Exhibit 43

IN THE UNITED STATES DISTRICT COURT
FOR THE NORTHERN DISTRICT OF CALIFORNIA

PEOPLE OF THE STATE OF CALIFORNIA, et al., Plaintiffs, v. *META PLATFORMS, INC, Instagram, LLC, Meta Payments, Inc., Meta Platforms Technologies, LLC.,* Defendants IN RE: SOCIAL MEDIA ADOLESCENT ADDICTION/PERSONAL INJURY PRODUCTS LIABILITY LITIGATION THIS DOCUMENT RELATES TO: 4:23-cv-05448	MDL No. 3047 Case No. 4:23-cv-05448-YGR Judge: Hon. Yvonne Gonzalez Rogers Magistrate Judge: Hon. Peter H. Kang

TRIAL REPORT OF ARVIND NARAYANAN
May 16, 2025

I. Executive Summary of Opinions

1. The following is a list of the expert opinions that I hold, am prepared to testify about, and which I explain in detail in this report. I hold these opinions to a reasonable degree of scientific certainty:

Opinion 1: Meta's recommendation algorithms are optimized to maximize user engagement which, in turn, drives advertising revenue.

Opinion 2: Meta's product testing practices encourage addictive design by prioritizing short-term engagement over all other considerations.

Opinion 3: The design of Meta's recommendation algorithms promotes problematic usage by programmatically exploring and then exploiting users' unstated proclivities.

Opinion 4: By optimizing for particular behaviors, the design of Meta's recommendation algorithms can cause a user's feed to become more homogenous over time and lead to rabbit holing.

Opinion 5: Meta has not adequately disclosed to users impacts that may result from engagement with its recommendation algorithms, such as users spending more time than intended on a platform, users being directed by algorithms down rabbit holes, and users' interests being shaped through feedback loops.

Opinion 6: Meta has not disclosed to users sufficient information to allow them to adequately control what their feed shows them, such as what information the platforms' recommendation algorithms infer about users, how users' actions impact the posts the algorithms show them, and why the algorithms present discrete posts to users.

Opinion 7: The recommendation algorithms used by Meta's social media platforms are separate from the content moderation algorithms used by those platforms, although the two interact in limited ways.

Opinion 8: The problems with recommendation algorithms cannot be solved through different or better content moderation policies or better enforcement of those policies.

- Experiments can be done before an update is deployed widely (pretest) or after (backtest). In a pretest, a small treatment group receives the update, whereas in a backtest, all users *except* a small treatment group (a holdout) do so.
- Experiments can be short term (typically lasting a few days) or longer term (months or sometimes years). Long-term experiments allow measuring whether updates have any effects on churn (users quitting their use of the app) and retention (users continuing to use the app). Since churn and retention accumulate over time, differences rates are usually too small to be discernible within a few days. On the other hand, changes in time spent on the app are more readily detectable.

V. Meta's Algorithms

A. Overview of recommendation algorithms

71. Meta's products include Instagram and Facebook, which I understand are the subject of the instant lawsuit and therefore my focus in this section. Instagram and Facebook each have multiple surfaces through which users can interact with the platform. Some of these surfaces are powered by ranking algorithms, and others, such as Direct Messages, are not. Table 2 identifies the main surfaces of Instagram and Facebook that utilize ranking algorithms.

CONFIDENTIAL

Table 2: Facebook and Instagram surfaces that use ranking algorithms

Product	Surface	Launch date	What is shown
Facebook	Feed	2006;[7] ranking introduced in 2010[8]	Posts from accounts the user does and doesn't follow
Facebook	Reels	2021 (launched with ranking)[9]	Short videos, mainly from accounts user doesn't follow
Instagram	Feed	2010;[10] ranking introduced 2016[11]	Mainly posts from accounts user follows
Instagram	Explore	2012 (launched with ranking)[12]	Mainly posts from accounts user doesn't follow
Instagram	Reels	2020 (launched with ranking)[13]	Short videos, mainly from accounts user doesn't follow

72. The posts shown on each of the surfaces identified in this table may be quite different, as are the surfaces' user interfaces. However, their underlying algorithmic logic is similar. Indeed, the descriptions in the previous section of how algorithmic recommendation works

[7] Meta. Our History. Meta.com. https://www.meta.com/about/company-info/. Last visited Mar. 13, 2025; Mosseri. Building a Better News Feed for You. about.fb.com. Jun. 29, 2016. https://about.fb.com/news/2016/06/building-a-better-news-feed-for-you/.

[8] META3047MDL-065-00143238; Kincaid. EdgeRank: The Secret Sauce That Makes Facebook's News Feed Tick. TechCrunch. Apr. 22, 2010. https://techcrunch.com/2010/04/22/facebook-edgerank/; Kincaid. Hacking The Graph: Live From Facebook's f8 Conference. Apr. 21, 2010. https://techcrunch.com/2010/04/21/hacking-the-graph-live-from-facebooks-f8-conference/.

[9] Facebook. Launching Reels on Facebook in the US. Sept. 29, 2001. https://about.fb.com/news/2021/09/launching-reels-on-facebook-us.

[10] Mosseri. Shedding More Light on How Instagram Works. about.instagram.com. Jun. 8, 2021. https://about.instagram.com/blog/announcements/shedding-more-light-on-how-instagram-works.

[11] Constine. Instagram is switching its feed from chronological to best posts first. TechCrunch. Mar. 15, 2016. https://techcrunch.com/2016/03/15/filteredgram/; See the Moments You Care About First. Instagram. Mar. 15, 2016. https://web.archive.org/web/20160315223405/http://blog.instagram.com/post/141107034797/160315-news.

[12] Constine. Instagram's New "Explore" Brings The Future Of Photo Discovery Into Focus. TechCrunch. Jun. 25, 2012. https://techcrunch.com/2012/06/25/instagram-explore/.

[13] Alexander. Instagram launches Reels, its attempt to keep you off TikTok. The Verge. Aug. 5, 2020. https://www.theverge.com/2020/8/5/21354117/instagram-reels-tiktok-vine-short-videos-stories-explore-music-effects-filters.

apply broadly to the surfaces in the table, as of the date that machine-learning-based ranking was introduced in each case.

73. Through discovery in this case, Meta has provided extensive detail about 15 ranking and recommendation systems that operate on Facebook and Instagram.[14] For Facebook, these are: (1) Feed, (2) In-Feed Recommendations, (3) Pages You May Like, (4) Suggested Groups, (5) Video, (6) Reels, (7) Stories, and (8) Notifications. For Instagram, these are: (9) Reels, (10) In-Feed Recommendations, (11) Feed, (12) Explore, (13) Stories, (14) Notifications, and (15) Search.

74. Meta's interrogatory response explains that its recommendation systems operate in two main stages: candidate generation (also called "retrieval") and ranking. The first stage retrieves thousands of posts from among the universe of "millions or even billions" that could be recommended.[15] The second stage then ranks this subset to present to the user.[16]

75. Some further detail about the second, ranking stage of the system is warranted. At this stage, the model considers a variety of *input signals* about the user and each of the posts gathered during the retrieval stage. These signals are used to make *predictions* regarding the user's likelihood of engaging with the post in various ways (i.e. "the probability that the [post] will be liked, shared, or commented on").[17] Those predictions are then weighted depending on importance, with the weights themselves sometimes derived by applying a machine learning system that has been tuned to prioritize engagement, and summed to obtain a final Ranking Score. The posts are presented in the order of their Ranking Score.

76. Meta's interrogatory response presents information about the top predictions used by each of Meta's ranking systems, including Facebook In-Feed Recommendations, Instagram In-Feed Recommendations, and Instagram Reels. My review of this information indicates that the

[14] Meta Third Supp. Response to Interrog. 2.

[15] Meta Third Supp. Response to Interrog. 2 at 9.

[16] Presentation of ranked content is subject to the outcomes of Meta's separate content moderation algorithms, discussed in later subsections.

[17] Meta Third Supp. Response to Interrog. 2 at 10.

vast majority of such predictions (or signals) concern user behavior and attributes rather than an evaluation of the "content" of the post (let alone its subjective quality).

77. As noted above, during the ranking stage of the two-step process, each of the above predictions is generated for each post identified during the retrieval stage. Importantly, though, these predictions are then *weighted* before they are summed to arrive at the post's Ranking Score. Meta's interrogatory response does not provide information about these weights for any of its recommendation systems. However, I have seen examples of such weights in internal company documents. For instance, the weights for the Instagram Reels algorithm at one point in time were specified as shown below.

Figure 3: Excerpt from META3047MDL-014-00405864 at META3047MDL-014-00405875

[redacted]

Weight	Term	Category	Weight	Term	Category
[redacted]	Comment	Consumer	[redacted]	Reshare	Consumer
[redacted]	Use audio	Producer	[redacted]	Like	Consumer
[redacted]	Save audio	Producer	[redacted]	Log time	Consumer
[redacted]	Follow	Consumer	[redacted]	Video complete	Consumer
[redacted]	Use effect	Producer	[redacted]	Repeat	Consumer
[redacted]	Save	Consumer	[redacted]	Transfer - cover video complete	Consumer
[redacted]	Profile visit	Consumer	[redacted]	Swipe forward	Consumer
[redacted]	Transfer - cover click	Consumer	[redacted]	Skip	Consumer

Table 1. The terms and their weights in the IG Reels Tab Value Model, ordered by weight. Actual production weights have higher decimal places accuracies than presented above.

78. This formula is relatively straightforward and illustrates the concept just discussed. Here, a given piece of content's Ranking Score is equal to the sum of various predicted probabilities (e.g. "p(comment)," a measure of how likely the user is to comment on a Reel) that are each given their own weight (e.g. 20.0). This formula for weighting the predictions to produce a single score is known as the **value model**. It represents the objective that the system attempts to optimize — the "what" of the recommendation algorithm.

79. The technical information presented above should not obscure the key point. The predictions of user behavior that a ranking algorithm considers, and the weights that are assigned to each such prediction, are not endogenously defined by the algorithm. Rather, they are the function of what objective the algorithm sets out to achieve. If the algorithm's objective is to maximize user engagement, then the predictions and weights that mathematically optimize that objective will follow suit.

80. The objectives set by the company are reiterated throughout the testimony of Joshua Simons who worked as a member of the Responsible A.I. Team at Facebook from 2018 to 2022. Discussing the machine learning systems for News Feed ranking, Simons explained that "the ranking system's impetus, what it's aiming to do, is set by the topline metric. And in Facebook's case, that was meaningful social interactions, different forms of engagement."[18] Simons further explained that optimizing for "meaningful social interactions" has the effect of provoking strong feelings in users like social comparison and feelings of "shame or disgust or fear or pride" which increase interaction with the platform and increase engagement, which results in more revenue for the company.[19] And yet, Meta has continued optimizing user feeds for MSI metrics, despite their own internal research showing harm to users' emotional wellbeing.[20]

81. Simons added to this point, saying that whatever the "system is asked and trained and built to optimize for, whatever forms of engagement you define as a thing that it optimizes for, shapes the content that's put under all of our noses and on the phones and computers of billions of people who use the platform, and that shapes how they feel and how they behave over time."[21] He explains that this ability to shape user behavior poses a particular risk of harm to vulnerable

[18] Simons Dep. 118:24-119:2; *see also* 68:14-15 ("[…] what you choose to train and build a system to do, creates the future on that platform").

[19] Simons Dep. 122:22-123:13.

[20] Simons Dep. 122:22-123:13.

[21] Simons Dep. 93:11-17; 116:6-10 ("What matters about the impact of the News Feed system on users is the way that those thousands of pieces of content are ordered […], and that is set by the objectives of the ranking system that Facebook built.").

populations like "kids who are not yet developed" because it "produces kids who grow and learn in ways that aren't […] commendable and [are] undesirable."[22] An internal document reflects nearly identical sentiment, stating "[t]eens are still developing, novelty-seeking and impulsive. . . Cool, so we probably shouldn't have ranking primarily driven by p(eyeballtime) and p(engagement), which is how it currently works."[23]

82. A few particularly concerning internal documents suggest Meta's awareness that its intentional design choices have a particular impact on adolescents. One document appears to indicate that the company's optimization strategy varies among groups of users, saying Facebook "ranks [teens'] content based on engagement much more than any other cohorts."[24] In another internal document, Facebook researcher [redacted] observed, "we literally optimize Teen feeds for engagement," and "it seems to genuinely be the case that optimizing for sessions just naturally distributes more harm to teenagers. . . it's been too consistent over time."[25]

83. Engagement is not the only way to rank information, or to optimize feed ranking. If the algorithm's objective is specified differently—for instance, if its objective is to maximize user welfare, user safety, or connections with friends and family members—the predictions and weights will change accordingly, to optimize that different objective. Because the choice of objective is key to the algorithm's functioning, I turn next to an examination of how Meta's recommendation algorithm relates to its business model.

[22] Simons Dep. 94:5-17.

[23] META3047MDL-047-00706929.

[24] META3047MDL-022-00025511; Cunningham Dep. 122:3-22 (discussing Ex. 13 [META3047MDL-047-00995848] which states in part "[t]he goal of teens team is to maximize aggregate teen time spent [. . .] ongoing projects aim to change teen behavior by increasing original post sessions, reshares, feedback given, page connotations [sic], and more.").

[25] META3047MDL-072-00032552; *see also* META3047MDL-046-00328644 ("I think we need to stop optimizing teen users' algos for engagement metrics to be in compliance but nobody seems to be talking beyond building new tools and strong enforcements.").

B. Recommendation algorithms and Meta's business model

84. Meta's business model relies on advertising revenue. As it has acknowledged in a recent Form 10-K filing, it "generate[s] substantially all of [its] revenue from selling advertising placements on [its] family of apps to marketers."[26]

85. Meta displays ads on each of the surfaces listed in Table 2. The amount of time that users collectively spend on these surfaces directly translates to ad revenue. As Meta's former VP of Advertising Technology put it succinctly at his deposition, "[m]ore usage correlates with more money."[27]

86. Meta's recommendation algorithm is an important part of its ad-based business model. In an April 2024 earnings call with investors, Meta CEO Mark Zuckerberg explained the business impact of recommendations in user's feeds:[28]

> AI is already helping us improve app engagement which naturally leads to seeing more ads, and improving ads directly to deliver more value. So if the technology and products evolve in the way that we hope, each of those will unlock massive amounts of value for people and business for us over time. We are seeing good progress on some of these efforts already. Right now, about 30% of the posts on Facebook feed are delivered by our AI recommendation system. That's up 2x over the last couple of years. And for the first time ever, more than 50% of the content people see on Instagram is now AI recommended. AI has also always been a huge part of how we create value for advertisers by showing people more relevant ads, and if you look at our two end-to-end AI-powered tools, Advantage+ Shopping and Advantage+ App Campaigns, revenue flowing through those has more than doubled since last year.

87. To put this more succinctly: Meta uses machine learning algorithms to rank posts that users see, in order to increase use of Instagram and Facebook. More usage means more views of ads, which generates more revenue for the company.

---

[26] Meta Response to Req. for Admission 45.

[27] Boland Dep. 94:3-4.

[28] Meta Platforms, Inc. (META). First Quarter 2024 Results Conference Call. Apr. 24, 2024. https://s21.q4cdn.com/399680738/files/doc_financials/2024/q1/META-Q1-2024-Earnings-Call-Transcript.pdf.

88. One major difficulty faced by Meta is that existing machine learning technology is not capable of effectively learning and predicting the relationship between how particular posts are ranked in a user's feed and metrics that bear directly on revenue.

89. As a matter of course, Meta uses a set of metrics to assess how well it is doing at maximizing revenue. Indeed, Instagram and Facebook each have a set of "core metrics" that are considered "canonical"—meaning, "every time you make any sort of change…you have to check your test against these metrics to confirm that … you didn't affect them in a negative way."[29] Some of these metrics measure revenue itself. Other metrics measure various kinds of growth in usage which, as explained above, is closely tied to revenue.[30] These include DAU (Daily Active Users — the number of users active on the app per day), MAU (Monthly Active Users — the number of users active on the app per month), time spent on the app, the number of sessions (with each use of the app from opening to closing counting as one session), and the aggregate number of times posts have been viewed.[31] For example, Tom Cunningham, a former senior data scientist at Facebook, defined "top-line metrics" for Facebook News Feed as metrics which evaluate the performance (or "health") of News Feed, like DAU, daily active users, time spent, or number of comments, as well as MSI and number of sessions.[32]

90. However, the relationship between these usage metrics and the choices made by Meta's ranking algorithms is imprecise at best. This is in part because the relationship between a user seeing a single post and the user's choice to stay active or not active on the app (thus affecting a metric such as DAU) is too diffuse. As of 2022, Meta had research projects that attempted to solve this problem but no solution has made it into its products as far as I am aware.[33]

---

[29] Volichenko Dep. 49:11-23.

[30] META3047MDL-019-00123767; Volichenko Dep. 50:5-9 (Q: "Were any of the Instagram critical metrics related to growth?" A: "Yes, yes, absolutely.").

[31] META3047MDL-019-00123767; Meta's Resp. to Interrog. 15; META3047MDL-050-00215913.

[32] Cunningham Dep. 104:14-18; 106:7-22 (discussing Ex. 11 [META3047MDL-047-00765317]).

[33] META3047MDL-014-00405380.

CONFIDENTIAL

91. As a result, Meta trains its ranking algorithms to optimize proxies rather than to directly optimize revenue-related metrics.[34] These proxies primarily consist of the behavior of users when viewing individual posts, such as liking, commenting, sharing, or watching a video to completion. Collectively, these actions are referred to as engagement, an important term of art at Meta. Meta's ranking algorithms aim to maximize users' engagement with the respective surfaces on which they operate. The proxies and the relative weights given to them are chosen in such a way that Meta engineers predict — in reality, hope — will optimize revenue-related metrics.[35]

92. One internal document states:[36]

> Instagram does not care about the number of comments or likes. Those events are proxies for long term value, which we measure in DAP (daily active people) and time spent (time per daily active). We rely on these proxies because it is difficult to make a prediction on whether an impression will result in you coming back to the app tomorrow or spend more time in the long term.
>
> …
>
> The value model is one of the most important and least scientific parts of ranking. When we discuss changes to the value model, we are arguing about what causes people to return to Instagram on a multi-year time horizon. It might be good in the short term to increase time spent by showing you all video, but in the long term people feel alienated and stop posting. Less posting means less interesting content, which makes our platform a worse place to be. We've learned many of these lessons from past launches and experiments Facebook's Blue App has made.

93. In another internal document, Facebook researcher [redacted] explained, with reference to this litigation, how Facebook's value model is tuned to rank posts with the single goal of maximizing engagement:[37]

> [R]elevant to the lawsuits, and some foreshadowing from Mark in Q&A just now – basically the company position is that time spent and engagement

---

[34] META3047MDL-014-00047744.

[35] META3047MDL-208-00057329; META3047MDL-223-00002454.

[36] META3047MDL-047-00544942.

[37] META3047MDL-047-00706686; *see also* META3047MDL-072-00139365 and META3047MDL-046-00328644.

> metrics proxy user value and we are 'not' just tuning the algorithm to keep people scrolling as much and as long as possible... except mechanically, we basically absolutely do through a process called the "guided value model" (GVM) process. [redacted] which is then fed into an algorithm that is told to optimize for (a) sessions; and (b) vpvs. It spits out a bunch of changes to the weights of different terms and then we ship it. So really, the algorithm is absolutely tuned to maximize engagement in an maximally empirical, principle-less way. I am not sure that the youth legal people realize this, and that it applies to teens just as much as gen pop... so the things contained in the lawsuits are somewhat hard to refute so long Teens are part of the GVM process.

94. I have seen many indications that Meta's efforts at maximizing engagement (as a proxy for maximizing usage and therefore revenue) subordinated its other efforts related to minimizing negative and potentially harmful user experiences. Former data scientist George Volichenko testified about this directly:

> "Q. If a proposed safety feature negatively impacted Instagram's critical metrics, in your experience, could that impact the potential launch of the feature?
>
> A. Yes."
>
> …
>
> "Q. [I]f a proposed safety feature did have a negative impact on the explorer [sic] team's metrics, would that be an issue in actually rolling out the safety feature?
>
> A. Yes. Yes, that would be an issue."

95. One former data scientist, Tom Cunningham, described a "very complicated long-running negotiation over how to make a trade-off between integrity and engagement objectives […]."[38] But testimony from another research scientist indicates that this trade-off was routinely resolved in favor of engagement. From his experience as a research scientist with Facebook from 2018 to 2022, Joshua Simons explained that what the platform presents to users on their phones or computers was "driven by the feed system and the engagement metrics and meaningful social interactions that it was aiming at," and "the integrity system, whether it's hate speech, the toxicity model that wasn't rolled out, or the other systems in it, made barely a dent on the core ranking

[38] Cunningham Dep. 102:6-13.

CONFIDENTIAL

system and the imperatives that drove it."[39] Simons referred to the integrity system as "picking out bits" from a "huge pile."[40]

96. According to Simons, incentives for product engineers (those designing the platform's machine learning systems) were not aligned with user wellbeing or safety but instead aligned with maximizing Meaningful Social Interaction metrics.[41] "[I]f the systems they were building boosted those metrics, they were rewarded and promoted and given headcount, and if it had a deleterious effect on meaningful social interactions…they were not."[42] Simons further testified that during his time at the company, not only were engineers not incentivized to take user safety into account in their design, but engineers' performance bonuses only considered whether certain increases in engagement metrics were realized.[43]

97. Simons' testimony is consistent with testimony given by Natalie Troxel, a user experience researcher at Meta:[44]

> Q. Did you encounter impediments to -- while you were at Meta to achieving your goal of having your work promote social -- pro-social development?
>
> A. Yeah.
>
> Q. And what were some of those impediments?
>
> A. Some of it was, I think, the nature of the company. I think that I had joined the company perhaps a little naive about what the company would want to do and some of what the company was building and how much of an influence I would be able to have on things. I think some of it was that technically children under 13 aren't supposed to be using the products. So there wasn't going to be a focus on young children's development, for

[39] Simons Dep. 53:9-23.

[40] Simons Dep. 53:9-23.

[41] Simons Dep. 102:23-103:15.

[42] Simons Dep. 102:23-103:15.

[43] Simons Dep. 104:4-13.

[44] Troxel Dep. 31:5–32:14.

> example. And I think some of it was just the focus, the prioritization of the company.
>
> Q. When you say the "prioritization of the company," what do you mean by that?
>
> A. That the company wasn't prioritizing positive child development.
>
> Q. In your experience, what was the company prioritizing?
>
> A. Revenue.
>
> Q. And by "revenue," do you mean making money?
>
> A. Yes.
>
> Q. And how, in your experience, did Meta seek to prioritize revenue and making money?
>
> A. It seemed like when I was there, that Meta was trying to increase engagement, trying to increase how much time people spent using their apps in an effort to have more eyeballs on the apps for a longer period of time, which can get them more advertising revenue.

98. The testimony above is consistent with examples I have seen in Meta's internal documents. For example, in a chat conversation between Mr. Cunningham and colleague [redacted] discussing the performance of integrity experiments involving "friend" recommendations to Facebook users, Cunningham asked if there is a rule of thumb for how trade off decisions were made. [redacted] retorted, "[y]es, if it hurts growth, don't ship it . . .[a]nd only shipping treatments that have very low impact on important metrics."[45]

99. Another example concerns a proposed Take A Break feature which, if operational for a user, would encourage them to pause usage of the platform after a certain amount of time spent. The articulated goal for this feature was to promote greater well-being by helping teens take a break from Instagram when they spend too much time in a single session.[46] Multiple versions of the Take A Break feature were considered; however, engineers were instructed to "disregard

---

[45] META3047MDL-047-00765317; *see also* Cunningham Dep. 107:24-110:12.

[46] META3047MDL-111-00189944.

variants with material ecosystem regressions."[47] Specifically, during testing of the feature, "guardrail" metrics were set to ensure the feature did not create regressions of time spent or sessions greater than 0.2%, or regressions of creator watch time greater than 0.5%.[48] Ultimately, employees involved in the project expressed disappointment at the version that was released, as less protective of teen well-being than other available alternatives.[49]

C. Overview of content moderation algorithms

100. Like other platforms, Meta takes various steps to police its community standards (the rules governing what is and is not allowed on its platforms). Meta then measures its success at enforcement by analyzing the prevalence of integrity violations and through surveys of its users to measure their subjective experiences.[50]

101. Meta's safety and integrity teams build classifiers to identify posts that violate policies on its platforms. The main policies relevant to the present matter are about Suicide and Self-Injury (SSI), Eating Disorder (ED), bullying and harassment, graphic and violent content, adult nudity, and child safety. Varying levels of detail about Meta's classifiers were shared with me in the form of Meta's responses to an interrogatory.[51]

102. Classifiers assign scores to each post that they process, representing the probability that the post is policy violating. If a score exceeds a specified threshold value, Meta takes action

[47] META3047MDL-031-00242248 at -00242256; *see also* Volichenko Dep. 93:13-94:4 (Q: "During your work on Take a Break, were you asked to test its impact against growth metrics?" A: "Yes…they were always a thing that we had to evaluate.").

[48] META3047MDL-031-00242248 at -00242252.

[49] Volichenko Dep. 109:20-24 (Q: "[W]ere you disappointed by the impact that Take a Break was having on youth safety and well-being?" A: "Yes, definitely."), 107:2-8 ("it would be more effective … if we turned it on for all teens, made it opt-out. But we were unable to test it…"), 113:23-114:10 (Q: "Do you think teen users would have been safer if Take a Break had been turned on as a default feature?" A: "I think so, yes." Q: "Did Instagram leadership agree to turn Take a Break on by default?" A: "No. I believe that never happened."). Bejar Dep. 574:9-14 ("I can't imagine this being effective at helping a teenager quit out the app or put down the phone and effectively take a break.").

[50] META3047MDL-019-00123767.

[51] Meta's Fourth Supp. Response to Interrog. 1.

on the post—which may range from automatic removal, enqueueing for human review, or "soft action." Soft action refers to reducing users' exposure to a post without removing the post. Types of soft action include downranking and adding warning screens.

103. Since 2020, Meta has also had "borderline" classifiers.[52] These classifiers identify posts that are not strictly policy violating but are nonetheless "potentially problematic or low quality." Borderline classifiers are not used to remove content, only downrank them.

104. Importantly, Meta's classifiers do not always automatically remove content—even content flagged as having a high likelihood of violating Meta's policies.

105. Consider, for instance, Meta's suicide and self-injury classifier ("SSI Classifier"). The SSI Classifier sets a threshold value for autodeletion of 0.94. This apparently means that, if a piece of content is flagged by the classifier as having a 93% likelihood of containing violative suicide promoting content, it is not removed from the platform.[53] This does not appear to be consistent with Meta's public policy, which states that "We remove any content that encourages suicide, self-injury or eating disorders, including fictional content such as memes or illustrations, and any self-injury content which is graphic, regardless of context."[54] For this statement to be accurate, it ought to read instead: "We remove any content that we are 94% or more confident encourages suicide, self-injury or eating disorders…" It is true that Meta uses human content reviewers as a backstop to its classifiers. However, I have been given no data concerning what percentage of enqueued content is actually reviewed, how quickly, or how accurately. For reasons discussed further below, it is my opinion that human content moderation is insufficient to identify policy-violating content. It is also true that Meta has lower threshold values for various soft actions, but soft actions fall short of what the policy states, which is removal.

[52] Meta's First Supp. Response to Interrog. 1 at 19.

[53] Meta's Fourth Supp. Response to Interrog. 1.

[54] Meta, Transparency Center. Policies-Community Standards: Suicide, Self-Injury, and Eating Disorders. https://transparency.meta.com/policies/community-standards/suicide-self-injury/. Last visited May 13, 2025.

106. Some of the other classifiers designed to enforce serious policies are not used for autodeletion at all. That is true for Meta's eating disorder classifier ("ED Classifier"), various Graphic Violence classifiers (such as the "Graphic Violence MAD [Mark as Disturbing] video classifier on IG"), and—most troublingly—Meta's classifiers designed to identify Child Sexual Abuse Material ("CSAM"), including its CSAM classifier, CSAM Solicitation, and Child Sexualization (CSx) classifier.[55] Meta's inability to identify and autodelete this content is, to my knowledge, not a public fact. It is inconsistent at least with Instagram's Community Standards, which, again, claim to "remove any content that encourages … eating disorders," and which have proclaimed since 2015, "We have zero tolerance when it comes to sharing sexual content involving minors."[56]

VI. Problematic features of Meta's algorithms

107. As discussed extensively above, Meta's recommendation algorithms are optimized to maximize user engagement over the short term, at the expense of all other considerations including user welfare. In this section, I discuss common ways that Meta's recommendation algorithms operate, from a technological perspective, in ways that promote problematic usage. In Subsections A and B, I discuss how the design of Meta's recommendation algorithms promotes problematic usage by programmatically exploring and then exploiting users' unstated proclivities. In Subsections C and D, I discuss how the design of Meta's recommendation algorithms can cause a user to experience forms of problematic use; for instance, how they can cause a user's feed to become more homogenous over time and lead to rabbit holing. In Subsection E, I discuss how Meta's product testing practices encourage addictive design.

A. Discovering behavioral proclivities

108. To predict engagement by a given user on a given post, recommendation algorithms typically work by trying to answer the question: How did users similar to this user engage with

[55] Meta's Fourth Supp. Response to Interrog. 1.

[56] Meta's Resp. to Req. for Admission 8.

posts similar to this post? The intuition behind this logic is straightforward: Two people who have something in common—a hometown, a hobby, a community they are embedded in, a celebrity they follow—will both engage with posts that relate to that shared interest. You don't need to know what the interest is or what the content of the posts are, you just need to know that the users' behavior is similar. This approach has repeatedly proven to work well in practice.[57]

109. Suppose that fans of a particular singer are more likely to be depressed than the general population, and that depressed users tend to have patterns of engagement that differ from other users. This is not a speculative assumption: the fact that people's associations with brands correlate with personality traits is well known from research.[58] Internal Instagram research reveals that followers of celebrities including Ariana Grande, Kylie and Kendall Jenner, and Addison Rae tend to feel worse about themselves than other users.[59]

110. Suppose a user follows the singer in question on a social media platform, but has never made searches or taken other actions that explicitly indicate depression. Based on the correlational logic discussed above, recommendation algorithms will nonetheless infer that the user is similar to other depressed users, and will take that into account when determining how to maximize the user's engagement on the platform.

111. This ability to infer hidden correlations has long been a core feature of recommendation algorithms, predating the application of recommendation algorithms to social media. For example, more than two decades ago, Amazon implemented the "if you liked that book, you might like this one" feature.[60] This type of inference does not require any knowledge of

---

[57] Narayanan. Understanding Social Media Recommendation Algorithms. Knight First Am. Inst. Mar. 9, 2023. https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms.

[58] Kosinski et al. Private traits and attributes are predictable from digital records of human behavior. PNAS. Mar. 11, 2013. https://www.pnas.org/doi/10.1073/pnas.1218772110.

[59] META3047MDL-003-00159559.

[60] Linden et al. Amazon.com Recommendations: Item-to-Item Collaborative Filtering. IEEE Computer Society. Jan.-Feb. 2003. https://www.cs.umd.edu/~samir/498/Amazon-Recommendations.pdf.

specific correlations from the developer of the algorithm. Nor does the algorithm itself "know" in any meaningful sense that the user is depressed or that the posts that are being recommended to them concern depressing or depression-oriented topics.

112. A second feature of recommendation algorithms that promotes problematic behavioral patterns is "exploration." When a user creates a social media account, the algorithm has minimal information on the user. To learn the user's interests or proclivities, the algorithm starts by recommending what is popular, while at the same time ensuring a diverse variety of posts, thus "exploring" the user's behaviors before ultimately narrowing what it recommends to the user. Exploration is explicitly programmed into recommendation algorithms—for all users, but especially for new users.

113. Exploration for new users is often referred to as a "cold start" where a platform is trying to gather engagement data to make predictions on what it thinks a user will engage with. Meta has stressed the importance and complexity of exploration.

114. Meta's internal documents repeatedly stress the importance of exploration: "Exploration by definition is uncertain and 'risky', and a core purpose is to try, get signals, and then adjust. To do this we should be able to observe how the user reacts to new items and adapt quickly, [for] which we need both low latency systems as well as adaptive algorithms."[61]

115. As an illustrative example of the effectiveness of exploration, a Wall Street Journal investigation created bots, or automated accounts, with various interests and programmed them to use TikTok.[62] The investigation revealed how the algorithm attempts to discover these interests. For one bot programmed to engage with videos related to sadness and depression, the fifteenth recommended video, a mere three minutes into its use of TikTok, was a hit, and after that TikTok's

[61] META3047MDL-014-00405308.

[62] Wall Street Journal. Investigation: How TikTok's Algorithm Figures Out Your Deepest Desires. WSJ Video. Jul. 21, 2021. https://www.wsj.com/video/series/inside-tiktoks-highly-secretive-algorithm/investigation-how-tiktok-algorithm-figures-out-your-deepest-desires/6C0C2040-FF25-4827-8528-2BD6612E3796.

algorithm gradually increased the fraction of such videos in the user's "For You" feed, eventually reaching 93%.

116. Meta's internal research identified cold start problems similar to those revealed in the WSJ investigation. For example, Meta set up test accounts that only followed suggestions provided by the platform, which very quickly led to an extreme barrage of harmful content. The document emphasized that "[i]n the 3 weeks since the account has been opened, by following just this recommended content, the test user's News Feed has become a near constant barrage of polarizing nationalist content, misinformation, and violence and gore."[63] Mark Zuckerberg himself apparently experienced this phenomenon with a competitor social media platform, writing privately to a coworker: "Random thing. I just logged into Snapchat for the first time in a while to see what progress they're making with AR. What surprised me is how porny and inappropriate the content they're featuring is in their equivalent of Reels (Spotlight) and their 3P content stories (Stories). I was completely blown away…especially given that their audience skews so much younger. It's not nudity but it's really raunchy. I would never let my kids anywhere near this."[64]

117. Exploration has an important role in preventing users' feeds from becoming too homogenous over time. But as I explain below, by helping algorithms learn users' unstated interests and behaviors, exploration *also* contributes to a phenomenon known as "rabbit holing." It is tricky to get this balance right, as I also explain below.

118. A third salient feature of recommendation algorithms is the use of implicit signals in addition to explicit signals. As discussed extensively above, signals are inputs to recommendation algorithms, such as a user's likes or geographic location. *Explicit* signals are users' intentional expressions of interests or preferences, while *implicit* signals are behaviors from which the algorithm can infer interests or preferences.[65] There is not an agreed-upon classification

---

[63] META3047MDL-207-00024997.

[64] Zuckerberg Dep., Ex. 83.

[65] Jannach, Lerche, & Zanker. Recommending based on implicit feedback. Social information access: systems and technologies. Springer International Publishing. 2018.

of signals as implicit or explicit, and the same signal may be implicit or explicit depending on the user's level of knowledge of the implications of the action. However, there is a clear spectrum of explicitness: search terms, "see more/less like this" buttons, and written reviews or comments are highly explicit; likes are less explicit; and dwell time is generally considered implicit. Dwell time refers to the amount of time the user views or lingers on a post before moving on to the next one.

119. Meta uses implicit signals in its recommendation algorithms. For example, the Instagram Reels algorithm uses thirteen consumer-oriented signals, six of which are "explicit interaction based signals" and five of which are "implicit time-spent related signals" (as well as other signals not directly related to the user).[66]

120. Implicit signals lead to discovery of unstated interests because they are more likely to be the result of "System 1" (impulsive) than "System 2" (deliberative) cognitive processes.[67] By the very nature of impulsive behavior, users are less likely to be able to take steps to control those inputs to the algorithm. System 1 and System 2 are a concept relevant to several disciplines – including cognitive psychology, behavioral economics, and artificial intelligence – which describes two human modes of thinking, one (System 1) instinctive and emotional, and the other (System 2) deliberate and logical.[68]

B. Exploiting behavioral proclivities

121. Many users report not only that they spend a large amount of time interacting with algorithmically curated feeds that they later regret, but also that these feeds become more homogenous over time in ways they did not intend. In this section I explain why this behavior is a plausible consequence of the way these algorithms are designed.

---

[66] META3047MDL-014-00405874. The remaining two consumer-oriented signals are described as "learning transferred signals with models transfer-learned from non-tab data."

[67] Moehring et al. Better Feeds: Algorithms That Put People First. Knight-Georgetown Institute. Mar. 4, 2025. https://kgi.georgetown.edu/research-and-commentary/better-feeds/.

[68] *Id.* One former Meta employee describes these systems as "considered goals" versus "base instincts" that all humans possess and explains that humans want to "develop in ways that reflect our higher-order goals" like "staying healthy and supporting our family, […] and not our base instincts to eat a cookie or McDonalds or whatever." Simons Dep. 97:4-17.

i. Engagement optimization

122. Engagement optimization prioritizes the content that users are most likely to engage with.[69] To understand the consequences of this logic, it is important to keep in mind that the average engagement rate on every major social media platform is low. The engagement rate of a post is the number of engagements (likes, comments, and so forth) divided by the number of views.

123. According to third-party engagement statistics, Instagram and Facebook have engagement rates under 1%.[70]

124. Low engagement numbers are not at odds with the fact that social media platforms are experienced by users as highly engaging and potentially addictive. It is important to note that even when a user enjoys a post, they may not engage with it in a manner that can be captured by third-party engagement statistics, such as liking or commenting. Besides, even if a post is correctly predicted to be of interest to the user, it may not interest them at a given moment.

125. Users might differ in terms of how often they engage with posts, but consider a user who is in the typical range, engaging with about 1-3% of posts in their feed. So if there is a particular type of content with which they are predicted to be 10% likely to engage, that content will score extremely highly compared to the alternatives. In other words, even if the user resists their impulse to engage with such content 90% of the time, but gives in to the impulse just 10% of the time, that will be more than enough to ensure that the algorithm ranks such content highly when creating their feed. With the rise in the importance of implicit signals, this does not even require active engagement such as a like, but simply dwelling on posts for a few extra seconds.

126. In short, the goal of engagement maximization algorithms is to maximize the time that users spend on the app (over some time horizon), and the social media platforms I examined

[69] META3047MDL-072-00032552; META3047MDL-047-00706929; META3047MDL-047-00706686; META3047MDL-072-00139365; META3047MDL-046-00328644.

[70] SocialInsider. 2025 Instagram Benchmarks. https://www.socialinsider.io/social-media-benchmarks/instagram. Last visited Mar. 14. 2025.; SocialInsider. 2025 Facebook Engagement Benchmarks. https://www.socialinsider.io/social-media-benchmarks/facebook. Last visited Mar. 14. 2025.

optimize for this narrow goal, regardless of whether it leads to experiences that users will later regret.

127. In a workplace chat from 2023, a researcher for Facebook expresses this idea well, noting he is curious about "what is going on with content that is getting high relevance scores on behavioral models like e(linear_vpvd/eyeballtime)" despite being content "users don't want to see (low wyt; p(smsl)<0.4) as I suspect that is where high-compulsivity content is."[71]

ii. Feedback loops

128. A second reason why feeds may become more homogenous over time is feedback loops, which are defined as cycles in which user behavior and recommendation algorithms influence each other. It is a well-known concern about recommender systems and has been a topic of research study.[72]

129. Meta's "Ranking and Value Model Deep Dive" document notes that "models are trained on data that's strongly influenced by themselves, and this forms a **strong feedback loop**."[73]

130. A feedback loop could occur because a user's interest in a category of content may become more pronounced over time due to the consumption of content on that topic, leading to an increase in the prevalence of that type of content in the user's algorithmic recommendations. This might not necessarily reflect unwanted content. For example, a user may develop a gradually deepening interest in history by watching history videos, and find the experience rewarding. But a

---

[71] META3047MDL-022-00025511.

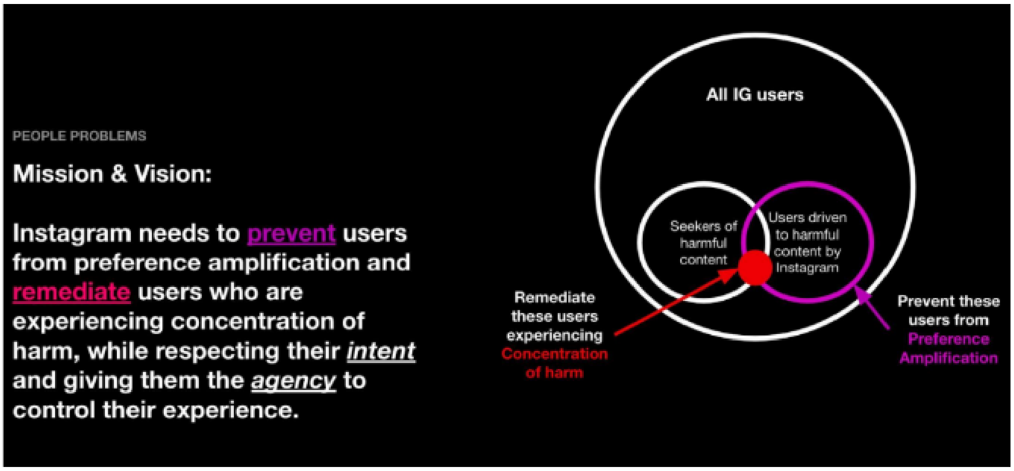
[72] Kalimeris et al. Preference Amplification in Recommender Systems. KDD '21: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. Aug. 14, 2021; Mansoury et al. Feedback Loop and Bias Amplification in Recommender Systems. CIKM '20: Proceedings of the 29th ACM International Conference on Information & Knowledge Management. Oct. 19, 2020; Tong et al. Navigating the Feedback Loop in Recommender Systems: Insights and Strategies from Industry Practice. RecSys '23: Proceedings of the 17th ACM Conference on Recommender Systems. Sept. 14, 2023; Jiang et al. Degenerate Feedback Loops in Recommender Systems. AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. Jan. 27, 2019; Matias & Wright. Impact Assessment of Human-Algorithm Feedback Loops. Just Tech. Mar. 1, 2022.

[73] META3047MDL-014-00405864 (emphasis in original).

feedback loop may also represent the result of the platform promoting content the user never sought out in the first place, leading to worsening addictive behavior over time and regretful use of social media. As discussed earlier, recommendation algorithms typically do not discern the difference between the two patterns.

131. Below are images from another Meta document discussing users whose feeds have a high prevalence of harmful content. As depicted below, this prevalence issue is not necessarily a result driven by user intent; and in fact, the document makes clear that users can be "driven to harmful content *by* Instagram."[74]

Figure 7: From META3047MDL-054-00000061



132. As former Facebook Responsible AI team member, Joshua Simons, testified, the objective the company sets for the system "shapes the feelings and the behaviors over time of the people who use it, and folks in computer science call this performative prediction, and it's performative in the sense that whatever you ask the system to do, over time people perform the very thing that you've asked the system to optimize for."[75] Based on his direct experience working with the designers of these systems, Simons understood that "by defining value in terms of those

[74] META3047MDL-054-00000061.

[75] Simons Dep. 93:18-24.

proxies—clicks, likes, shares, plays—you ended up incentivizing repeated patterns of behavior," that were ultimately "undesirable, and sometimes actively harmful for the users."[76]

133. These loops are not created by the algorithms affirmatively selecting for particular topics. Instead, it is one of the predictable outcomes of designing algorithmic recommendation systems to maximize particular user behaviors.

iii. Creator incentives

134. A third reason why feeds may become more homogenous over time is the way in which algorithms shift the type of content that is created on the platform. Content creators, by and large, have strong incentives to maximize the reach, or viewership, of the content they publish. For some of them it is their livelihood, or aspired livelihood.[77] For others, there may be non-monetary yet significant rewards. In fact, there are many courses on online learning platforms such as Skillshare promising to teach influencers on Facebook, Instagram, and other platforms how to maximize the reach of their content by appealing to the algorithm.[78] These courses collectively have tens of thousands of paying students.

135. Shaping the incentives for creators is a key goal in the design of Meta's recommender systems. For example, the system is designed to reward new creators with increased distribution if they are successful at engaging users. Meta's documents also discuss advanced strategies such as "build[ing] models to predict what content to show to a user to motivate the user to create."[79] Another example is setting up a "flywheel" where consumption drives production through "mimicry" — users creating posts similar to those which have been recommended to them

---

[76] Simons Dep. 81:19-25.

[77] Hughes. Food Businesses Lose Faith in Instagram After Algorithm Changes. The New York Times. Mar. 22, 2022. https://www.nytimes.com/2022/03/22/dining/instagram-algorithm-reels.html.

[78] Huang. TikTok Class Is in Session. The New York Times. Dec. 27, 2022. https://www.nytimes.com/2022/12/21/technology/tiktok-influencer-classes.html.

[79] META3047MDL-014-00405308.

in their feeds — which in turn influences the "inventory" of posts on the platform and consumption of those posts.[80]

136. Some creators are particularly attuned to the incentives of the recommendation algorithms because Meta offers them monetary incentives by splitting advertising revenue with them.[81]

137. If the recommendation algorithm learns to maximize engagement by triggering addictive behavioral patterns or amplifying extreme content, creators will be incentivized to create content that satisfies these demands. Many types of creators ranging from children's video channels[82] to politicians[83] have been known to modify their posting behaviors over time in response to algorithmic incentives in ways that are harmful to users and society.

C. Problematic use as a consequence of algorithm design

138. Meta's engagement-optimized ranking algorithms often operate in ways that functionally mirror the psychological mechanism of *intermittent variable rewards,* an idea that is foundational to persuasive technology.[84]

139. Imagine a vending machine that learns your stated and unstated preferences (or weaknesses) over time. But instead of simply giving you your favorite snack every time you press a button—which would quickly become predictable and dull—it introduces an element of surprise.

---

[80] META3047MDL-014-00405864.

[81] Scanlon. What it takes to get paid by YouTube, TikTok and other social platforms. Digiday. Feb. 28, 2025. https://digiday.com/marketing/what-it-takes-to-get-paid-by-youtube-tiktok-and-other-social-platforms; Instagram Help Center. About monetization for ads in Instagram profile feed. https://help.instagram.com/5485466918184985.

[82] Townsend. ElsaGate: The Problem With Algorithms. United States Cybersecurity Magazine. https://www.uscybersecurity.net/elsagate/.

[83] Uberti. How Fox News dominates Facebook in the Trump era. Vice. Apr. 29, 2019. https://www.vice.com/en/article/how-fox-news-dominates-facebook-in-the-trump-era/; Facebook internal report. "Political Party Response to '18 Algorithm Change." Apr. 1, 2019. https://fbarchive.org/doc/odoc889111.

[84] Fogg. Persuasive Technology: Using Computers to Change What We Think and Do. (Interactive Technologies). Morgan Kaufmann. Dec. 30, 2002; Eyal & Hoover. Hooked: How to Build Habit-Forming Products. Portfolio. Nov. 4, 2014.

It deliberately delays the best snacks, occasionally serves something unusual to gauge your reaction, and every so often, unexpectedly delivers your favorite candy bar. Over time, you find yourself engaging more with the machine itself than actually enjoying the snack.

140. The effect of an engagement-optimized recommendation system on users mirrors that of this vending machine. An engagement-optimized feed doesn't present content in a linear, predictable fashion. Unpredictability is baked into the system. Due to the use of machine learning to select posts from a vast inventory, engaging with a post from a creator is no guarantee that the next post from that creator will reach your feed. The exploration aspect of recommendation algorithms creates further unpredictability. In doing so, it mirrors behavioral conditioning patterns, reinforcing usage through intermittent variable rewards.

141. "Problematic use" is a term of art gleaned from internal company documents. Facebook researchers published internal research on problematic use in 2019.[85]

142. Even before that, documents show it was a concern of the company. An internal presentation on "Facebook 'Addiction'" starts out with the statement that "There is growing concern in the public that technology companies are deliberately manipulating users by designing addictive products that can harm well-being and subvert self-control."[86]

143. A 2020 Meta presentation deck on the topic defines problematic use this way:[87]

> We define Problematic Use (PU) as experiencing both of the following issues "very often" or "all the time":
>
> - Lack of control or feelings of guilt over Facebook use.
> - Negative impact in at least one of the following areas: productivity, sleep, parenting, or relationships.

---

[85] Cheng et al. Understanding Perceptions of Problematic Facebook Use: When People Experience Negative Life Impact and a Lack of Control. CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems Proceedings. May 2, 2019. https://doi.org/10.1145/3290605.3300429.

[86] META3047MDL-020-00563113.

[87] META3047MDL-079-00000177 at -00000181; Zuckerberg Dep. 121:16-124:2 (discussing this Meta document's definition); META3047MDL-035-00005017 (document from 2022 providing a very similar definition).

This is sometimes referred to as "social media addiction" externally.

144. From the same 2020 presentation above, Meta reports the finding that problematic use—as they have defined it— is affecting "about 12.5% of people on Facebook."[88]

145. [redacted]—who spent more than a decade with Meta and built the company's machine learning ranking and recommendation systems before moving into responsible AI near the end of his tenure—expressed concern in one internal document that "[s]ophisticated reinforcement learning might increase the problem of tech addiction in Stories, News Feed, or Instagram."[89]

146. [redacted] concerns were not without merit according to findings of another user survey on problematic use. The results of that study, which took place between December 2022 and January 2023, linked perceived problematic use with platform products or surfaces that utilize ranking and IVR tactics.[90] For Instagram, users identified as the top three problematic surfaces where they felt least in control of their time: 1) Discover, 2) Stories, and 3) Reels. For Facebook, users' top three were: 1) Notifications, 2) Stories, and 3) News Feed and/or Reels.[91]

147. Indeed, I observed in Meta's documents a number of emails sent directly to Zuckerberg from users telling the founder that they were addicted to his platform.[92]

---

[88] META3047MDL-079-00000177 at -00000183; Zuckerberg Dep. 124:3-125:5; *see also* META3047MDL-065-00316061 (internal Meta study from 2018 titled "User Perception of Time-Spent on Instagram" showing that more than a third of the survey respondents feel they have spent more time than they would have liked on Instagram, and that about one in five respondents think they have a problem controlling their time).

[89] META3047MDL-014-00233670. Simons, who worked with [redacted] at Facebook, testified that [redacted] comment reflected the concern that using this form of AI to optimize ranking for engagement "might supercharge problems of addiction across those products." Simons Dep. 108:8-24.; *see also* Simons Dep. 67:2-17 ("[T]he best way to increase engagement is addiction.").

[90] META3047MDL-074-00027496.

[91] META3047MDL-074-00027496.

[92] Two examples of user emails to Zuckerberg are META3047MDL-040-00028597 and META3047MDL-040-00028639.

148. Problematic use is not just a phenomenon affecting users of Meta's products. It occurs as a known result of design choices, including engagement optimized ranking, and intermittent variable reward mechanisms, and is reflected in documents I have reviewed from Meta.

D. Rabbit holing as a consequence of algorithm design

149. Another related consequence of the design of Meta's recommendation algorithms is a phenomenon known as "rabbit holing." Rabbit holing is not a technical term, but drawing on the common understanding of the phenomenon as well as the research literature, I define it as a pattern of social media use in which a user, without intending to, spends a large amount of time consuming an algorithmically recommended feed and later regrets it, with the feed becoming more homogenous over time. Rabbit holing can be considered a form of problematic use characterized by exposure to an increasingly homogenous feed. Rabbit holing can refer to a single session, or it may refer to a pattern that intensifies over a longer period, weeks or months.

150. Meta has not provided adequate access to external researchers to study the effects of their products on users and society. As a result, it has not been possible to rigorously test the causal effect of recommendation algorithms on outcomes such as rabbit holing. Despite the inability to test the "end to end" effect of recommendation algorithms, researchers have a good understanding of many specific problematic aspects of the interaction between users and recommendation algorithms, which cumulatively provide a compelling explanation for why rabbit holing is likely to arise. This is also supported by many internal documents from Meta which recognize the "rabbit hole" phenomenon or related concepts.

151. Given the scope of my report, I limit my discussion of rabbit holing to its algorithmic component although it has a psychological component as well. My key assumption regarding the psychological component is that some users can develop a behavioral proclivity to consume certain content if it is presented in a certain way, even though they do not want to do so,

would not actively seek it out, and would regret consuming it. This may arise due to a dissonance between the user's "System 1" thinking and "System 2" thinking.[93]

152. Developers of recommendation algorithms are well aware of the tendency of these algorithms to inundate users with a single type of content and, as a result, to suck some of its users into rabbit holes.

153. "Carol's journey to QAnon" refers to a 2019 internal user account study on the Facebook platform whereby a fake user profile of a 41 yr old southern female was created to understand how the ranking and recommendation evolved based on a new user profile with modest inputs regarding interest.[94] An internal report on the findings notes that "[w]ithin 2 days" the platform was suggesting "more exclusively right wing content, and beg[an] to lean towards conspiracy content."[95]

154. Meta has explicitly acknowledged the phenomenon of rabbit holing. Around 2020, the company created a project called Rabbithole to address the problem.[96] However, based on the documents I reviewed, I did not find any evidence that this team was able to devise or deploy effective solutions to the problem. As of 2021, Meta's documents acknowledged that "repetitiveness of content" and "over saturation of trends and audios" is a "major pain point" of Reels compared to TikTok.[97] A 2022 document lists and prioritizes various problems related to teen wellbeing on Instagram.[98] Teens rabbitholing into SSI [suicide and self-injury] and ED [eating disorder] content is the lowest on the list of 24 problems ranked by priority. "Current product investment" is listed as "low" for this problem.

---

[93] Kahneman. Thinking, Fast and Slow. Farrar, Straus and Giroux 2011. Apr. 2, 2013.

[94] META3047MDL-207-00033754.

[95] *Id*.

[96] META3047MDL-003-00043614; META3047MDL-049-00000126; META3047MDL-047-00990131; META3047MDL-072-00176016.

[97] META3047MDL-014-00405308.

[98] META3047MDL-034-00161622.

155. If the specific pattern of spiraling behavior involved in rabbit holing impacts only a small number of users, no matter how severely, it will not show up in aggregate statistics of problematic use. To the extent that Meta has made efforts related to mental wellbeing and safety aimed to improve experiences broadly across all users or some large subset of users such as teens, those efforts may not have impacted users affected by rabbit holing. What was distinctive about the Rabbithole project was the recognition of this phenomenon and the attempt (though under-resourced and unfinished) to identify this subset of users and improve experiences for them.

156. A standard feature of recommendation algorithms is called "diversity" and it has the potential to mitigate rabbit holing. This is closely related to the "exploration" aspect described above. However, the strength of the diversity feature implemented by platforms is often insufficient. For example, in the aforementioned Wall Street Journal investigation of TikTok, for a bot user that was programmed to engage with only one type of content, 93% of the recommendations were of this type of content and only 7% were other types of content.[99] The difficulty is that the algorithms can only discern engagement, not user's internal experience of that engagement, which could be fulfillment, regret, delight, anger, and so forth. While too little diversity leads to rabbit holing, too much diversity reduces engagement, which is the primary goal social media platforms are targeting.

157. Rabbit holing can be avoided through design and engineering choices, but Meta has not implemented such changes, as described below in the section on failure to adopt alternatives.

E. Meta's product testing practices encourage addictive design

158. Given the phenomena just described, one might expect that affected users would, over time, leave the platforms—which, in turn, would incentivize the platforms to change their algorithms to mitigate the problems. However, as I discuss in this section, the incentive structure is more complicated given the nature of product testing at these companies.

---

[99] Barry et al. How TikTok Serves Up Sex and Drug Videos to Minors. Wall Street Journal. Sept. 8, 2021. https://www.wsj.com/tech/tiktok-algorithm-sex-drugs-minors-11631052944.

159. As discussed earlier, changes to algorithms are rigorously tested to assess their impact on core metrics of interest. Suppose a proposed change to the algorithm increases the addictiveness of the app. In short-term experiments, this will increase key metrics, because it keeps users on the app for longer. Thus, engineers will be able to make a strong case for deploying the change. The metrics themselves do not necessarily provide a way to distinguish between changes that merely increase addictiveness and those that increase long-term user value.[100]

160. Suppose that the increased addictiveness leaves users unsatisfied, and hence more likely to quit using the app over a period of a few months. In theory, the developer can discover this negative impact on long-term user value due to a drop in users in a holdout experiment. So one might think this would cause the developers to engage in remediation techniques.

161. In practice, though, there are many hurdles. Users are a finite resource and there is a huge demand for experiments. Long-term experiments take up proportionally more of the holdout user budget than short-term experiments, so they are not always conducted.

162. Even when long-term experiments are conducted, there may be a single holdout group of users from whom all changes to the app over a given time period are withheld, of which there may be dozens. Thus, even if there is a drop in users among those exposed to the changes, there may not be any way to attribute the increase in addictiveness to a specific change.

163. Finally, even if a set of algorithm changes brings negative value to a particular user in the form of addiction, even in the long run it might not result in a decrease in their usage of the app, due to the very nature of addiction.

164. There is extensive evidence that Meta's pursuit of engagement in its algorithms came at the expense of long-term user value for the reasons described above.

[100] Illustrating this point, one Meta document states that "Consumer Value reflects the value that we provide to the viewer, which is measured by Consumer DAP." DAP, of course, is simply a measure of "Daily Active People"—how many people are using the platform. That metric may or may not correlate to the actual "value" that consumers experience from their use of the product, though it does correlate to the revenue that the platform is able to generate from their usage. META3047MDL-014-00405878 at META3047MDL-014-00405878.

165. In 2019, Tom Cunningham, who worked on the Core Data Science team at Facebook, wrote an internal report discussing three gaps between engagement and user value.[101] First, he analogized the Facebook experience to broadcast television shows that attempt to maximize advertising revenue by stretching out the viewing time, such as by having recaps and cliffhangers, even if it makes the overall experience less valuable. Second, he explained that Facebook creates an incentive for clickbait and misleading posts, and that the company's product testing practices are not adequate to demote such posts because "many features are evaluated just on short-run engagement — if we mislead users it'll surely cause a lack of trust down the line, but that can be slow and cumulative, and it's very hard to detect with just short-run experiments." Finally, when users have self-control problems (that is, their System 1 behavior diverges from their System 2 preferences), they end up spending more time on Facebook than they like.

166. Mr. Cunningham also wrote:

> I sometimes wonder what the alternative is to making decisions based on AB-tests of engagement metrics, and I scratch my head, then I remember that just over the road WhatsApp built a 1Bn user product without running any tests, they made all their product decisions just based on their perceptions of user quality. … Of course WhatsApp cares about growth and engagement, but the fact they didn't run experiments until recently implies that they don't think short-run AB-test measures were very valuable measures of long-run value."

167. In 2020, [redacted] expressed a worry that Instagram's ranking team would "optimize the shit out of [sessions, a short term metric] and game it too much", and notes that this had previously happened with Time Spent, another short-term metric.

168. Once a platform is locked into a culture of all changes needing to be justified by experiments, it creates a strong misalignment of incentives between engineers and users. Engineers who focus on improvements that improve short term metrics will appear far more productive and will be rewarded over those whose changes to the code take weeks, months, or years to bear fruit.

[101] META3047MDL-020-00575591; META3047MDL-037-00028264.