Filed Under Seal

EXHIBIT D

EXPERT REPORT OF

Arvind Narayanan, Ph.D.

on behalf of the State of New Mexico

July 11, 2025

43. There are many kinds of models, and social media platforms contain a multitude. “Neural networks” are the most common. Neural networks consist of simple computational units arranged in layers, with each layer transforming its input and passing it to the next. While these models function as networks, they are stored as sequences of numbers. Some models are tiny, consisting of only a handful of numbers, while others have billions. Importantly, in most cases, these numbers are not intended to be human-interpretable and indeed cannot be interpreted by humans. That is one reason why these types of algorithms are sometimes referred to as black boxes. If one were to ask the robot programmers, “*why* did your model tell the robot to take a bigger step to walk over a specific crack in the sidewalk?” the programmers would not have an easy way of answering that question—other than to explain the goal of the algorithm and how the training data was collected and processed. Reverse-engineering the actual numbers that dictated the specific movement of the robot would be an extreme challenge if not impossibility.

B. Social media recommendation algorithms

44. In my opinion, the recommendation algorithms behind the major social media applications, including those of Meta, are strikingly similar to each other. The differences between them, while sometimes significant for users, are relatively minor from a computer science perspective.[3] Therefore, I will first describe social media recommendation algorithms generally before analyzing Meta’s algorithms.

45. In the context of social media, the components in Figure 1 are typically implemented as follows:

- **Goal**: As discussed in the context of the robot example, a key lever of control a developer exercises over the system is by mathematically specifying the objective for the algorithm. The choice of the objective determines what the resulting model (and algorithm) aims to optimize. In the example, the programmers set a goal of getting the robot to walk (or, more precisely, optimizing the likelihood that the robot walks). Social media companies

[3] Narayanan. TikTok’s Secret Sauce. Knight First Am. Inst. Dec. 15, 2022.

also articulate a goal for their recommendation algorithms. The most common objective for social media platforms is to maximize engagement. As I discuss below, Meta has set the goal of recommending posts to users that will maximize the likelihood of those users' engagement with the platform (where engagement refers to any action taken by a user in response to a post). Just as in the robot example, this goal is a choice made by humans who work for the companies—it is exogenous to the algorithm or model itself.

- **Training data**: A core component of the training data consists of all users' past interactions with posts that they have seen. It is important to grasp the volume of engagement data tracked by platforms. Here is a simple calculation: If a user spends an hour a day on a short-form video app for four years, the average time spent per video is 20 seconds, and they skip half the videos, that means that they interact with about six videos per minute (including the videos skipped, which is valuable information to the platform). That means that over the four-year period, the platform would collect interaction records on over half a million videos for that single user. This number must be multiplied by billions of users to arrive at the volume of the entire dataset for the largest video apps. Training data for social media recommendation algorithms, like Meta's, include individual user information such as user demographics, history of past engagement, interests, and contextual data, such as the device used by the user.
- **Training process**: The training process is the stage at which the developers' objective is applied to the training data, so that the model can learn. Much of the academic research and commercial innovation in social media algorithms goes into figuring out how to make the training process more efficient and effective at learning the patterns in the training data. These technical details are less relevant for our purposes.
- **User, models, and algorithm**: A user interacting with a social media platform (a user scrolling through their feed would constitute the "input" in the top right corner of the figure) is really interacting with a set of machine learning models (illustrated in the middle right of the figure). Those models, together with some manually specified logic,

together comprise the social media platform's algorithm. This is usually called a recommendation algorithm, because the algorithm recommends posts for the user to engage with.

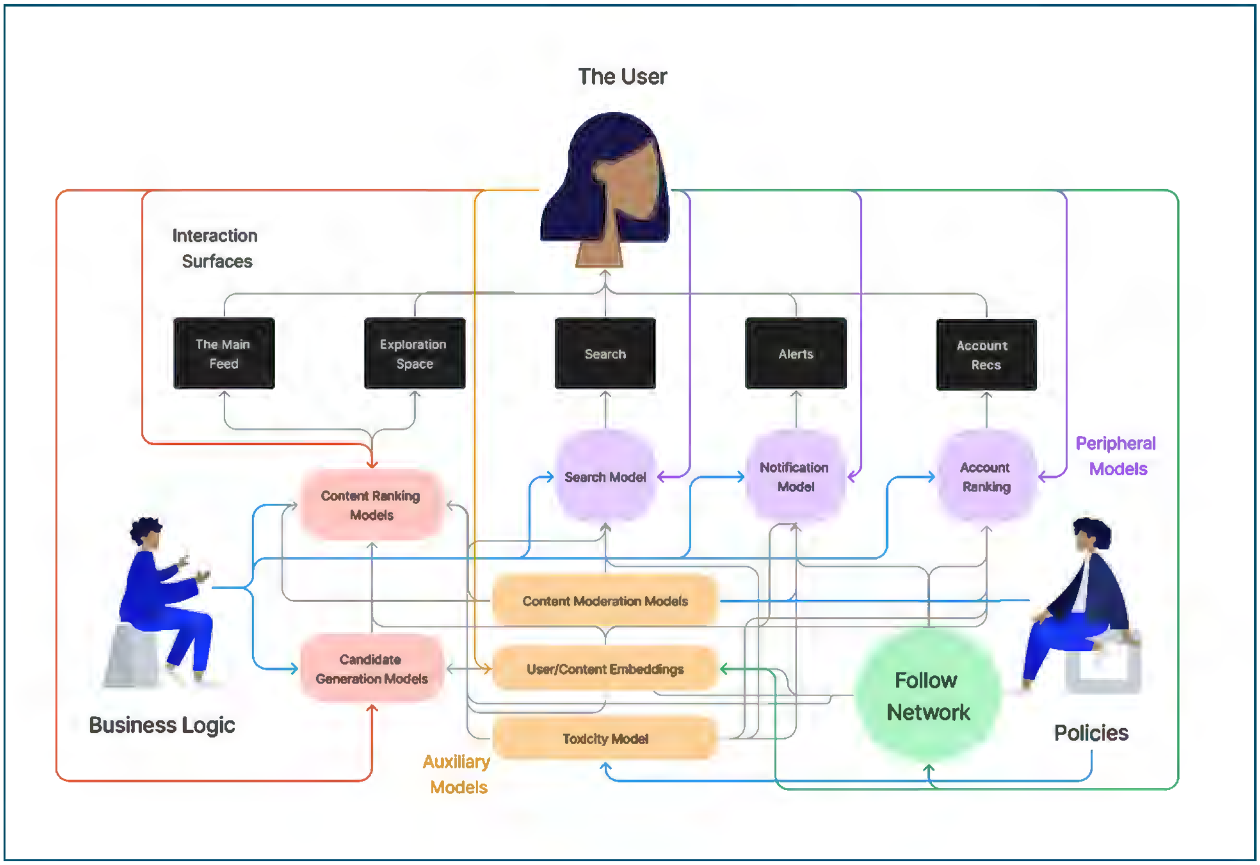
- **Decision**: Based on the user's characteristics as interpreted by the model, the ranking algorithm will output a decision—what posts to show the user in what order when they are interacting with the application.

46. Importantly, and just as with the robot, this design allows developers to focus on the "what" and ignore the "how." Developers specify a goal such as maximizing the user's likelihood of engagement with the posts that they see, but they do not manually program any judgments for which kinds of posts should be shown to users in order to satisfy that objective. To put the matter plainly, there is no point in the process at which developers say about any specific user or set of users, "they have shown an interest in puppies so we should program our algorithm to show them more puppies." Such decision-making logic exists only within the black box of the algorithm itself and consists of a complex sequence of numbers that is not human-interpretable.

47. I now situate social media ranking algorithms in the broader context of the platforms as a whole. Social media applications or websites have many surfaces that users can interact with. Surfaces refer to different tabs or components of applications. Scrolling through a feed of posts is usually the primary surface (and it is the one referred to in the preceding discussion), but there are others such as searching for specific content and receiving notifications. Each surface has its own algorithm that determines what to show, with one or more models behind it. This is one reason for the multiplicity of models.

48. The figure below shows an illustrative design of social media algorithms. It was created by Kristian Lum and Tomo Lazovich, former machine learning engineers at Twitter.[4]

Figure 2: Illustrative design of social media algorithms



49. The row immediately underneath the user illustrates the different interaction surfaces mentioned in the preceding paragraph. In addition to the user's main feed, this can include tabs for exploration, search, alerts or notifications, and recommendations to follow other accounts. The next row illustrates the various machine learning models that power one or more of these surfaces. For instance, in this diagram search has its own model, notifications have their own model, and the main and explore tabs share a ranking model.

50. The figure usefully illustrates another important concept that governs social media ranking algorithms. Ranking models have two main stages, **candidate generation** and **content**

[4] Lum & Lazovich. The Myth of The Algorithm: A system-level view of algorithmic amplification. Knight First Am. Inst. Sept. 13, 2023.

ranking. The goal of candidate generation is to select roughly a few hundred or a few thousand posts that might potentially rank high for that user from among the millions or billions of posts on the platform. Examples of potentially good candidates are posts from accounts the user follows, posts that are highly popular, and posts in the user's geographic proximity. In the ranking stage, previously trained machine learning models compute a score for each candidate that is personalized to the user, based on the optimization objective chosen by the developers. The top scoring posts are then shown to the user. The benefit of a two-stage pipeline is to achieve a good tradeoff between speed and accuracy. The first stage is fast but less accurate. The second stage is slower and more accurate, but has far fewer posts to consider.

51. Finally, the figure usefully situates social media ranking models alongside content moderation models, correctly depicting these as separate functions governed by different logic. Ranking is governed by "business logic;" again this is the goal set by a company (generally prioritizing engagement-maximization) that governs the decisions made by the ranking algorithm. Content moderation and related auxiliary models are governed, instead, by "policies"—the decisions regarding what sort of content will and will not be allowed on the platform, often memorialized in lengthy, detailed, and publicly-available community standards or guidelines.

52. It bears repeating that the recommendation algorithm code created by the developers is essentially content agnostic. The question of what to show a user is delegated to the models and algorithms, which are trained to select whatever the data say will optimize the business objective of increasing engagement. By contrast, the content moderation algorithms, which I turn to next, encode a company's judgments or preferences about content to be prioritized or demoted. For example, posts judged to be from "authoritative" sources or posts of "high quality" might be prioritized, posts judged to be "spam" or "borderline" policy violating might be demoted, while posts judged to be outright policy violating might be deleted.

53. Importantly, while the ranking model interacts with the content moderation model in the course of selecting what posts to present to a user, it operates as an independent system.

Defects with the recommendation system cannot be solved by changing or improving the platform's content moderation algorithm or its content moderation policies.

C. Content moderation algorithms

54. As noted, most social media platforms publish detailed policies specifying types of content that users are not allowed to post. For example, Facebook's "community standards" document is 18,000 words long. Examples of categories that many platforms prohibit include pornography, illegal content, or content that promotes self-harm. These policies have evolved in scope and complexity throughout the history of social media, often driven by complaints from users, journalists, and governments about the spread of content perceived by some to be problematic.[5]

55. Enforcing these policies is called content moderation. It is done by a combination of AI algorithms and people. The specific decision flows for content moderation can and do vary across platforms and within platforms for different kinds of policy violations. Generally speaking, when content is posted, it is scanned by one or more content moderation algorithms.

56. There are two types of content moderation algorithms. The first is called fingerprint matching, which is capable of detecting posts (typically images or videos) whose specific contents are *already* known to be violative. This is done by comparing the "fingerprints" of the new post against those of known-violative posts stored in a database. Matches may result from identical copies or copies with small modifications such as cropping an image. Since this technology is incapable of detecting newly created violative content, its use is limited to certain categories of policy violations, mainly child sexual abuse imagery, copyright violations, and terroristic content.

57. The second, much more broadly applicable type of content moderation algorithm is based on machine learning and uses "classifiers" to check for policy violations. Platforms

---

[5] Roberts. Behind the Screen: Content Moderation in the Shadows of Social Media. Yale University Press. Jun. 25, 2019.; Gillespie. Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press. Jun. 26, 2018.

- 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.

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] METANMAG-010-00154293; 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. [redacted]

76. [redacted]

[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.

[redacted]

[redacted] [redacted]

Figure 3: Excerpt from METANMAG-002-00633168 at METANMAG-002-00633179

VM base score = 1.5 * p(like) + 1.6 * p(reshare) + 3 * p(profile visit) + 3.5 * p(save) + 9 * p(follow) + 20 * p(comment) + 3 * p(transferred cover click) + 0.4 * p(transferred cover video complete) + 1.2 * p(video complete) + 0.2 * p(swipe forward) + 0.9 * p(repeat) + 1.3 * E(log time) - 1.0 * p(skip) + [20 * p(use audio) + 15 * p(save audio) + 5 * p(use effect) + 0.03] *p(producer) boost

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. [redacted] 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, [redacted], 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 Tom Wynter 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] METANMAG-005-01392553.

[24] METANMAG-002-01529600; 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] METANMAG-011-00034219; *see also* METANMAG-005-00517359 ("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] METANMAG-002-00769186; Volichenko Dep. 50:5-9 (Q: "Were any of the Instagram critical metrics related to growth?" A: "Yes, yes, absolutely.").

[31] METANMAG-002-00769186; Meta's Resp. to Interrog. 15; METANMAG-007-00215718.

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

[33] METANMAG-002-00632684.

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 Tom Wynter 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] METANMAG-002-00279446.

[35] METANMAG-055-00145763; METANMAG-062-00015080.

[36] METANMAG-005-01230566.

[37] METANMAG-005-01392310; *see also* METANMAG-011-00141032 and METANMAG-005-00517359.

> 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. GVM halves and doubles the weights on every major relevance term, 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.

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.

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] METANMAG-011-00034219; METANMAG-005-01392553; METANMAG-005-01392310; METANMAG-011-00141032; METANMAG-005-00517359.

[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] METANMAG-002-01529600.

[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] METANMAG-002-00633168 (emphasis in original).

recommendation surfaces such as Explore.[123] By December 2020, internal notes from an employee indicate that SSI was still only "on the verge of becoming a priority pillar because of vulnerabilities on IG."[124] I have seen no indication that these various vulnerabilities were ever disclosed to the public, despite Instagram's promise in February 2019 to "remove all graphic self-harm images from [its] site."[125] Again, Ms. Jayakumar acknowledged at her deposition that "there's a disconnect between the suicide and self-injury policy and the aggregated content that actually gets recommended by Instagram's algorithm."[126]

186. In the absence of effective classifiers, it is simply infeasible for platforms to effectively enforce their advertised policies at the scale at which they operate. Simply put, human labor is expensive on a scale of millions or billions of posts.[127] I have seen no indication that these limitations were disclosed by Meta.

VIII. Failure to adopt alternatives

A. Content moderation cannot solve the problems of recommendation algorithms

215. As described earlier in this report, platforms deploy content moderation algorithms to tackle various categories of problematic content. One might be tempted to think, then, that improving platforms' content moderation is the best way, or the only way, to reduce negative user experiences on the platform. That perspective is mistaken. As I have explained throughout this report, Meta's content moderation algorithms and its recommendation algorithms operate as separate systems. Changing or improving content moderation will not curb the tendency of recommendation algorithms to discover and exploit users' behavioral proclivities, will not reduce or eliminate addictive usage, and will not prevent users from being algorithmically directed down

[123] METANMAG-002-00044234.

[124] Jayakumar Dep., Ex. 19.

[125] Jayakumar Dep. 212:25-215:6.

[126] Jayakumar Dep. 222:6-12.

[127] Narayanan & Kapoor. AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference. Princeton University Press. Sept. 24, 2024.

rabbit holes. Also, content moderation systems constantly lag behind known harms. As new categories of harmful content crop up, integrity efforts are already way behind. Further, because content moderation applies at the level of individual posts, it cannot anticipate and address what happens when posts (each of which may not be policy-violating) are aggregated by the recommendation algorithm.

187. For all these reasons, and as discussed in this section, the problems with recommendation algorithms cannot be solved through different or better content moderation policies.

188. Rabbit holes are not necessarily caused by certain types of content. While some content can be considered harmful even in isolation, even benign content can lead to rabbit holes if it takes over a user's feed due to the design of the recommendation algorithm. What may not be harmful on its own can be negatively influential in the aggregate, and content moderation operating on one piece of content at a time cannot determine that certain kinds of content might be problematic in the aggregate, as part of a homogeneous feed.

189. In 2021, Vaishnavi Jayakumar, Instagram's head of safety and well-being policy, explained that "the challenge is that when you zoom in, any one of those photos probably wouldn't be triggering. But in an explore grid with all these photos stacked against one another, it's pretty overwhelming".[128]

190. Further, even when there is something problematic about the content, it is highly contextual and user-dependent. What is problematic for one user may not be so for another. Unfortunately, content moderation operates at a global level and cannot account for user differences and other nuances.

191. Meta belatedly recognized this problem. In 2020 there was an internal proposal for "integrity personalization" that would attempt to tailor sensitivity thresholds to each user in the

[128] METANMAG-003-00868309.