Ashley M. Simonsen (Bar No. 275203)
COVINGTON & BURLING LLP
1999 Avenue of the Stars
Los Angeles, California 90067-4643
Telephone: + 1 (424) 332-4800
Facsimile: + 1 (424) 332-4749
Email: asimonsen@cov.com

Attorneys for Defendants Meta Platforms, Inc. f/k/a Facebook, Inc.; Facebook Holdings, LLC; Facebook Operations, LLC; Facebook Payments, Inc.; Facebook Technologies, LLC; Instagram, LLC; Siculus, Inc.; and Mark Elliot Zuckerberg
Additional counsel listed on signature pages

UNITED STATES DISTRICT COURT
FOR THE NORTHERN DISTRICT OF CALIFORNIA
OAKLAND DIVISION

IN RE: SOCIAL MEDIA ADOLESCENT ADDICTION/PERSONAL INJURY PRODUCTS LIABILITY LITIGATION	MDL NO. 3047 Civil Case No. 4:22-md-03047-YGR: Honorable Yvonne Gonzalez Rogers
THIS DOCUMENT RELATES TO: ALL ACTIONS	**META DEFENDANTS' SUPPLEMENTAL RESPONSES AND OBJECTIONS TO PLAINTIFFS' FIRST INTERROGATORY**

Defendants Meta Platforms, Inc., Facebook Payments, Inc., Siculus, Inc., Facebook Operations, LLC, and Instagram, LLC, (collectively, "Defendants" or "Meta"), by and through their attorneys, Covington & Burling, LLP, submit their supplemental responses and objections to Plaintiffs' Interrogatory No. 1 ("Interrogatory") served on May 16, 2024.

Meta's responses to the Interrogatory are made to the best of its current knowledge, information, and belief. Meta reserves the right to supplement or amend any of its responses should future investigation indicate that such supplementation or amendment is necessary.

INTERROGATORIES

INTERROGATORY NO. 1:

Please itemize and describe each Level 2 integrity classifier used by You to track or classify content on the Facebook Platform and/or Instagram Platform. For purposes of this request, "itemize and describe" means to identify the classifier by both its plain language names and alphanumeric code, the date You created the classifier, the Identity of the person who created the classifier, the purpose for which the classifier was created, the characteristics of content the classifier is or was designed to identify, the probability thresholds associated with the classifier at any given time, the related actions taken at each such threshold (including automatic removal or manual review of content) at any given time, and the weight given to each classifier at any given time.

FOURTH SUPPLEMENTAL RESPONSE TO INTERROGATORY NO. 1:

Meta objects to the use of the undefined phrase "integrity classifier," making this Interrogatory vague and ambiguous. Meta construes "integrity classifier" to be classifiers that review and detect content for potential violations of Facebook's Community Standards and Instagram's Community Guidelines (collectively, "Policies"), to the extent they pertain to the following policy-violating content, also known as "harm areas": suicide and self-injury ("SSI") and eating disorder ("ED"), bullying and harassment, graphic and violent content, adult nudity, and child safety. Each of these classifiers is described in further detail in the Response below.

Meta further objects to the use of the undefined phrases "plain language names" and "alphanumeric code," making this Interrogatory vague and ambiguous. Meta construes "alphanumeric code" to refer to the 16-digit numeric code assigned to a classifier for their identification.

Meta further objects to the use of the undefined phrase "probability thresholds," making this Interrogatory vague and ambiguous. Meta construes "probability thresholds" to mean the threshold value assigned to each of the harm areas—for example, SSI, ED, bullying and harassment, graphic and violent content, adult nudity, and child safety—against which Confidence Level ("CL") is measured to determine whether or not content should be removed, as further described below.

Meta further objects to the use of the undefined phrase "weight given to each classifier," making this Interrogatory vague and ambiguous. Meta construes "the weight given to each classifier" to mean "predictive integrity value (pIV)," based on which Meta prioritizes the order in which content is reviewed by the human reviewers.

Meta further objects to this Interrogatory to the extent it seeks "the Identity of the person who created the classifier," to the extent it asks for a single person who created the classifier, making this Interrogatory ambiguous. Because classifiers are created through a collaborative effort between Meta's Product, Operations, Content Policy, Product Policy, and Safety Policy teams, its creation cannot be attributed to a single individual.

Meta further objects to this Interrogatory to the extent it seeks "the date You created the classifier," to the extent it asks for a single date on which any particular classifier was created, making this Interrogatory ambiguous. Each classifier was developed through an iterative process and through a collaborative effort between Meta's safety and integrity teams, such that any particular classifier could have multiple historic iterations (and, indeed, may have existed prior to being referred to as a "classifier"). In any event, Meta will construe "the date You created the classifier," to mean the date on which the latest iteration of a classifier is launched.

Meta further objects to this Interrogatory as overbroad, unduly burdensome, and not proportional to the needs of the case because this Interrogatory seeks information "at any given time," as these thresholds values given to each classifier are re-assessed periodically and subject to change at any

given time. Meta further objects to this Interrogatory on the grounds that it is unduly burdensome and overly broad insofar as it contains numerous subparts, and on that basis constitutes more than a single interrogatory.

After a reasonable investigation, Meta is unable to ascertain whether or not classifiers are assigned "Levels" such that there are "Level 2" classifiers. Meta will supplement this response if it identifies this information.

Meta further objects to Plaintiffs' request on September 3, 2024, as vague, overbroad, unduly burdensome, and not proportional to the needs of the case to the extent it seeks information as to whether any classifiers created by Well-Being, Social Impact, and Safety teams are "Integrity Classifiers."[1] Plaintiffs did not seek this information in their First Set of Interrogatories, and the requested information is outside the scope of and irrelevant to this litigation, and Plaintiffs' demand is thus not proportional to the needs of this case. As explained in Meta's June 28, 2024 response, the only Integrity Classifiers relevant to this litigation are those that address the harm areas identified in the response. To the extent Plaintiffs seek information as to which individuals or teams have created the current iterations of each relevant Integrity Classifier, Meta will supplement its response if and when it learns this information.

Meta further objects to Plaintiffs' request on September 3, 2024, for "the recall rate for each of the classifiers in each of the harm areas listed in Meta's Responses and Objections," Letter from Nelson Drake to Kurt Calia at 7 (Sept. 3, 2024), because it is vague and because Plaintiffs did not seek this information in their First Set of Interrogatories. To the extent any recall rate for any Integrity Classifier in the identified harm areas are relevant to Plaintiffs' request for information on those classifiers (including plain language names, alphanumeric codes, dates of creation, the individuals or teams responsible for their creation, and

[1] Letter from Nelson Drake to Kurt Calia at 2 (Sept. 3, 2024).

the threshold values for classifiers to take action on content), Meta will supplement its response if and when it discovers this information.

Meta further objects to Plaintiffs' request on September 3, 2024, for Meta to explain "the purpose of borderline classifiers" as vague, overbroad, and duplicative. Letter from Nelson Drake to Kurt Calia at 8 (Sept. 3, 2024). As explained in Meta's initial response:

> A "borderline classifier" is an artificial intelligence tool that reviews and detects content on Facebook and Instagram that is not strictly policy-implicating, but nonetheless considered potentially harmful to users. For example, a borderline graphic and violent content classifier would consider content of puss-filled or oozing wounds, people undergoing surgical operations, fight videos, animal abuse, and fictional violence, not strictly policy-violating, but as "borderline content." Where a piece of content is determined to be "borderline content," then that content is not removed, but will be downranked.

MDL Rog. 1 Resp. at 13.

Subject to and without waiving these objections, Meta provides below a Response which generally describes its current classifier system and notes that variations may exist across Instagram and Facebook. The description contained in this Response is therefore not exhaustive or inclusive of every detail comprising every process that may support the classifiers on Facebook or Instagram, and indeed more detailed information would be found in Meta's source code, which is maintained in Meta's source code repositories. Meta's source code will be made available for inspection by Plaintiffs subject to the entry of a suitable Source Code Protective Order.

I. Classifiers Generally

A "classifier" refers to an artificial intelligence tool or algorithm that automatically reviews and categorizes content or accounts on Facebook and Instagram for potential further action, including but not limited to content removal, downranking, demoting, filtering, age gating, the addition of warning screens or captions, or non-action. Classifiers evaluate the content or accounts potentially against Meta's Policies,

and can make determinations about whether a particular piece of content or account is considered "policy-violating," and otherwise form the basis for potential actioning of content, as further explained below.

In general, Meta's content moderation efforts begin with the development of its policies in consultation with experts around the world. Those policies drive the identification of prohibited content and ultimately inform what type of content and accounts are excluded and removed from Meta's Platforms. Meta's safety and integrity teams, responsible for scaling the detection and enforcement of Meta's policies, build machine learning models to predict whether content may violate one or more of these policies. Meta's content moderation efforts involve the use of several systems to detect prohibited content and a network of human reviewers.

Meta's safety and integrity teams develop machine learning models that can look for certain signals to recognize what is posted on its platforms and evaluate whether it may violate one or more of Meta's policies. These tools are called "classifiers." Classifiers assess content—including text, video, and photos—on both Facebook and Instagram. As part of the development of a classifier as well as its ongoing maintenance, Meta trains the classifier using labeled data, working to improve its detection accuracy by introducing more content of a specific type relevant to that classifier. This improves the classifier's performance for that specific type of content.

This artificial intelligence (AI) system runs in the background and automatically monitors Meta's Platforms for any policy-violating content—including comments, posts, and pictures—before users have an opportunity to see or report content that is concerning to them. When clearly prohibited content is detected, the AI system automatically removes it; the vast majority of content removed for being policy-violating is removed before anyone sees it. In that sense, if the content is obviously actionable, the system is self-executing. If the AI system cannot determine whether content is prohibited, then the content is routed to a team of human reviewers. Meta has a network of over 15,000 human reviewers who are trained to review individual content and determine whether such content violates Meta's policies. The human

reviewers' decisions also inform the AI detection of policy-violating content, so the AI grows more sophisticated over time. Classifiers are constantly refined.

As noted above, the AI detection system runs proactively and automatically, with technology working behind the scenes to remove prohibited content, often before any user sees it. Meta also continuously trains its artificial intelligence system for more accurate and nuanced detection of prohibited content. In addition to making improvements to classifiers using human review mentioned above, Meta incorporates additional signals into its classifiers as they are developed. For example, Meta enhanced its technology to enable classifiers to take all the components of a post—including text, pictures, videos, and comments—into consideration to better understand the meaning of the post as a whole, instead of looking at only one component part in a vacuum.

A. Confidence Levels and Threshold Value

Classifiers assign "Confidence Levels" ("CL") to each piece of content it reviews. A CL is a statistical value that refers to the probability with which a piece of content is policy-violating (*e.g.*, p-value = 0.95), whereby the higher the CL value, the more likely it is that the content is policy-violating. The reverse is also true: the lower the CL value, the less likely it is that the content is policy-violating.

CL is measured against a "threshold value" to determine what type of action (*e.g.*, content removal, human reviewer queuing, soft action), if any, to take on that content. "Soft action" refers to an action that Meta takes on a piece of content short of removal to reduce its exposure to Meta's users, such as downranking or demoting, filtering, age gating, adding warning screens or captions, non-recommendations, in-feed recommendation (IFR) filtering, or search engine results page (SERP) filtering. Not all forms of soft action—*i.e.*, action on content short of removal—are used for each classifier or for each harm area. If a piece of content is assigned a CL that exceeds a certain threshold value, then that

content is deemed to violate Meta's content policy and is automatically removed.[2] If a piece of content is assigned a CL that is lower than the removal threshold value, but higher than the threshold value that would trigger soft action, then that content is routed for soft action and possible human review for additional assessment. Where the confidence level is relatively low, content is not sent for human review and is instead downranked and/or age gated, or subject to some other form of soft action.

Different threshold values are set for different types of policy-violating content (*e.g.*, SSI, ED, bullying and harassment, violent and graphic content, adult nudity, and child safety) by the teams in charge of creating, monitoring, and updating each harm-specific classifier. The threshold values may also vary across Facebook and Instagram. The threshold values are re-assessed both each time classifiers are updated and on an ongoing basis in order to maintain a consistent accuracy bar. Threshold values may be written as a decimal value or as a "p-value" (e.g., 0.5 or p50) when the threshold value is equivalent to the targeted marginal precision value of the given classifier.[3] The threshold values for soft actions vary, both over time and between classifier use cases. These are set by surface teams.

In the case where content violates multiple Community Standards, if any of the classifiers meet the threshold for deletion, the content will be removed. Additionally, multiple classifiers can trigger soft actions at the same time, resulting in an even stronger demotion. There are also some "multi-class" classifiers that can identify multiple different policy violations.

---

[2] *See* discussion *infra* Part II.

[3] Precision is a metric Meta uses to assess the accuracy of integrity classifiers in correctly identifying and acting on potentially policy-violating content on Facebook and Instagram. Precision is measured by first taking an unweighted sample of the content that integrity classifiers have detected and assigned a confidence level. Next, humans review the decisions made by the integrity classifiers and label the decisions as true positives (TP) or false positives (FP). Next, precision is calculated by solving for the total number of true positives divided by the total number of true positives and false positives (*i.e.*, TP / (TP + FP)). Precision can be calculated either cumulatively or marginally. If cumulative, the precision metric captures the chance of there being a true positive among all samples with a score *above* a threshold confidence interval. If marginal, the precision metric captures the chance of there being a true positive among all the samples with a score *at or slightly above* a threshold confidence interval.

After a reasonable investigation, Meta determined the alphanumeric codes, the specific threshold value associated with content removal, human reviewer queuing, and soft action as stated in the Response below. However, because the threshold value range changes at different times, for different harm types, and for different actions, Meta has provided current information related to its classifiers and states that historical information concerning its classifiers is embedded in source code, which Meta will make available for inspection upon entry of a suitable Source Code Protective Order.

B. Predictive Integrity Value ("pIV")

Once a piece of content is enqueued for human review, then that piece of content is assigned a pIV, which is used to determine which piece of content should be reviewed first by the human reviewers. pIV is calculated based on factors including likelihood of violation, virality, and violation severity, as further explained below:

First, Meta uses a range of signals, including Whole Post Integrity Embeddings ("WPIE"), a basic classification layer that helps predict whether a piece of content violates any integrity problem, and problem-specific classifiers to predict how likely a piece of content is to be violating for any problem area. WPIE assesses the basic details of the content being detected to determine what kind of violation may be at issue.

Second, Meta assigns a severity weight to the piece of content based on the problem area it was most likely to violate for, *i.e.,* if we suspect the content represents a very high severity problem it will get a high severity weight, while if it's high or medium severity the assigned weight factor will be lower. Meta does not assign "weight" to or otherwise "rank" classifiers. Rather, content is ranked for human review, according to its severity. Meta's Multi-Armed Bandit ("MAB") system assesses the content's severity level. MAB assigns content a severity score based on its "predictive severity level."

Third, there are systems that evaluate content virality. The Virality, Estimations, Prediction, and Applications ("VESPA") system assesses the piece of content's virality by counting each time a user

views it for a meaningful amount of time. VESPA assigns a "predicted virality" value to a content. "Predicted virality value" reflects how widely viewed a particular piece of content is expected to be. VESPA used to use WPIE as an input feature to learn from, but VESPA no longer utilizes WPIE. The High Risk Early Review Operations (HERO) system is distinct from VESPA, though it may review content flagged by VESPA, and is an end-to-end system used for human review to remove high viral violating content.

Finally, based on these evaluations of harm classification, severity, and virality, content is assigned a pIV to prioritize the piece of content for review. A priority formula then orders content for review based on pIV. The highest pIV tasks move to the top of the queue so that the most viral and potentially harmful content gets reviewed sooner. The queue is designed to be dynamic, such that it continually brings the highest priority items to the top of the queue for review. Some issues (for example, Child Sexual Abuse Material ("CSAM")) are always prioritized because they are "Extremely High Severity" content—and even if they have really low virality, the severity rating will carry it to the top of the queue.

C. Live Content Moderation

Live content moderation also involves the use of classifiers. However, there is a relative lack of data associated with Live content (in contrast to other content) because Live is not as heavily utilized by users. [redacted] In contrast, regular videos uploaded to Meta's platforms can be analyzed by a classifier just once. Furthermore, because of the nature of live content (involving severe, rare, and immediate harms), moderation must be done near-instantaneously. Moderation of Live is also more complex because it requires its own custom classifiers.

Live content classifiers are newer relative to their static content counterparts, and they frequently result in content being enqueued for human review, except for a few high-severity harms. Generally,

enqueuing happens at around a 60-65% confidence level, with autodeletion when classifiers are at approximately above 85-90%.

II. Harm-Specific Classifiers

Classifiers run automatically and review content on Facebook and Instagram. Different classifiers review and detect different types of policy-violating content, also known as "harm areas." Those harm areas include, but are not limited to, SSI, ED, bullying and harassment, graphic and violent content, adult nudity, and child safety.

A. SSI Classifiers

Meta first implemented a classifier to action SSI content at least as early as February 2019, and Meta has continuously refined its SSI classifiers since then. The latest iteration of SSI classifiers were launched at different times in 2024.[4] SSI classifiers are intended to detect suicide and self-injury content on Facebook and Instagram that may violate Meta's Policies.[5] SSI classifiers consider the following content policy-violating under Meta's Policies and subject to removal: "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. . . . as well as real time depictions of suicide or self-injury."[6] However, not all policy-violating content is subject to removal, such as "[c]ontent about recovery from suicide, self-injury or eating disorders that is allowed, but may contain imagery that could be upsetting (such as a healed scar)."[7]

Meta removes content that encourages suicide, self-injury or eating disorders, including fictional content such as memes or illustrations and any such content which is graphic, regardless of context. Meta

---

[4] *See infra* Table 1.

[5] *Suicide & Self Injury*, Meta: Transparency Center, https://transparency.meta.com/policies/community-standards/suicide-self-injury/ (last visited Feb. 6, 2025).

[6] *Id.*

[7] *Id.*

also removes content that mocks victims or survivors of suicide, self-injury or eating disorders, as well as real-time depictions of suicide or self-injury. Content about recovery from suicide, self-injury or eating disorders that is allowed, but may contain imagery that could be upsetting, such as a healed scar, is placed behind a sensitivity screen."[8]

Meta uses its AI technology, machine learning, and image-based technology (including pattern-recognition signals, such as phrases and comments of concern) to proactively identify and take action on clearly violating SSI content. Meta refines its classifiers by providing both positive examples (actual SSI content) and contrasting negative examples (not actual SSI content—e.g., "I have so much homework, I want to kill myself"—so it can learn to distinguish the two.[9]

SSI Classifiers do not just review the content itself but instead may review other information—*e.g.*, comments to a post where a user is seriously at risk, the classifier may detect and consider comments from human commenters such as "Tell me where you are" or "has anyone heard from [user]?"[10]

Training automated technology to identify and distinguish between SSI content that is violative or non-violative is incredibly complex. This is particularly the case on Instagram, where much of the content is image-based and therefore the intended meaning substantially relies on subtle context. For instance, a picture of train tracks may have to do with travel (and therefore be innocuous content), or it may relate to SSI. Humans can quickly tell the difference between these images based on the context, but training AI to do so is challenging, particularly at scale.

The current threshold values for autodeletion and various soft actions on SSI content are listed below.[11] Content containing SSI is prioritized during human review because Meta has a vested interest

---

[8] *Id.*

[9] *How Facebook AI Helps Suicide Prevention*, Meta, https://about.fb.com/news/2018/09/inside-feed-suicide-prevention-and-ai/ (last visited Feb. 6, 2025).

[10] *Id.*

[11] *See infra* Table 1.

in preventing as much human harm as possible. Meta uses artificial intelligence to prioritize the order in which potentially violating SSI content is reviewed to ensure efficient enforcement of policies and to get resources to users in need quickly.

Table 1. SSI Classifiers and SSI Borderline Classifiers.

Plain-Language Classifier Name	Alpha-numeric code	Launch date of current iteration	Threshold value for autodeletion	Threshold value(s) for soft action(s)

SSI Classifier	f590828509	08/12/2024	0.94	Facebook: various enforcements >P50
				• EU Demotion on Newsfeed
				  ○ Suicide: 0.8
				  ○ Self-injury: 0.8
				• EU Deboost on Newsfeed
				  ○ Suicide: 0.3
				  ○ Self-injury: 0.4
				• Non-EU Demotion on Newsfeed
				  ○ Suicide: 0.5
				  ○ Self-injury: 0.5
				• Non-EU Deboost on Newsfeed
				  ○ Suicide: 0.3
				  ○ Self-injury: 0.4
				• Age-based Demotion on Newsfeed
				  ○ Suicide: 0.3
				  ○ Self-injury: 0.3
				• Age-based Deboost on Newsfeed
				  ○ Suicide: 0.05
				  ○ Self-injury: 0.05
				• Age gating on Newsfeed
				  ○ Suicide: 0.1
				  ○ Self-injury: 0.1
				• Non-rec filtering
				  ○ Suicide: 0.105
				  ○ Self-injury: 0.25
				• Filtering demotion on Stories
				  ○ Self-injury: 0.25

Plain-Language Classifier Name	Alpha-numeric code	Launch date of current iteration	Threshold value for autodeletion	Threshold value(s) for soft action(s)

				• Age gating on Stories
				  ○ Self-injury: 0.25
				• Age gating on Stories (reduce more)
				  ○ Self-injury: 0.1
				
				Instagram: various enforcements > P40
				• Age gating
				  ○ Suicide: 0.89
				  ○ Self-injury: 0.71
				• Age gating High COE
				  ○ Suicide: 0.6
				  ○ Self-injury: 0.24
				• Non-rec filtering
				  ○ Suicide: 0.6
				  ○ Self-injury: 0.24
				• Hashtag filtering
				  ○ Suicide: 0.80
				  ○ Self-injury: 0.80
				• Media SERP
				  ○ Suicide: 0.85
				  ○ Self-injury: 0.85
				• Age-filtering
				  ○ Suicide: 0.6
				  ○ Self-injury: 0.24
				• Demotion Feed teen non-EU
				  ○ Suicide: 0.7
				  ○ Self-injury: 0.7
				• Demotion Feed teen EU
				  ○ Suicide: 0.85
				  ○ Self-injury: 0.85
				• Demotion Stories teen
				  ○ Suicide: 0.7
				  ○ Self-injury: 0.7

SSI Borderline Classifier	f551409441	04/15/2024	N/A	Facebook: various enforcements above >P10 • Age gating on Newsfeed: 0.1 • P-non MSI: 0.3

Plain-Language Classifier Name	Alpha-numeric code	Launch date of current iteration	Threshold value for autodeletion	Threshold value(s) for soft action(s)
				• Non-rec filtering: 0.3 • Age gating on Stories: 0.75 • Search demotion: 0.5 Instagram: various enforcements > P40 • Age gating: 0.91 • Age gating High CoE: 0.54 • Non-rec filtering: 0.34 • Media SERP: 0.85 • Hashtag filtering: 0.80 • Age filtering: 0.25

B. ED Classifiers

Meta launched a classifier to action ED content specifically at least as early as March 2021. Prior to the launch of an ED-specific classifier, ED content fell under the purview of SSI policies and enforcement technology. The latest iteration of ED classifiers were launched at various times in 2024.[12] ED classifiers are intended to detect ED content on Facebook and Instagram that may violate Meta's Policies.[13] ED classifiers consider the following content as policy-violating and subject to removal: "[c]ontent that focuses on depiction of ribs, collar bones, thigh gaps, hips, concave stomach, or protruding spine or scapula when shared together with terms associated with eating disorders."[14] Not all policy violating ED content is subject to removal, such as content that "depicts ribs, collar bones, thigh gaps, hips, concave stomach, or protruding spine or scapula in a recovery context."[15] Such content is

[12] *See infra* Table 2.

[13] *Suicide & Self Injury*, Meta: Transparency Center, https://transparency.meta.com/policies/community-standards/suicide-self-injury/ (last visited Feb. 6, 2025).

[14] *Id.*

[15] *Id.*

placed behind a sensitivity screen. Generally, content promoting or encouraging EDs are removed, but users are allowed to post content that touches on their own experiences around self-image and body acceptance.[16]

Because Meta's platforms see less ED content than SSI content, there are fewer datapoints for the AI system to learn from in the ED sphere than for SSI. To combat this issue, Meta generates synthetic training data the classifiers can learn from.

ED Classifiers do not currently perform automatic deletions and therefore there is no threshold value for autodeletion. All content with a value above P70 is demoted and enqueued for human review and potential manual deletion. Any content at or below a value of P70 but higher than the respective minimum values set forth below in Table 2 is demoted. By extension, Meta periodically re-assesses the threshold value at which content is enqueued for human review, and the threshold value is subject to change at any given time.

Table 2. ED Classifiers & ED Borderline Classifiers.

Plain-language classifier name	**Alpha-numeric code**	**Launch date of current iteration**	**Threshold value for autodeletion**	**Threshold value(s) for soft action(s)**
ED Classifier	f638461582	09/01/2024	N/A	Facebook • EU Demotion on Newsfeed: 0.8 • EU Deboost on Newsfeed: 0.15 • Non-EU Demotion on Newsfeed: 0.5 • Non-EU Deboost on Newsfeed: 0.15 • Age based Demotion on Newsfeed: 0.3 • Age based Deboost on Newsfeed: 0.05

[16] *Id.*

Plain-language classifier name	Alpha-numeric code	Launch date of current iteration	Threshold value for autodeletion	Threshold value(s) for soft action(s)
				• Age gating on Newsfeed: 0.1 • Non-rec filtering: 0.061 Instagram: various enforcements >P40 • Non-rec filtering: 0.81 • Media SERP: 0.85 • Hashtag filtering: 0.80 • Demotion Feed teen non-EU: 0.7 • Demotion Feed teen EU: 0.85 • Demotion Stories teen: 0.7
ED Borderline Classifier	f551542423	04/15/2024	N/A	Facebook • Age gating on Newsfeed: 0.1 • P-non MSI: 0.3 • Non-rec filtering: 0.33 • Search demotion: 0.5 Instagram: various enforcements >P40 • Age gating: 0.907 • Non-rec filtering: 0.7 (P60) • Media SERP: 0.85 • Hashtag filtering: 0.80 • Age filtering: 0.26 (P40)

C. Bullying and Harassment Classifiers

Meta first launched a classifier that actioned "bullying and harassment" ("B&H") content at least as early as May 2018. The latest iteration of B&H classifiers were launched at various points in 2022, 2023, and 2024.[17] B&H classifiers are intended to detect bullying and harassment content that may violate

[17] *See infra* Table 3.

Meta's Policies.[18] Bullying and harassment classifiers consider severe attacks on a public figure as policy-violating, but not necessarily content that merely degrades or shames a public figure.[19] Under Meta's Policies, public figures are state and national level government officials, political candidates for those offices, people with over one million fans or followers on social media, and people who receive substantial news coverage. Bullying and harassment classifiers consider content meant to degrade or shame any private individual as policy-violating.[20] Meta recognizes that bullying and harassment can have more of an emotional impact on minors, which is why Meta's policies provide heightened protection for anyone under the age 18, regardless of user status.

Some content is removed for all users: repeatedly degrading or shaming contact, attacks based on the target's experience of sexual assault or domestic abuse, calls for SSI to a specific individual or group of individuals, derogatory terms, claims that a violent tragedy did not occur, claims that individuals are lying about being a victim of a violent tragedy, threats to release private information, statements to engage in a sexual activity, severe sexualized commentary, derogatory sexualized photoshop, calls for bullying or harassment of people, and degrading people who are depicted vomiting, defecating, urinating, or menstruating.

In addition to Meta's tools to take action on a user's behalf, Meta also provides tools that help users control their own experiences on platform, such as the ability to block certain key words from being posted in the comments to their content.

The confidence level ranges for soft actions vary, both over time and between classifier use cases. These are set by surface teams. The levels also vary based on user demographics, resulting in stricter soft actions for youth users. While the threshold for soft actions amongst general adult users typically rests

---

[18] *Bullying & Harassment*, Meta: Transparency Center, https://transparency.meta.com/policies/community-standards/bullying-harassment/ (last visited Feb. 6, 2025).

[19] *Id.*

[20] *Id.*

between 70 and 80, for youth it is considerably lower. This results in less accuracy (i.e., more false positives) but in a more protected experience for youth users.

Table 3. B&H Classifiers & B&H Borderline Classifiers.

Plain-language classifier name	Alpha-numeric code	Launch date of current iteration	Threshold value for autodeletion	Threshold value(s) for soft action(s)
FB Comment	8165225660248547	10/20/2024	Bullying and Harassment = 0.95 Borderline Hostile Speech = n/a	Bullying and Harassment: • Filtering on Comments: 0.432 • Age filtering comments: 0.275 Borderline Hostile Speech: • Filtering on Comments: 0.6 • Age filtering on Comments: 0.2
FB Post	8655505921181700	10/23/2024	Bullying and Harassment = 0.95 Borderline Hostile Speech = n/a	Bullying and Harassment: • Age-based Non-rec Filter on Newsfeed: P30 • Age-based Deboost on Newsfeed: P10 • Age-based Demotion on Newsfeed: P30 • Deboost on Newsfeed: P50 • Demotion on Newsfeed: P50 • EU Demotions on Newsfeed: P80 • Demotion on IFR: 0.50 • Non-rec Filter on IFR: 0.60

Plain-language classifier name	Alpha-numeric code	Launch date of current iteration	Threshold value for autodeletion	Threshold value(s) for soft action(s)
				• Age gating on Newsfeed: P10 • ARC Demotion on Newsfeed: P50 Borderline: • Age-based Non-rec Filter on Newsfeed: P10 • Demotion on IFR: 0.50 • Non-rec Filter on IFR: P60 • ARC Demotion on Newsfeed: P50 • ARC Deboost on Newsfeed: P50
IG Comment	5295222083900364	06/06/2022	Bullying and Harassment = 0.95 Borderline Hostile Speech = n/a	Bullying and Harassment: • Comment filter (auto-hide): P80 • Comment cover: P70 • Advanced comment filtering: P60 • Comment downranking (demotion): P50 • Preview comment filter: P50 Borderline Hostile Speech: • Comment filter (auto-hide): P95 • Comment cover: P80 • Advanced comment filtering: P70

Plain-language classifier name	Alpha-numeric code	Launch date of current iteration	Threshold value for autodeletion	Threshold value(s) for soft action(s)
				• Comment downranking (demotion): P70 • Preview comment filter: P70
IG Post	24098062513172542	11/05/2023	Bullying and Harassment = 0.95 Borderline Hostile Speech = n/a	Bullying and Harassment: • Youth non-rec filter: P50 • Youth demotions on connected stories: P60 • Youth demotions on connected feed: P60 • EU youth demotions on connected feed: P80 Borderline Hostile Speech: • Youth non-rec filter: P40
Reels	7408044142611719	04/27/2024	Bullying and Harassment = 0.95 Borderline Hostile Speech = n/a	Facebook Bullying and Harassment: • Age-based Non-rec Filter on Reels: P15 • Non-rec Filter on Reels: P25 • Non-rec Filter on Watch: P25 Borderline Hostile Speech:

Plain-language classifier name	Alpha-numeric code	Launch date of current iteration	Threshold value for autodeletion	Threshold value(s) for soft action(s)
				• Age-based Non-rec Filter on Reels: P15 • Non-rec Filter on Reels: P25 • Non-rec Filter on Watch: P25 Instagram Bullying and Harassment: • Youth demotions on connected feed: P60 • EU youth demotions on connected feed: P80 Borderline Hostile Speech: • Youth non-rec filter: P40

D. Graphic and Violent Content Classifiers

Meta first launched a classifier that actioned "graphic and violent content" ("GV") content at least as early as December 2016. The latest iteration of GV classifiers were launched at various times in 2021 and 2024. GV classifiers are intended to detect graphic and violent content on Facebook and Instagram that may violate Meta's Policies.[21] GV classifiers consider, for example, photos and videos of wounded

[21] *Violent & Graphic Content*, Meta: Transparency Center, https://transparency.meta.com/policies/community-standards/violent-graphic-content/ (last visited Feb. 6, 2025).

or dead people if they show dismemberment, visible innards, charred or burning people, victims of cannibalism, and/or throat slitting, as policy-violating.[22]

To protect users, and consistent with Meta's content policies, Meta employs a GV classifier to autodelete the most graphic and policy-violating content.[23] Other GV classifiers add a "disturbing" label to less graphic content so that users are aware that the content may be sensitive or disturbing.[24] When content is marked as disturbing by a GV classifier, it is routed for human review for potential deletion. Meta allows some GV content to remain on the platform out of recognition that users may share content to shed light on or condemn acts.[25] Nevertheless, even for less graphic content, various GV classifiers age gate content from youth users.[26]

Table 4. GV Classifiers & GV Borderline Classifiers.

Plain-language classifier name	Alpha-numeric code	Launch date of current iteration	Threshold value for autodeletion	Threshold value(s) for soft action(s)

Graphic Violence MAD photo classifier on FB	f647163274	10/01/2024	N/A	• Mark as disturbing ("MAD"): [redacted]
				• Various enforcements > P50
				  ○ Reels: [redacted]
				  ○ CFR Demotion: [redacted]
				  ○ CFR Deboost: [redacted]
				  ○ Stories demotion: [redacted]
				  ○ Search filtering : [redacted]

Graphic Violence MAD video	f645517662	9/25/2024	N/A	• MAD: [redacted] • Various enforcements > P50 ○ Reels [redacted]

[22] *Id.*

[23] *See infra* Table 4 (GV Video Deletion classifier on IG); *see also Violent and Graphic Content*, Meta Transparency Center, https://transparency.meta.com/policies/community-standards/violent-graphic-content/ (last visited Feb. 2, 2025).

[24] *Violent and Graphic Content*, Meta Transparency Center, https://transparency.meta.com/policies/community-standards/violent-graphic-content/ (last visited Feb. 6, 2025).

[25] *Id.*

[26] *Id.*

Plain-language classifier name	Alpha-numeric code	Launch date of current iteration	Threshold value for autodeletion	Threshold value(s) for soft action(s)
classifier on FB				○ CFR Demotion: [redacted] ○ CFR Deboost: [redacted] ○ Stories demotion: [redacted] ○ Search filtering : [redacted]
Graphic Violence MAD photo classifier on IG	f647151014	09/30/2024	N/A	• MAD: [redacted] • EU age gating on Feed: [redacted] • Non-EU age gating on Feed: [redacted]
Graphic Violence MAD video classifier on IG	f639562790	09/11/2024	N/A	• MAD: [redacted] • EU age gating on Feed: [redacted] • Non-EU age gating on Feed: [redacted]
GV Video Deletion classifier on IG	3614043688651385	02/21/2021	[redacted]	N/A
Borderline GV classifier			N/A	Facebook Youth: Various enforcements > P10 • Feed age gating: [redacted] (P10) • Feed demotion: [redacted] (P30) • Reels age gating: [redacted] (P20) • Reels age filtering: [redacted] (P25) • Watch/IFR filtering: [redacted] (P25) Instagram Youth: various enforcements > P40 • Filtering: [redacted] (P50) • Demotion: [redacted] (P40) • Age gating: [redacted] (P80)

E. Adult Nudity & Sexual Activity Classifiers

Meta first launched a classifier to action "adult nudity & sexual activity" ("ANSA") content at least as early as January 2016. The current iteration of ANSA classifiers were launched at various times in 2024.[27] ANSA classifiers are intended to detect images and videos of adult nudity and sexual activity on Facebook and Instagram that may violate Meta's Policies.[28] Adult nudity classifiers consider the following adult nudity content as policy-violating: "visible genitalia"; "visible anuses and/or fully nude close-ups of buttocks"; "uncovered female nipples"; "explicit sexual activity and stimulation"; "implicit sexual activity and stimulation"; "erections"; "presence of by-products of sexual activity"; "sex toys placed upon or inserted into mouth"; "stimulation of visible human nipples"; "squeezing female breasts"; "imagery depicting fetish that involves [ ] acts that are likely to lead to the death of a person or animal," "dismemberment," "cannibalism," "feces, urine, spit, snot, menstruation or vomit," "bestiality," and "incest"; "digital imagery of adult sexual activity"; and "extended audio of sexual activity."[29]

Table 5. ANSA Classifiers & ANSA Borderline Classifiers.

Plain-language classifier name	Alpha-numeric code	Launch date of current iteration	Threshold value for autodeletion	Threshold value(s) for soft action(s)
Violating ANSA photo classifier on FB and IG	f644294876	09/18/2024	• [redacted] for FB • [redacted] for IG	Instagram • Age gating: [redacted] Facebook Youth filter on stories: [redacted] Youth filter on reels: [redacted] Youth filter on watch: [redacted] Youth filter on feed: [redacted]

[27] *See infra* Table 5.

[28] *Adult Nudity & Sexual Activity*, Meta: Transparency Center, https://transparency.meta.com/policies/community-standards/adult-nudity-sexual-activity/ (last visited Feb. 6, 2025).

[29] *Id.*

Plain-language classifier name	Alpha-numeric code	Launch date of current iteration	Threshold value for autodeletion	Threshold value(s) for soft action(s)
Violating ANSA video classifier on FB and IG	f643324421	09/16/2024	• [redacted] for FB • [redacted] for IG	Instagram • Age gating: [redacted] Facebook • Youth filter on stories: [redacted] • Youth filter on reels: [redacted] • Youth filter on watch: [redacted] • Youth filter on feed: [redacted]
Violating ANSA linkshare classifier on FB and IG	f635467640	08/26/2024	[redacted] for FB	Instagram • Age gating: [redacted] Facebook • Youth filter on stories: [redacted] • Youth filter on reels: [redacted] • Youth filter on watch: [redacted] • Youth filter on feed: [redacted]
Borderline ANSA Photo Classifier on FB	f645114003	9/21/2024	N/A	• Youth filter on Reels: [redacted] • Youth filter on Watch: [redacted] • Youth filter on Feed: [redacted] • Youth filter on IFR: [redacted] • Youth filter on Story: [redacted]
Borderline ANSA Video Classifier on FB	f645120189	9/21/2024	N/A	• Youth filter on Reels: [redacted] • Youth filter on Watch: [redacted] • Youth filter on Feed: [redacted]

Plain-language classifier name	Alpha-numeric code	Launch date of current iteration	Threshold value for autodeletion	Threshold value(s) for soft action(s)
				• Youth filter on IFR: [redacted] • Youth filter on Story: [redacted]
IG Non-rec ANSA Photo Classifier on IG	f639524943	09/04/2024	N/A	• Age gating: [redacted] • Age filtering: [redacted] • Age demotion: [redacted]
IG Non-rec ANSA video classifier on IG	f641128648	09/10/2024	N/A	• Age gating: [redacted] • Age filtering: [redacted] • Age demotion: [redacted]

F. Child Safety Classifiers

Meta first launched a classifier to action "child safety" ("CS") content at least as early as 2018. The current iterations of CS classifiers were launched at various times in 2020, 2023, and 2024.[30] CS classifiers are intended to proactively identify and action content and behaviors that may violate Meta's policies on Child Sexual Exploitation, Abuse, and Nudity.[31] Child Safety classifiers consider the following content, inter alia, as violating: content, activity, or interactions that threaten, depict, praise, support, provide instructions for, make statements of intent, admit participation in, or share links of the sexual exploitation of children; content that solicits sexual content or activity depicting or involving children, nude imagery of real or non-real children, and/or sexualized imagery of real or non-real children; content that solicits sexual encounters with children; content that constitutes or facilitates inappropriate

[30] *See infra* Table. 6.

[31] *Child Sexual Exploitation, Abuse, and Nudity*, Meta: Transparency Center, https://transparency.meta.com/policies/community-standards/child-sexual-exploitation-abuse-nudity/ (last visited Feb. 7, 2025).

interactions with children; content that attempts to exploit real children by coercing money, favors or intimate imagery with threats to expose intimate imagery or information or by sharing, threatening, or stating an intent to share private sexual conversations or intimate imagery; content that sexualizes real or non-real children; Groups, Pages, and profiles dedicated to sexualizing real or non-real children; content that depicts real or non-real child nudity; videos or photos that depict real or non-real non-sexual child abuse regardless of sharing intent; or content that praises, supports, promotes, advocates for, provides instructions for or encourages participation in non-sexual child abuse.[32]

The names of the below classifiers generally align with their intended purpose. For example, the CSAM classifier is intended to identify CSAM, while the CSAM solicitation classifier is intended to identify the solicitation of CSAM. The Google Content Safety API classifier is intended to detect CSAM and is supported by Google's Application Programming Interface. The Child Sexualization ("CSx") classifier is intended to identify child sexualization content. The MetaGen for Child Safety classifier is intended to identify child safety violating content, generally. All of the classifiers identified in Table 6 are run simultaneously, with the exception of the Google Content Safety API classifier which is used to prioritize human review queues. There are no borderline content detection classifiers in this harm area. The threshold values for enqueueing CS content for human review and manual removal change frequently to ensure we can enqueue as much potentially violating CS content for human review as possible. The pIV ranking system works to ensure that potential CS content most likely to violate policy moves to the top of the review queue.[33]

[32] *Id.*

[33] *See* discussion *supra* Section I.B.

Table 6. Child Safety Classifiers.

Plain-language classifier name	Alpha-numeric code	Launch date of current iteration	Threshold value for autodeletion	Threshold value(s) for soft action(s)
CSAM classifier	FB: f588692710; IG: f565741344	08/29/2024	N/A	Facebook • Age filtering: [redacted] (P20) • Filtering for GenPop: [redacted] (P25) Instagram • Filtering for Teens: [redacted] (P25)
CSAM Solicitation classifier	FB: f636246456; IG: f566561739	09/16/2024	N/A	Facebook • Age filtering: [redacted] (P20) • Filtering for GenPop: [redacted] (P25) Instagram • Filtering for Teens: [redacted] (P25)
CSAM-S comment classifier	f636721225	08/29/2024	N/A	Facebook • Filtering for GenPop: [redacted] (P25) Instagram • IG: no soft actions
Child Sexualization (CSx) classifier	f590979030	08/28/2024	N/A	Facebook • Age filtering: [redacted] (P20) • Filtering for GenPop: [redacted] (P25) Instagram • Filtering for Teens: [redacted] (P25)

Plain-language classifier name	Alpha-numeric code	Launch date of current iteration	Threshold value for autodeletion	Threshold value(s) for soft action(s)
CSE-I Classifier	f573679609	06/18/2024	N/A	Facebook • Age filtering: [redacted] (P20) • Filtering for GenPop: [redacted] (P25) Instagram • IG: no soft actions
Child subject to objectionable content classifier	f589982602	08/20/2024	N/A	Facebook • Age filtering: [redacted] (P20) • Filtering for GenPop: [redacted] (P30) Instagram • Filtering for Teens: [redacted] (P50)
Google Content Safety API classifier	N/A	06/02/20	N/A	No soft actions

III. Borderline Classifiers

In addition to its policies regarding clearly violating content, Meta also has had borderline content policies since 2020. A "borderline classifier" is an artificial intelligence tool that reviews and detects content on Facebook and Instagram that is not strictly policy-implicating, but nonetheless considered potentially problematic or low quality. For example, a borderline graphic and violent content classifier would consider content of puss-filled or oozing wounds, people undergoing surgical operations, fight videos, animal abuse, and fictional violence, not strictly policy-violating, but as "borderline content."

Where a piece of content is determined to be "borderline content," then that content is not removed, but will be downranked.

Some borderline classifiers can detect a single type of harmful content, while others can detect multiple types of harmful content. For example, a single borderline classifier can detect both borderline SSI and borderline ED content.

There is a substantially larger amount of material that falls into the borderline category relative to content that is clearly policy-violating. As a consequence, Meta takes exponentially higher amounts of soft actions than it does hard actions like outright deletion.

There are age-related differences in the application of Meta's classifiers. Instagram has an Age Appropriate Content vision, where youth users should be less likely to see content that violates Meta's Community Standards or to be recommended content that has been deemed non-recommendable—in other words, young users see less borderline content. This essentially results in stricter enforcement of policies for teen users.

When Meta's content moderation technology detects content that is sensitive but not technically a violation of policy (*i.e.*, healed scars or content trivializing SSI), Meta makes this content harder for users to find. For example, Meta has an SSI Borderline Policy that covers things that are close to prohibited, or SSI-adjacent like dark and depressing content, which has been in place since 2020. Content that is "aimed to be informative, diagnostic, to discuss recovery or to provide help and encouragement to those experiencing [SSI]" is excluded from the borderline policy, and thus allowed on Meta's platforms.

Meta also has an ED Borderline Policy intended to limit user exposure to content posted by non-connected users that could cause harm when viewed either by a vulnerable user or as a part of an aggregate viewing experience. Meta also has a GV Borderline Policy, covering things like pus or oozing wounds, people undergoing surgical operations, fight videos, animal abuse, and fictional violence. Further, Meta has an ANSA Borderline Policy covering things like implied sexual intercourse and stimulation.

Meta has identified above the relevant borderline classifiers and their respective alphanumeric codes and probability thresholds.

IV. Meta Has Changed Its Prioritization of Certain Harm Areas During the Relevant Time Period.

Meta has invested in protecting teen mental health since 2009. Since the inception of its investment in this area, addressing SSI and Child Safety content have remained top priorities among leadership at Meta responsible for content moderation. Prior to 2019, Meta's SSI content policies encompassed ED and Credible Intent of Suicide ("CIS") content, in addition to other SSI content. In Q3 2019, Meta updated its policies, classifiers, and enforcement practices to begin addressing ED content separately from other SSI content. In Q1 2023, Meta updated its policies, classifiers, and enforcement practices to begin addressing CIS content separately from other SSI content.

V. Core Members of the Content Moderation Team

There are multiple teams within Meta whose responsibilities relate to the content moderation topics described above, and so a complete list of individuals with knowledge of these topics would be extremely large. Without purporting to identify all such individuals, the following are core members of the various teams involved in content moderation: Aju Bardardeen, Matthew [redacted], Robbie [redacted], Zaur [redacted], Fatima [redacted], Arcadiy [redacted], Liz [redacted], George [redacted], Christopher [redacted], Ryan [redacted], Zain [redacted], Michael [redacted], Richard [redacted], and Jon [redacted].

Dated: February 21, 2025

Respectfully submitted,

COVINGTON & BURLING LLP

/s/ Ashley M. Simonsen
Ashley M. Simonsen (State Bar No. 275203)
asimonsen@cov.com
COVINGTON & BURLING LLP
1999 Avenue of the Stars
Los Angeles, CA 90067
Telephone: (424) 332-4800
Facsimile: + 1 (424) 332-4749

Emily Johnson Henn (State Bar No. 269482)
ehenn@cov.com
COVINGTON & BURLING LLP
3000 El Camino Real
5 Palo Alto Square, 10th Floor
Palo Alto, CA 94306
Telephone: + 1 (650) 632-4700
Facsimile: +1 (650) 632-4800

Phyllis A. Jones, *pro hac vice*
pajones@cov.com
Paul W. Schmidt, *pro hac vice*
pschmidt@cov.com
Michael X. Imbroscio, *pro hac vice*
mimbroscio@cov.com
COVINGTON & BURLING LLP
One CityCenter
850 Tenth Street, NW
Washington, DC 20001-4956
Telephone: + 1 (202) 662-6000
Facsimile: + 1 (202) 662-6291

Attorneys for Defendants Meta Platforms, Inc. f/k/a Facebook, Inc.; Facebook Holdings, LLC; Facebook Operations, LLC; Facebook Payments, Inc.; Facebook Technologies, LLC; Instagram, LLC; Siculus, Inc.; and Mark Elliot Zuckerberg

CERTIFICATE OF SERVICE

The undersigned hereby certifies that a copy of the foregoing was served via electronic mail on February 21, 2025 on the following:

PSCServiceMDL3047@motleyrice.com

~SnapMDL3047JCCP5255@mto.com

john.beisner@skadden.com

YT-MDL3047JCCP5255@list.wsgr.com

YT-ML-MDL3047JCCP5255@morganlewis.com

W&C-YT-PL-Cases@wc.com

TikTokMDL3047JCCP5255@faegredrinker.com

TikTokMDL3047JCCP5255@kslaw.com

SM.MDLAGLeads@coag.gov

mdl3047coleadfirms@listserv.motleyrice.com

Defendants-MDL3047JCCP5255@skaddenlists.com

I declare under penalty of perjury that the foregoing is true and correct. Executed on February 21, 2025.

DATED: February 21, 2025 By: */s/ Ashley M. Simonsen*
Ashley M. Simonsen

VERIFICATION

I, Arcadiy Kantor, am a Director of Product Integrity at Meta. I am authorized to make this Verification for and on behalf of Meta. I have read META'S THIRD SUPPLEMENTAL RESPONSES & OBJECTIONS TO PLAINTIFFS' FIRST INTERROGATORY, and know the contents thereof. I affirm that I have personal knowledge that the facts set forth therein are true and correct, or as to matters that are not within my personal knowledge, that I have made a reasonable inquiry and I am informed and believe that those facts are true and correct.

I declare under penalty of perjury under the laws of the United States that the foregoing is true and correct this 12th day of February, 2025, in Atlanta, Georgia.

Signed by:
/s/ Arcadiy Kantor
BB72D31B359C4A0...
Arcadiy Kantor