AMENDED Exhibit 166

PLAINTIFFS' OMNIBUS OPPOSITION TO DEFENDANTS' MOTIONS FOR SUMMARY JUDGMENT

HIGHLY CONFIDENTIAL (COMPETITOR)

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.

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.

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: 0.1 • Youth filter on Story: 0.5
IG Non-rec ANSA Photo Classifier on IG	f639524943	09/04/2024	N/A	• Age gating: 0.623 • Age filtering: 0.159 • Age demotion: 0.104
IG Non-rec ANSA video classifier on IG	f641128648	09/10/2024	N/A	• Age gating: 0.623 • Age filtering: 0.237 • Age demotion: 0.158

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: 0.875 (P20) • Filtering for GenPop: 0.891 (P25) Instagram • Filtering for Teens: 0.943 (P25)
CSAM Solicitation classifier	FB: f636246456; IG: f566561739	09/16/2024	N/A	Facebook • Age filtering: 0.982 (P20) • Filtering for GenPop: 0.984 (P25) Instagram • Filtering for Teens: 0.965 (P25)
CSAM-S comment classifier	f636721225	08/29/2024	N/A	Facebook • Filtering for GenPop: 0.9975 (P25) Instagram • IG: no soft actions
Child Sexualization (CSx) classifier	f590979030	08/28/2024	N/A	Facebook • Age filtering: 0.942 (P20) • Filtering for GenPop: 0.945 (P25) Instagram • Filtering for Teens: 0.979 (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: 0.83 (P20) • Filtering for GenPop: 0.847 (P25) Instagram • IG: no soft actions
Child subject to objectionable content classifier	f589982602	08/20/2024	N/A	Facebook • Age filtering: 0.949 (P20) • Filtering for GenPop: 0.951 (P30) Instagram • Filtering for Teens: 0.957 (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."

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