
R^2	0.54	0.45	0.43	0.50	0.10	0.05	0.03	0.05	0.10	0.02

NOTE.— This table reports OLS regression estimates of reported feed share by treatment arm, as specified in Equation (4). For each listed characteristic X , participants were asked to answer on a scale from 0 to 100, “What share of the videos featured influencers who are X ?” Regressions were estimated on the neutral, weight-loss, and body-positive treatment arms, as specified in our pre-analysis plan. All regressions control for age and BMI and include indicators for race, country, gender, and LGBTQ identity. Treatment effects are reported in standard deviations. Standard errors are robust.

Table B8: Treatment Effects on Body Image and Mood—Components and Heterogeneity

	(1) Body Image	(2) Positive Mood	(3) Negative Mood
<i>Panel A. Full Sample</i>			
Body-Positive Treatment	0.22** (0.11)	0.34*** (0.12)	0.00 (0.10)
Weight-Loss Treatment	-0.20* (0.11)	0.26** (0.11)	0.56*** (0.12)
Observations	455	455	455
R^2	0.07	0.08	0.09
<i>Panel B. Exclude Weight Loss</i>			
Body-Positive Treatment	0.03 (0.17)	0.29* (0.16)	-0.12 (0.15)
Weight-Loss Treatment	-0.14 (0.17)	0.12 (0.16)	0.64*** (0.20)
Observations	198	198	198
R^2	0.10	0.06	0.16
<i>Panel C. Include Weight Loss</i>			
Body-Positive Treatment	0.38*** (0.14)	0.44*** (0.16)	0.05 (0.13)
Weight-Loss Treatment	-0.23 (0.16)	0.42*** (0.15)	0.45*** (0.16)
Observations	257	257	257
R^2	0.11	0.11	0.06

Notes: This table reports treatment effects on the components of the body image and mood index in the mechanism experiment. Body image is measured using the Body Appreciation Scale, while positive and negative mood are measured using the Positive and Negative Affect Schedule. All outcomes are standardized using the full sample; higher values indicate better body image, more positive mood, and more negative mood, respectively. The mixed treatment arm is excluded, as pre-registered. Panel A includes the full sample. Panels B and C restrict the sample to participants who chose before treatment to exclude or include weight-loss content, respectively. All regressions control for age and BMI and include indicators for race, country, gender, and LGBTQ identity (with separate categories for Yes, No, Unsure, and Prefer not to answer). Heteroskedasticity-robust standard errors are in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

C Appendix: Model

C1 Proof of Proposition 1

Proposition 1. *A pure-strategy perfect Bayesian equilibrium exists.*

Proof. The first-order condition and the nonnegativity constraint give consumption of the c -th genre- k video as

$$\sigma_k^*(c; \theta) = \frac{\max\{0, u_k + \tau \mathbf{1}\{k = P\}\}}{\gamma c}.$$

This is the unique best response because the user's short-run payoff is strictly concave in consumption. In particular, $x_1 = \max\{0, u_H\}/\gamma$. If $u_H > 0$, the platform observes positive consumption and learns the user's taste. If $u_H \leq 0$, it observes zero consumption and learns only that her taste is nonpositive.

The platform's beliefs $\mu^*(\cdot | x_1)$ are obtained by conditioning the prior f on consumption under this rule. The model's moment conditions ensure finite posterior expected engagement for almost every signal. At any remaining signals, let the platform retain the prior f . For the observed signal x_1 , define

$$e_k(x_1) = \mathbb{E}_{\mu^*(\cdot | x_1)}[\max\{0, u_k + \tau \mathbf{1}\{k = P\}\}], \quad e_k^\varnothing = \mathbb{E}[\max\{0, u_k + \tau \mathbf{1}\{k = P\}\}].$$

Expected engagement with the c -th genre- k video is then $e_k(x_1)/(\gamma c)$ after observing x_1 , and $e_k^\varnothing/(\gamma c)$ under the prior.

Given its posterior, the platform chooses the remaining feed to maximize

$$\sum_{n=2}^{\infty} \Pr(N \geq n) \mathbb{E}_{\mu^*(\cdot | x_1)}[\sigma_{r_n}^*(c_{r_n}(n); \theta)] = \sum_{n=2}^{\infty} \kappa^{n-1} \frac{e_{r_n}(x_1)}{\gamma c_{r_n}(n)}.$$

This objective is finite because the engagement values are finite and $\sum_{n=2}^{\infty} \kappa^{n-1} < \infty$. Since $0 < \kappa < 1$, the probability weights decrease with the slot number. Thus, the platform maximizes the weighted sum by ordering the remaining videos in decreasing order of expected engagement, breaking ties according to the order H, P, O . This ordering respects the exposure order within each genre because $e_k(x_1)/(\gamma c)$ is nonincreasing in c .

This ranking defines a pure strategy ρ^* . The user's consumption strategy is myopically optimal, the platform's ranking is optimal given its beliefs, and those beliefs follow Bayes' rule wherever determined. Thus $(\sigma^*, \rho^*, \mu^*)$ is a pure-strategy perfect Bayesian equilibrium. $\square$

C2 Proof of Proposition 2

Proposition 2. *For a given user, genre P is parasitic to genre H for all sufficiently large N if and only if:*

1. (High taste for genre H) u_H is sufficiently high; and
2. (Strong temptation for genre P) $\tau > |u_P|$.

We first prove two supporting lemmas.

Statements about users are understood almost surely under f , and the sufficient session length may depend on the user. We therefore consider signals for which Bayesian conditioning gives finite posterior expected engagement.

Lemma 1. *There exists a finite cutoff $\bar{u}_H \geq 0$ such that $c_P(N) > c_P^\emptyset(N)$ for all sufficiently large N if and only if $u_H > \bar{u}_H$. If $u_H < \bar{u}_H$, then $c_P(N) < c_P^\emptyset(N)$ for all sufficiently large N .*

Proof. We work towards bounding $c_k(N)$ for a genre k . Temporarily remove the requirement that the first video be from genre H and let $\tilde{c}_k(N)$ be the number of genre- k videos among the first N slots. Define a_N to be expected engagement with the video in slot N , where $a_N > 0$ since genre O has infinitely many videos with positive expected engagement. As described in the proof of Proposition 1, every video with expected engagement strictly above a_N is shown in the first N slots, and none with expected engagement below a_N is shown.

If $\tilde{c}_k(N) > 0$, the last genre- k video shown in the first N slots is the $\tilde{c}_k(N)$ -th video of that genre. Since it is shown, its expected engagement is at least a_N , that is, $\frac{e_k(x_1)}{\gamma \tilde{c}_k(N)} \geq a_N$, which can be rearranged to give

$$\tilde{c}_k(N) \leq \frac{e_k(x_1)}{\gamma a_N}.$$

Similarly, the $(\tilde{c}_k(N) + 1)$ -th video of genre k is not shown in the first N slots. Its expected engagement must therefore be at most a_N , that is $\frac{e_k(x_1)}{\gamma(\tilde{c}_k(N)+1)} \leq a_N$, which can be rearranged to give

$$\tilde{c}_k(N) \geq \frac{e_k(x_1)}{\gamma a_N} - 1.$$

Combining the two inequalities, we obtain

$$\frac{e_k(x_1)}{\gamma a_N} - 1 \leq \tilde{c}_k(N) \leq \frac{e_k(x_1)}{\gamma a_N}. \quad (7)$$

If $e_k(x_1) = 0$, its count is zero and the same bounds hold.

To replace a_N , we sum $\tilde{c}_k(N)$ across k , note that $e_O(x_1) = e_O^\emptyset$ by Assumption 1, and use $\sum_k \tilde{c}_k(N) = N$ to obtain

$$N \leq \frac{e_H(x_1) + e_P(x_1) + e_O^\emptyset}{\gamma a_N} \leq N + 3.$$

Multiplying each part of this inequality by $e_k(x_1)/(e_H(x_1) + e_P(x_1) + e_O^\emptyset)$ gives

$$\frac{e_k(x_1)}{e_H(x_1) + e_P(x_1) + e_O^\emptyset} N \leq \frac{e_k(x_1)}{\gamma a_N} \leq \frac{e_k(x_1)}{e_H(x_1) + e_P(x_1) + e_O^\emptyset} (N + 3).$$

Using the left-hand inequality above and the lower bound in (7), we obtain

$$\tilde{c}_k(N) \geq \frac{e_k(x_1)}{\gamma a_N} - 1 \geq \frac{e_k(x_1)}{e_H(x_1) + e_P(x_1) + e_O^\emptyset} N - 1.$$

Similarly, using the right-hand inequality above and the upper bound in (7), we obtain

$$\tilde{c}_k(N) \leq \frac{e_k(x_1)}{\gamma a_N} \leq \frac{e_k(x_1)}{e_H(x_1) + e_P(x_1) + e_O^\emptyset} (N + 3) \leq \frac{e_k(x_1)}{e_H(x_1) + e_P(x_1) + e_O^\emptyset} N + 3,$$

where the final inequality is true because $e_k(x_1)/(e_H(x_1) + e_P(x_1) + e_O^\emptyset) \leq 1$. Thus we have

$$\frac{e_k(x_1)}{e_H(x_1) + e_P(x_1) + e_O^\emptyset} N - 1 \leq \tilde{c}_k(N) \leq \frac{e_k(x_1)}{e_H(x_1) + e_P(x_1) + e_O^\emptyset} N + 3.$$

Now restore the initial genre- H video by placing it first and leaving all other videos in their previous order. If there was already a video of genre H in the first N slots, this does not change the genre counts and $\tilde{c}_k(N) = c_k(N)$ for all k . Otherwise, the initial genre- H video displaces one video, and each genre's count changes by at most one, that is,

$$\frac{e_k(x_1)}{e_H(x_1) + e_P(x_1) + e_O^\emptyset} N - 2 \leq c_k(N) \leq \frac{e_k(x_1)}{e_H(x_1) + e_P(x_1) + e_O^\emptyset} N + 4.$$

Using the same argument as above, we can obtain analogous bounds for $c_k^\emptyset(N)$. Using these bounds and subtracting $c_k^\emptyset(N)$ from $c_k(N)$ for $k = P$ gives

$$\begin{aligned} \left(\frac{e_P(x_1)}{e_H(x_1) + e_P(x_1) + e_O^\emptyset} - \frac{e_P^\emptyset}{e_H^\emptyset + e_P^\emptyset + e_O^\emptyset} \right) N - 6 \leq c_P(N) - c_P^\emptyset(N) \leq \\ \left(\frac{e_P(x_1)}{e_H(x_1) + e_P(x_1) + e_O^\emptyset} - \frac{e_P^\emptyset}{e_H^\emptyset + e_P^\emptyset + e_O^\emptyset} \right) N + 6. \end{aligned} \quad (8)$$

Whenever the share difference is nonzero, it determines the sign of $c_P(N) - c_P^\emptyset(N)$ for sufficiently large N . With some simplification, it can be shown that this share difference has the same sign as

$$\frac{e_P(x_1)}{e_H(x_1) + e_O^\emptyset} - \frac{e_P^\emptyset}{e_H^\emptyset + e_O^\emptyset}. \quad (9)$$

It is convenient to characterize the first term above, $\frac{e_P(x_1)}{e_H(x_1) + e_O^\emptyset}$. Suppose $x_1 > 0$. It follows from the consumer's consumption strategy that $u_H = \gamma x_1$, so $e_H(x_1) = \gamma x_1$ and $e_P(x_1) = \mathbb{E}[\max\{0, u_P + \tau\} \mid u_H = \gamma x_1]$. By Assumption 2, $\frac{e_P(x_1)}{e_H(x_1) + e_O^\emptyset}$ is continuous and strictly increasing. If $\Pr(u_H \leq 0) > 0$, zero consumption induces a pooled posterior over $u_H \leq 0$. Thus, $e_H(0) = 0$ and $e_P(0) = \mathbb{E}[\max\{0, u_P + \tau\} \mid u_H \leq 0]$. For any $\varepsilon > 0$ and $u \leq 0$, Assumption 2 gives $\frac{\mathbb{E}[\max\{0, u_P + \tau\} \mid u_H = u]}{e_O^\emptyset} < \frac{\mathbb{E}[\max\{0, u_P + \tau\} \mid u_H = \gamma\varepsilon]}{\gamma\varepsilon + e_O^\emptyset}$. Averaging the left-hand side over $u_H \leq 0$ yields $\frac{e_P(0)}{e_H(0) + e_O^\emptyset} < \frac{e_P(\varepsilon)}{e_H(\varepsilon) + e_O^\emptyset}$. Thus, $\frac{e_P(x_1)}{e_H(x_1) + e_O^\emptyset}$ is strictly increasing in x_1 and continuous for all $x_1 > 0$. If instead $\Pr(u_H \leq 0) = 0$,

initial consumption is positive almost surely; no pooling at zero is needed, and the zero-signal cases below can be omitted.

We next show that (9) takes both positive and negative values. To see this, let $X_1 = \max\{0, u_H\}/\gamma$ be a random variable and note that (9) can be expressed

$$\frac{e_P(X_1)(e_H^\emptyset + e_O^\emptyset) - e_P^\emptyset(e_H(X_1) + e_O^\emptyset)}{(e_H(X_1) + e_O^\emptyset)(e_H^\emptyset + e_O^\emptyset)},$$

which is 0 when the numerator is 0. By the law of iterated expectations,

$$\mathbb{E}[e_P(X_1)(e_H^\emptyset + e_O^\emptyset) - e_P^\emptyset(e_H(X_1) + e_O^\emptyset)] = e_P^\emptyset(e_H^\emptyset + e_O^\emptyset) - e_P^\emptyset(e_H^\emptyset + e_O^\emptyset) = 0.$$

Thus, the numerator has expectation zero. The numerator cannot be zero for every user since (9) is strictly increasing in X_1 and X_1 is nonconstant; it must therefore be positive for some values of x_1 and negative for others. Likewise, (9) takes both positive and negative values.

If (9) equals zero at some $\bar{x} > 0$, define $\bar{u}_H = \gamma\bar{x}$. Strict monotonicity ensures that $\bar{x}$ is unique. If (9) never equals zero for $x_1 > 0$, it must be strictly positive for every $x_1 > 0$. We have already shown that it is positive at some positive value of x_1 . If it were negative at another positive value, continuity would imply that it equals zero somewhere between them, a contradiction. In this case, define $\bar{u}_H = 0$. Either way, $\bar{u}_H \geq 0$. We now consider three cases.

Case 1: $u_H > \bar{u}_H$. By strict monotonicity of (9) for $x_1 > 0$ (and thus $u_H > 0$), (9) is strictly positive. Thus, $c_P(N) - c_P^\emptyset(N) > 0$ for sufficiently large N .

Case 2: $u_H < \bar{u}_H$. If $0 < u_H < \bar{u}_H$, then $0 < x_1 = u_H/\gamma < \bar{x}$, so (9) is strictly negative by strict monotonicity. If $u_H \leq 0$, then $x_1 = 0$. Since (9) is strictly increasing in x_1 , it takes its smallest value at $x_1 = 0$; furthermore, since we showed that (9) is negative for some consumption levels, it is strictly negative at $x_1 = 0$. Thus, in either case, $c_P(N) - c_P^\emptyset(N) < 0$ for sufficiently large N .

Case 3: $u_H = \bar{u}_H$. Suppose $\bar{u}_H = 0$. Then $x_1 = 0$. Using the same argument as in case 2, $c_P(N) < c_P^\emptyset(N)$ for sufficiently large N .

Suppose $\bar{u}_H > 0$. Then $x_1 = \bar{u}_H/\gamma$, and (9) equals 0 at x_1 . Equation (8) now gives only $-6 \leq c_P(N) - c_P^\emptyset(N) \leq 6$. We therefore show separately that $c_P(N) = c_P^\emptyset(N)$ at arbitrarily large N .

Let $q = \frac{e_H(x_1) + e_P(x_1) + e_O^\emptyset}{e_P(x_1)} = \frac{e_H^\emptyset + e_P^\emptyset + e_O^\emptyset}{e_P^\emptyset}$. Temporarily remove the requirement that the first item be from genre H (note that this is wlog since $x_1 > 0$. Thus, for sufficiently large N , both feeds will show at least one video from genre H). For either feed, write e_H, e_P, e_O for its engagement values. Under the tie order H, P, O , the c -th genre- P item occupies slot

$$n(c) = c + \left\lfloor \frac{ce_H}{e_P} \right\rfloor + \left\lceil \frac{ce_O}{e_P} \right\rceil - 1,$$

where the second term is the number of genre- H items with weakly higher engagement, and the

third and fourth terms represent the number of genre- O items with strictly higher engagement.

It is useful to bound $n(c)$. Using the properties of the floor and ceiling functions, $ce_H/e_P - 1 < \lfloor ce_H/e_P \rfloor \leq ce_H/e_P$ and $ce_O/e_P - 1 \leq \lceil ce_O/e_P \rceil - 1 < ce_O/e_P$. Adding these inequalities and c gives $c + ce_H/e_P + ce_O/e_P - 2 < n(c) < c + ce_H/e_P + ce_O/e_P$. Since $q = 1 + e_H/e_P + e_O/e_P$, this becomes $qc - 2 < n(c) < qc$. For the upper bound, $n(c) < qc < \lfloor qc \rfloor + 1$. Because $n(c)$ is an integer, it must therefore satisfy $n(c) \leq \lfloor qc \rfloor$. For the lower bound, $\lfloor qc \rfloor - 2 \leq qc - 2 < n(c)$. An integer strictly greater than $\lfloor qc \rfloor - 2$ must be at least $\lfloor qc \rfloor - 1$. Thus, $\lfloor qc \rfloor - 1 \leq n(c) \leq \lfloor qc \rfloor$, and the same bounds hold in both feeds because they have the same q .

Choose $N = \lfloor qc \rfloor$ with $\lfloor q(c+1) \rfloor \geq N + 2$. For such N , in both feeds, the c -th genre- P item has appeared by slot N , but the next has not. To see this, our upper bound gives $n(c) \leq \lfloor qc \rfloor = N$, so the c -th item has appeared. Applying the lower bound to the next item gives $n(c+1) \geq \lfloor q(c+1) \rfloor - 1$. By our choice of N , the right-hand side of the inequality is at least $(N+2) - 1 = N+1$, so the next item has not appeared by slot N . Both feeds therefore contain exactly c genre- P items at session length N .

It remains to show that such N exist, that is $\lfloor q(c+1) \rfloor \geq \lfloor qc \rfloor + 2$, at arbitrarily large integers c . Since c is an integer, we can write $\lfloor qc \rfloor = c + \lfloor (q-1)c \rfloor$. Applying the same identity at $c+1$ and subtracting the expression at c , we obtain

$$\begin{aligned} \lfloor q(c+1) \rfloor - \lfloor qc \rfloor &= \lfloor (c+1) + \lfloor (q-1)(c+1) \rfloor \rfloor - \lfloor c + \lfloor (q-1)c \rfloor \rfloor \\ &= 1 + \lfloor (q-1)(c+1) \rfloor - \lfloor (q-1)c \rfloor. \end{aligned}$$

Thus, the difference is at least two as desired whenever $\lfloor (q-1)(c+1) \rfloor - \lfloor (q-1)c \rfloor \geq 1$. This occurs at arbitrarily large integers c . To see this, fix any positive integer C . Since $q > 1$, $\lfloor (q-1)c \rfloor$ grows without bound, so there is a smallest integer $D > C$ such that $\lfloor (q-1)D \rfloor > \lfloor (q-1)C \rfloor$. Because the sequence is nondecreasing and D is the first such integer, $\lfloor (q-1)(D-1) \rfloor = \lfloor (q-1)C \rfloor$. The value at D therefore exceeds the value at $D-1$, and since both are integers,

$$\lfloor (q-1)D \rfloor - \lfloor (q-1)(D-1) \rfloor \geq 1.$$

Taking $c = D - 1$ gives the required inequality. Such choices of c , and hence session lengths $N = \lfloor qc \rfloor$, are arbitrarily large.

Hence $c_P(N) = c_P^\emptyset(N)$ at arbitrarily large N .

Together, the three cases establish that $c_P(N) > c_P^\emptyset(N)$ for sufficiently large N if and only if $u_H > \bar{u}_H$, and that $c_P(N) < c_P^\emptyset(N)$ for all sufficiently large N when $u_H < \bar{u}_H$. $\square$

Lemma 2. *For every feed r satisfying $c_P(N) > 0$, $V^{\text{block}}(r) > V^{\text{open}}(r)$ if and only if $\tau > |u_P|$.*

Proof. Blocking genre P does not change the user's consumption of genres H and O . It is therefore enough to compare her long-run payoff from genre P under the two conditions.

If $u_P + \tau \leq 0$, the user consumes no genre- P content, so $V^{block}(r) = V^{open}(r)$.

If $u_P + \tau > 0$, her consumption of the c th genre- P video is

$$\frac{u_P + \tau}{\gamma c}.$$

The contribution of this video to her long-run payoff is

$$u_P \frac{u_P + \tau}{\gamma c} - \frac{\gamma}{2} c \left(\frac{u_P + \tau}{\gamma c} \right)^2 = \frac{u_P^2 - \tau^2}{2\gamma c}.$$

Consequently,

$$V^{block}(r) - V^{open}(r) = \frac{\tau^2 - u_P^2}{2\gamma} \sum_{c=1}^{c_P(N)} \frac{1}{c}.$$

Because $c_P(N) > 0$, the sum is positive. Hence blocking strictly raises long-run payoff if and only if $\tau^2 > u_P^2$, or equivalently $\tau > |u_P|$. This inequality also implies $u_P + \tau > 0$, so it is consistent with positive consumption. $\square$

Proof. Suppose first that $u_H > \bar{u}_H$ and $\tau > |u_P|$, where $\bar{u}_H$ is the cutoff in Lemma 1. The first inequality implies $u_H > 0$, so consumption of the initial genre- H video is

$$x_1 = \frac{u_H}{\gamma} > \frac{\bar{u}_H}{\gamma}.$$

Lemma 1 then gives, for all sufficiently large N ,

$$c_P(N) > c_P^\emptyset(N).$$

This implies $c_P(N) > 0$. Lemma 2 and $\tau > |u_P|$ therefore give $V^{block}(r) > V^{open}(r)$. Both criteria in the definition of parasitic content hold for all sufficiently large N .

Conversely, suppose $u_H < \bar{u}_H$. By Lemma 1, $c_P(N) < c_P^\emptyset(N)$ for all sufficiently large N , so the **host-parasite criterion** fails. If $u_H = \bar{u}_H > 0$, the equality case in the proof of that lemma gives $c_P(N) = c_P^\emptyset(N)$ at arbitrarily large N . If $u_H = \bar{u}_H = 0$, it gives $c_P(N) < c_P^\emptyset(N)$ for all sufficiently large N . Thus, at the cutoff, parasitism cannot hold for all sufficiently large N .

Finally, suppose $u_H > \bar{u}_H$ but $\tau \leq |u_P|$. For sufficiently large N , Lemma 1 gives $c_P(N) > 0$, and Lemma 2 implies $V^{block}(r) \leq V^{open}(r)$, so the **well-being criterion** fails. $\square$

C3 Proofs of the Implications

Implication 1 (Consumption). *If genre P is parasitic to genre H for a user sophisticated about her temptation, she prefers to block genre P but consumes a positive amount when it is served.*

Proof. By the definition of parasitic content, $c_P(N) > c_P^\emptyset(N)$ and $V^{block}(r) > V^{open}(r)$. Sophistication means that the user anticipates her temptation-driven consumption and the resulting long-run payoff, so she prefers to block genre P ex ante.

Since $c_P(N) > 0$, Lemma 2 implies $\tau > |u_P|$, and hence $u_P + \tau > 0$. If the user is served the c th genre- P video, her short-run self consumes

$$\sigma_P^*(c; \theta) = \frac{u_P + \tau}{\gamma c} > 0.$$

□

Implication 2 (Average Effect of Personalization on Well-Being). *Let $p > 0$ be the share of users for whom genre P is parasitic to genre H , and let ℓ be their average long-run well-being loss from the additional genre- P videos induced by algorithmic personalization. If $p\ell$ is sufficiently large, algorithmic personalization lowers average long-run well-being:*

$$\mathbb{E}[V^{open}(r) - V^{open}(r^\emptyset)] < 0.$$

Proof. We wish to show that

$$\mathbb{E}[V^{open}(r) - V^{open}(r^\emptyset)] < 0$$

when $p\ell$ is sufficiently high. We first write this payoff difference in terms of the videos supplied in each feed. First fix the session length N and take expectations over user types; p and ℓ then refer to users at that session length.

For $k = H, O$, optimal consumption gives a long-run payoff of $\max\{0, u_k\}^2/(2\gamma c)$ from the c th genre- k video. For genre P , the corresponding payoff is $(u_P^2 - \tau^2)/(2\gamma c)$ when $u_P + \tau > 0$ and zero otherwise, as shown in Lemma 2. Adding these payoffs across the videos in each feed gives

$$\begin{aligned} V^{open}(r) - V^{open}(r^\emptyset) &= \frac{\max\{0, u_H\}^2}{2\gamma} \left(\sum_{c=1}^{c_H(N)} \frac{1}{c} - \sum_{c=1}^{c_H^\emptyset(N)} \frac{1}{c} \right) \\ &\quad + \mathbb{1}\{u_P + \tau > 0\} \frac{u_P^2 - \tau^2}{2\gamma} \left(\sum_{c=1}^{c_P(N)} \frac{1}{c} - \sum_{c=1}^{c_P^\emptyset(N)} \frac{1}{c} \right) \\ &\quad + \frac{\max\{0, u_O\}^2}{2\gamma} \left(\sum_{c=1}^{c_O(N)} \frac{1}{c} - \sum_{c=1}^{c_O^\emptyset(N)} \frac{1}{c} \right). \end{aligned}$$

A sum is zero when its upper limit is zero.

By Definition 1 and Lemma 2, the set of users experiencing parasitic content at session length N is

$$\mathcal{P}_N = \{\theta : c_P(N) > c_P^\emptyset(N), \tau > |u_P|\}.$$

This definition uses the actual feed counts and does not require a common sufficiently large session

length for all user types. For these users, $u_P + \tau > 0$ and $c_P(N) > c_P^\emptyset(N)$. The first $c_P^\emptyset(N)$ terms in the genre- P sums cancel, leaving only the additional videos. Their contribution to the population-average payoff difference is therefore

$$\mathbb{E} \left[\mathbb{1}\{\theta \in \mathcal{P}_N\} \frac{u_P^2 - \tau^2}{2\gamma} \sum_{c=c_P^\emptyset(N)+1}^{c_P(N)} \frac{1}{c} \right] = -p\ell,$$

where $p = \Pr(\theta \in \mathcal{P}_N) > 0$ and ℓ is the conditional mean loss from these additional videos, as in the implication.

Taking expectations in the payoff difference above and separating the genre- P contribution of users in $\mathcal{P}_N$ gives

$$\begin{aligned} & \mathbb{E}[V^{\text{open}}(r) - V^{\text{open}}(r^\emptyset)] \\ &= \mathbb{E} \left[\frac{\max\{0, u_H\}^2}{2\gamma} \left(\sum_{c=1}^{c_H(N)} \frac{1}{c} - \sum_{c=1}^{c_H^\emptyset(N)} \frac{1}{c} \right) \right. \\ & \quad + \mathbb{1}\{\theta \notin \mathcal{P}_N, u_P + \tau > 0\} \frac{u_P^2 - \tau^2}{2\gamma} \left(\sum_{c=1}^{c_P(N)} \frac{1}{c} - \sum_{c=1}^{c_P^\emptyset(N)} \frac{1}{c} \right) \\ & \quad \left. + \frac{\max\{0, u_O\}^2}{2\gamma} \left(\sum_{c=1}^{c_O(N)} \frac{1}{c} - \sum_{c=1}^{c_O^\emptyset(N)} \frac{1}{c} \right) \right] - p\ell. \end{aligned}$$

Denote the expectation on the right-hand side by B_N . It includes the payoff changes from genres H and O for all users and from genre P for users outside $\mathcal{P}_N$.

The residual B_N can have either sign. It is finite by the second-moment assumptions and the finite harmonic sums. Therefore, if $p\ell > B_N$, then $\mathbb{E}[V^{\text{open}}(r) - V^{\text{open}}(r^\emptyset)] < 0$.

The same decomposition applies when averaging over the random session length: use the event $\{\theta \in \mathcal{P}_N\}$ jointly over (θ, N) , and define p and ℓ for that joint population. Finiteness follows from independence of N and θ and the geometric distribution of N : writing $h_N = \sum_{c=1}^N 1/c$, the absolute payoff of either feed is bounded by

$$\frac{h_N}{2\gamma} (\max\{0, u_H\}^2 + \max\{0, u_O\}^2 + u_P^2 + \tau^2),$$

which has finite expectation. Thus the population-average payoff difference is again a finite residual minus $p\ell$, and is negative when $p\ell$ exceeds that residual. $\square$

Implication 3 (Revenue). *Choose x_1 such that $c_P(N) > c_P^\emptyset(N)$. For $c \in \{c_P^\emptyset(N) + 1, \dots, c_P(N)\}$, the c -th genre- P video generates weakly higher expected engagement and thus revenue than the next*

video from genres H or O that could be served instead:

$$\mathbb{E}_{\mu^*(\cdot|x_1)} [\sigma_P^*(c; \theta)] \geq \mathbb{E}_{\mu^*(\cdot|x_1)} [\sigma_k^*(c_k(N) + 1; \theta)], \quad k = H, O.$$

Proof. As shown in the proof of Proposition 1, the platform ranks the remaining videos by weakly decreasing posterior expected engagement. Every additional genre- P video indexed by $c \in \{c_P^\emptyset(N) + 1, \dots, c_P(N)\}$ is shown by slot N , whereas the next genre- k video, indexed by $c_k(N) + 1$, is not. Consequently,

$$\frac{e_P(x_1)}{\gamma c} \geq \frac{e_k(x_1)}{\gamma(c_k(N) + 1)}, \quad k = H, O,$$

which is the claimed comparison. Expected revenue is proportional to expected consumption by assumption, so the same ordering applies to revenue. $\square$