



18 January 2024

Attention: Prabhat Agarwal, Head of Unit

Your ref: Ares(2023)8217917

Response to the Request for information pursuant to Article 67(1), Article 72(1) and Article 34(3) of Regulation (EU) 2022/2065 regarding Meta's compliance with Articles 14, 34 and 35 of Regulation (EU) 2022/2065

Dear Mr Agarwal,

Meta Platforms Ireland Limited (“MPIL”) notes that, independently of MPIL’s obligations under Regulation (EU) 2022/2065 (the Digital Services Act or (“DSA”), Meta already had, and continues to maintain and evolve, effective policies and procedures to assess and mitigate any risks in relation to the protection of minors, particularly related to CSAM, and serious negative consequences to a person’s physical and mental well-being, among other potential risks. MPIL understands that this Request for information (“RFI”) is limited to MPIL’s risk assessment and mitigation measures taken after the date of application of Articles 34 and 35 DSA in relation to the designated Facebook and Instagram VLOP services offered to recipients established or located in the EU, and this is reflected in its responses. Nothing in this response or in the measures MPIL has undertaken to comply with the DSA, including its risk assessment and mitigation process, should suggest that MPIL believes that using Facebook or Instagram causes, or is likely to cause, any particular physical or mental well-being harms.

MPIL’s response to this RFI should be read in light of MPIL’s response to the Request for information pursuant to Article 67(1), Article 72(1) and Article 34(3) of Regulation (EU) 2022/2065 regarding Meta’s compliance with Articles 14, 34 and 35 of Regulation (EU) 2022/2065 of 15 December 2023 (the “**15 December Response**”), which overlaps with the subject-matter of the present RFI, and in some cases duplicates the questions. This response will cross-refer as appropriate to the 15 December Response and its annexes.

Pursuant to discussions with the Commission on 19 December 2023, MPIL understands that all references in this RFI to “*underaged users*” are intended to refer to minors (*i.e.*, under 18s) who are at least 13 years old.¹ MPIL refers to such users as ‘teen’ users. For other defined terms used herein, please see the Glossary located at the end of this response.

Further, MPIL notes that it had originally asked for an extension of the deadline to respond to 31 January 2024. MPIL has used its best endeavours, in the limited period of time given to it by a shorter extension granted, to provide a complete response and collect relevant documents where requested. MPIL would be pleased to supplement this response should the Commission require additional information.

Finally, MPIL notes that, over and above its confidential and commercially sensitive nature, this response contains extremely sensitive information on how MPIL detects and enforces against bad actors. The broader release of this information could prejudice the effective application of MPIL’s policies and protocols, negatively impact MPIL’s efforts to safeguard users and the public at large, and ultimately place persons at risk. MPIL trusts that the Commission will take all necessary precautions to ensure the information contained herein is not shared beyond those persons who need to have access.

Overview of MPIL’s risk assessment for Instagram and Facebook in relation to the protection of minors and the person’s physical and mental well-being, in particular as regards minors

¹ 14 years old in Spain.

1. Before addressing the specific questions posed in the RFI, MPIL refers to the overview contained in the 15 December Response of MPIL’s risk assessment for Instagram and Facebook in relation to, *inter alia*, the protection of minors.
2. That overview set out how the 2023 DSA Systemic Risk Assessment for Instagram and Facebook (described in, respectively, the “**Instagram SRA Results Report**” and the “**Facebook SRA Results Report**”²) approached the assessment of systemic risks identified in Article 34(1)(b) DSA (which lists any actual or foreseeable negative effects for the exercise of the rights of the child as enshrined in Article 24 of the European Charter of Fundamental Rights) and Article 34(1)(d) DSA (which lists any actual or foreseeable negative effects in relation to the protection of minors as well as serious negative consequences to the person’s physical and mental well-being).
3. As explained in more detail in the 15 December Response, MPIL developed a detailed risk assessment framework and approach in line with industry standards, including ISO-31000: Risk Management,³ best practices, and other previously established risk management practices at Meta. The various stages of the risk assessment process are described in detail in paragraphs 6-13 of the 15 December Response.
4. As part of this process, MPIL leveraged Meta’s industry-leading content moderation processes that it has developed over many years and which continue to evolve. Leveraging this long-established integrity ecosystem, and in parallel to its DSA compliance obligations, Meta continues to build foundational integrity risk and compliance programme components that enable it to scale and build a risk and compliance programme fit for managing systemic risks in the future. This includes building a comprehensive Integrity Risk Register that catalogues the integrity risks that could potentially arise on Facebook and Instagram. To create the register, Meta’s Integrity Governance Risk & Compliance team (“**IGRC Team**”) surveyed and consulted a variety of teams within Meta and analysed the ecosystem which included existing external engagement programmes to understand the risk landscape (e.g., via policy development processes) and identify integrity concerns.
5. As regards the dissemination of child sexual abuse materials (“**CSAM**”) on Instagram and Facebook, the SRAs addressed this under the Problem Area of Child Sexual Exploitation, Abuse & Nudity (“**CSEAN**”). This Problem Area refers to the risk of Meta’s systems being used to promote or disseminate content or activity of non-sexual child abuse, child nudity, child endangerment, Child Sexual Exploitation (“**CSE**”), CSAM, potentially including self-generated CSAM (“**SG-CSAM**”) and solicitation of CSAM, sexualization of minors, and exploitative intimate imagery and sextortion of minors. The risk assessment is documented in the Workbook for CSEAN.⁴
6. Specifically, MPIL identified seven L4 risks under this Problem Area:
 - a. **Child Endangerment:** Risk of Meta’s platforms being used to facilitate activities that put minors in dangerous, unhealthy, or inappropriate situations, potentially contributing or leading to detrimental external party impact.

² See Annexes 1-2 of the 15 December Response.

³ Relevant information is available at <https://www.iso.org/iso-31000-risk-management.html>.

⁴ See Annex 10 of the 15 December Response.

- b. **Child Nudity:** Risk of Meta’s platforms being used to share content that depicts real or non-real child nudity with a human likeness, such as close-ups of children’s genitalia, nude toddlers and minors, and non-real imagery such as in art, AI-generated content, fictional characters, dolls, etc., unless the non-real imagery is for health or educational purposes or is a depiction of child nudity in real-world art, potentially contributing or leading to detrimental external party impact.
 - c. **Child Sexual Abuse Materials:** Risk of Meta’s platforms being used to promote or disseminate content or activity that threatens, depicts, praises, supports, provides instructions for, makes statements of intent, admits participation in or shares links of the sexual exploitation of children, potentially contributing or leading to detrimental external party impact.
 - d. **Inappropriate Interactions with Children:** Risk of Meta’s platforms being used to share content that constitutes or facilitates inappropriate interactions with children, including arranging or planning real-world sexual encounters with children, purposefully exposing children to sexually explicit language or sexual material, engaging in implicitly sexual conversations in private messages with children, or obtaining or requesting sexual material from children in private messages, potentially contributing or leading to detrimental external party impact.
 - e. **Non-Sexual Child Abuse:** Risk of Meta’s platforms being used to share imagery that depicts real or non-real non-sexual child abuse or content that praises, supports, promotes, advocates for, provides instructions for, or encourages participation in non-sexual child abuse unless the imagery is from real-world art, cartoons, movies or video games, potentially contributing or leading to detrimental external party impact.
 - f. **Sexualization of Children:** Risk of Meta’s platforms being used to share content (including photos, videos, real-world art, digital content, and verbal depictions) that sexualizes real or non-real children or Groups, Pages and profiles dedicated to sexualizing real or non-real children, potentially contributing or leading to detrimental external party impact.
 - g. **Solicitation of Child Sexual Abuse Materials:** Risk of Meta’s platforms being used to facilitate the solicitation of sexual content depicting or involving real or non-real children (e.g., images, videos, links, instruction manuals, etc.), nude imagery of children, sexualized imagery of children, and real-world sexual encounters with children, potentially contributing or leading to detrimental external party impact.
7. As explained in the 15 December Response, MPIL pulled relevant controls from its inventory of controls and mapped these to the identified L4 risks. These are either Problem Area controls (identified with a unique ID in the format “CIDPA-XXX”), or Ecosystem controls (identified with a unique ID in the format “CIDE-XXX”). For convenience, MPIL has prepared a list of descriptions of the controls that were determined to directly contribute to mitigation of the risk statements within the Problem Area of CSEAN⁵ (“**CSEAN Control Descriptions**”), provided in Annex 1. As controls are routinely reviewed and adjusted by MPIL, these descriptions are subject to further updating.

⁵ This is not a document underlying the SRAs for the purpose of Article 34(3) DSA, but was prepared for the purpose of responding to this RFI.

Private and Confidential – Contains Commercially Sensitive Information

8. MPIL mapped the Problem Area and Ecosystem Controls on to the risks as defined/identified by the DSA they are designed to mitigate, and then carried out the scoring exercise described in the 15 December Response. This exercise is recorded in the tab ‘Detailed Control Scoring’ in the Workbook for CSEAN.⁶
9. MPIL’s approach to the protection of minors more generally is detailed in the 15 December Response. First, as indicated in figure 3 of the Facebook and Instagram SRA Results Reports, the Systemic Risk Area of ‘Protection of Minors’ was addressed under seven Problem Areas: (i) Adult Sexual Exploitation & Adult Nudity, (ii) Bullying & Harassment, (iii) CSEAN, (iv) Fraud & Deception, (v) Human Exploitation, (vi) Privacy & Security and (vii) Suicide & Self-Injury. Second, the Instagram SRA and Facebook SRA took into account the specific impact on minors when conducting the Severity analysis of all identified risks. Where a particular risk was deemed to have a potentially elevated impact on minors (or indeed other vulnerable groups), it received the highest possible tier rating (Tier 5) in the “Impacted Demographic” factor of the Severity Rubric. In addition, in the same Severity Rubric, where a risk was deemed to have a potentially elevated impact on the rights of the child (or other fundamental rights), it also received the highest possible tier rating (Tier 5) in the “Nature of Impact” factor. All risks identified under the Problem Area of CSEAN received the maximum severity rating (Tier 5) for Impacted Demographic and Nature of Impact (as can be seen in the tab ‘Risk Register & Scoring’ in the Workbook for CSEAN.⁷
10. With respect to physical and mental well-being, this is relevant to, and was therefore assessed under, multiple Problem Areas, as detailed further in this response and the 15 December Response (in particular in response to question 7).

As regards the risk management framework (compliance with Articles 34 and 35 of Regulation (EU) 2022/2065)

A. In relation to the dissemination of CSAM and in particular SG-CSAM on Instagram

1. The 7 June 2023 Stanford Internet Observatory study suggests that large networks of accounts have been operated by minors, openly advertising SG-CSAM for sale on Instagram. Instagram was identified as the most prominent platforms for these networks. After the above mentioned report was published on 7 June 2023, Meta claims that it set up a task force and employed mitigation measures such as child safety related controls. Information on the functioning and resources of this child safety task force were shared with the Commission and discussed in a bilateral meeting between Meta and the Commission on 24 August 2023.

According to Meta’s “Response to findings by the Stanford Internet Observatory and Wall Street Journal” submitted to the Commission on 29 August 2023, the task force consists of 40 child safety experts from various Meta departments. The task force focused on understanding and mitigating the issues related to the studies, and improving Meta’s internal processes to identify, monitor, and react more quickly when dealing with child sexualisation content. However, in Meta’s risk assessments for Facebook and Instagram there is no information on (the estimated effectiveness of) the task force and the measures it has implemented.

⁶ See Annex 10 of the 15 December Response.

⁷ See Annex 10 of the 15 December Response.

a) **Has the child safety task force been put in place permanently or has it been or will it at some point be dissolved? If it is no longer active, please indicate the reasons for the discontinuation of the task force.**

11. Notwithstanding the diligent assessment of the risk of the dissemination of CSAM carried out for the 2023 SRA under the CSEAN Problem Area, the existence of sophisticated bad actors who constantly evolve their tactics to avoid detection requires continuous monitoring, adaptation and investment to ensure child safety.
12. As explained in Meta's "Response to findings by the Stanford Internet Observatory and Wall Street Journal" submitted to the Commission on 29 August 2023 ("**Response to EC of 29 August 2023**"), provided in Annex 2, as soon as MPIL was notified of the matters identified by the Stanford Internet Observatory ("**SIO**") and Wall Street Journal ("**WSJ**") in June 2023, MPIL took immediate action to investigate and remediate several issues raised.⁸ As part of this MPIL also immediately recognised that additional work was needed to understand why its systems were not acting as expected in some cases. The Child Safety Task Force ("**CSTF**") was therefore set up in June 2023 as a responsive measure to investigate and, to the extent necessary, remediate and set additional remediation goals and objectives in response to the specific issues raised by the SIO and WSJ. The CSTF sought to concentrate and expedite the ongoing work Meta has been doing in this space for years to address these issues and was overseen by Meta's long-established Child Safety Team, which continues to tackle these issues even after the formal structure of the CSTF was wound down at the end of August 2023.
13. As mentioned in the Response to the EC of 29 August 2023, as part of its initial scoping work, the CSTF put in place an initial workstream to look in detail at three key issues: (i) handling of user reports ("**Theme 1**"); (ii) enforcement against bad actors ("**Theme 2**"); and (iii) recommendation, searches and hashtags ("**Theme 3**"). The CSTF also mapped out a fourth area focusing on monitoring and mitigation (process improvements, future proofing and auditability) focused on long term sustainability to help ensure that the issues focussed on by the CSTF continue to be monitored and improved upon going forward ("**Theme 4**").
14. The formal structure of the CSTF was wound down at the end of August as it had fulfilled Meta's initial goal of investigating and implementing immediate improvements to address the issues raised by the SIO and the WSJ.
15. The winding down of the CSTF was determined by reference to a number of exit criteria established for each Theme, which were tracked through weekly updates. At the end of August 2023, audits were conducted to review the CSTF's work and evaluate if the exit criteria had been met. It was determined that the exit criteria had been met sufficiently across the four themes, resulting in the formal structure of the CSTF being wound down.
16. However, it is important to note that the work carried out by the CSTF continues in relation to all of the themes and goals being subsumed into the wider work of the Child Safety Team and also the structures MPIL has put in place to meet its obligations under DSA, only now with the benefit of the CSTF's additional efforts. Further details of the CSTF's final evaluation resulting in the improvement of Meta's internal processes are set out in response to question 1(c). It should also be noted that the CSTF mainly focused on addressing the issues identified by the SIO and WSJ, while Meta's work on preventing dissemination of CSAM and SG-CSAM on Instagram is broader in scope.

⁸ See for example, responses to questions 5 and 6.

17.

18.

19.

20.

9

Control ID CIDPA-001.

10

Instagram Community Guidelines available at: [[HYPERLINK "https://help.instagram.com/477434105621119" \h](https://help.instagram.com/477434105621119)]. Control ID CIDE-034.

11

Facebook Community Standards available at: [[HYPERLINK "https://transparency.fb.com/en-gb/policies/community-standards/" \h](https://transparency.fb.com/en-gb/policies/community-standards/)]. Control ID CIDE-033.

12

That said, from an enforcement standpoint, SG-CSAM and malicious sharing of CSAM are treated with distinct violation types and enforcement outcomes. See response to question 5(a) below.

13

See further, response to question 5(a) below.

14

See Annex 10 of the 15 December Response.

21. [REDACTED]
22. [REDACTED]
23. [REDACTED]
24. [REDACTED]

¹⁵ Control ID CIDE-055.

¹⁶ The NCMEC is a US based private, non-profit 501(c)(3) corporation whose mission is to help find missing children, reduce child sexual exploitation, and prevent child victimisation; see [HYPERLINK "<https://www.missingkids.org/home>" \h].

¹⁷ Control ID CIDPA-003.

¹⁸ Control ID CIDE-052

¹⁹ Further details with respect to these tools are provided in response to questions 8 and 9.

²⁰ Control ID CIDE-048.

²¹ Control ID CIDE-194.

²² Control ID CIDE-099.

²³ Control ID CIDPA-287.

²⁴ CSx is defined as content (including photos, videos, real-world art, digital content, and verbal depictions) that sexualizes real or non-real children or Groups, Pages and profiles dedicated to sexualizing real or non-real children, potentially contributing or leading to detrimental external party impact.

²⁵ MPIL recalls that it gave a demonstration of the SRT to the European Commission's DSA team on 18 July 2023.

25.



Classifier	Surfaces	Input	Output	Action	Training
CSAM content	Facebook, Instagram	Content	Binary score indicating likelihood of being CSAM	Enqueues content that passes score threshold for review	Ad hoc retraining using jobs reviewed in SRT and MMS bank hits
CSAM solicitation content	Facebook, Instagram	Content	Binary score indicating likelihood of being CSAM solicitation	Enqueues content that passes score threshold for review	Ad hoc retraining using jobs reviewed in SRT
CSAM Facebook groups	Facebook Groups	Group	Score indicating Facebook Group's purpose is CSAM or Minor Sexualisation sharing	Top score predictions are enqueued for review and actioning daily	Facebook Groups labelled as CSAM/Minor Sexualisation sharing
Child Sexualisation content	Facebook, Instagram	Content	Binary score indicating likelihood of being CSx	Enqueues content that passes score threshold for review	Ad hoc retraining using jobs reviewed in SRT and MMS bank hits
Google Content Safety API	Facebook, Instagram	Image	Score indicating CSAM or Minor Sexualisation sharing	Enqueues images that bypass score thresholds; used in SRT queue ranking to prioritise potential CSAM reviews	Unknown / confidential within Google

²⁶ Control IDs CIDE-058 and CIDE-056.

²⁷ For more detail on the Unified Child Safety Keyword List, see response to question 2 below.

Meta's enforcement through expert teams²⁸ and technology²⁹

26. In addition to the scaled review and enforcement, Meta also has specialist teams focused on understanding evolving criminal behaviours to review and eliminate abusive content, accounts and networks. Meta's investigation teams³⁰ receive information off-platform (from various sources including open source, intelligence vendors and partners) and then dedicate time investigating the activity on Meta's platforms to conduct takedowns of entire networks of bad actors (**Strategic Network Disruptions** or "SND"). For instance, Meta's [[HYPERLINK "https://docs.google.com/document/d/1vcpuCTV_S6ldKdXIQeCvzVLNJSxUxn1W7FZpxCvmcQ/edit"](https://docs.google.com/document/d/1vcpuCTV_S6ldKdXIQeCvzVLNJSxUxn1W7FZpxCvmcQ/edit) \l "heading=h.24tmklp28dcz" \h] [[HYPERLINK "https://docs.google.com/document/d/1vcpuCTV_S6ldKdXIQeCvzVLNJSxUxn1W7FZpxCvmcQ/edit"](https://docs.google.com/document/d/1vcpuCTV_S6ldKdXIQeCvzVLNJSxUxn1W7FZpxCvmcQ/edit) \l "heading=h.24tmklp28dcz" \h] [[HYPERLINK "https://docs.google.com/document/d/1vcpuCTV_S6ldKdXIQeCvzVLNJSxUxn1W7FZpxCvmcQ/edit"](https://docs.google.com/document/d/1vcpuCTV_S6ldKdXIQeCvzVLNJSxUxn1W7FZpxCvmcQ/edit) \l "heading=h.24tmklp28dcz" \h] established "source pipelines" with partners to regularly intake new leads and track identified risks. These investigations uncover behavioural signals and keywords that also inform Meta's scaled systems (e.g., machine learning) used to proactively detect CSAM. Between 2020 and 2023, Meta's teams disrupted 32 abusive networks and removed over 160,000 accounts on Facebook and Instagram associated with those networks. In August 2023 alone Meta disabled more than 500,000 accounts globally for violating our child sexual exploitation policies. And, since August 2023, Meta has automatically blocked more than 250,000 devices on Instagram for violations of its child safety policies. Improvements in device blocking have led to more than 10,000 additional enforcements on Instagram and Facebook per day.³¹
27. Meta also uses a combination of technology and behavioural signals to detect and prevent potentially inappropriate interactions between teens and adults.³² A significant focus of Meta's work is on helping keep teens safe by stopping unwanted contact between teens and adults they do not know. When teens join Instagram, Meta automatically places these teens into private accounts and does not allow people the teen does not follow to tag, to mention them, or to include their content in Reels Remixes or Guides.
28. In addition, Meta has recommendation and search filtering protocols in place to attempt to address 'bad actors to minor' and 'bad actor to bad actor' connections. Leveraging a framework to tackle Malicious Child Safety Actors ("**MaCSA Framework**"),³³ Meta launched 14 interventions to prevent MaCSA accounts from connecting with each other.³⁴ Meta also works to prevent adults identified through the MaCSA framework from finding, following or interacting with teens, and notifying teens if these adults attempt to follow or message them. For instance, Meta seeks to prevent Suspicious Adults from (i) following, liking, commenting on, replying to or sending a request to a teen account and *vice versa*; and (ii) discovering a teen profile in search type-ahead, unless they type an exact match for the profiles and *vice versa*. Further, Suspicious Adults should not be recommended content in any recommendation surface

²⁸ Further details on Meta's teams behind the reviews available at: [[HYPERLINK "https://transparency.fb.com/en-gb/enforcement/detecting-violations/people-behind-our-review-teams/"](https://transparency.fb.com/en-gb/enforcement/detecting-violations/people-behind-our-review-teams/) \h].

²⁹ Further details on How Meta invests in technology available at: [[HYPERLINK "https://transparency.fb.com/en-gb/enforcement/detecting-violations/investing-in-technology/"](https://transparency.fb.com/en-gb/enforcement/detecting-violations/investing-in-technology/) \h].

³⁰ Control ID CIDE-098.

³¹ Further details on 'Our Work to Fight Online Predators', 1 December 2023, available at <https://about.fb.com/news/2023/12/combating-online-predators/>.

³² Control IDs and CIDPA-014 and CIDE-076.

³³ For more details on the MaCSA Framework, see response to question 1(b) (paragraph 28) above and the response to question 9 below.

³⁴ These are described in more detail in response to question 1(c) (paragraph 59) below.

Private and Confidential – Contains Commercially Sensitive Information

(e.g., Reels) from unconnected teen accounts and *vice versa*; and Meta endeavours to auto-hide comments from Suspicious Adults to teens and *vice versa* and remove the message button for unconnected Suspicious Adults on teen profiles and *vice versa*.

29. Meta has specially trained teams with backgrounds in law enforcement, online safety, analytics, and forensic investigations to review potentially violating content and report findings to the NCMEC,³⁵ in accordance with applicable law.
30. Technology-driven resources also help Meta identify and take action against violating content and accounts, and prioritise enqueuing certain content for human-review.³⁶ Meta uses for instance:
 - AI and machine learning to proactively detect and take action against child sexualisation;
 - PhotoDNA and other photo- and video-matching technologies that detect identical or near-identical photos and videos of known CSAM;
 - Google's Content Safety API to better prioritise content that may contain child exploitation for human-reviewers to assess; and
 - Technology which detects and removes Instagram Reels and Stories that violate Instagram Community Guidelines, including by scanning for child sexualisation terms.
31. Finally, in 2022, Meta implemented the MaCSA Framework.³⁷ This framework uses such machine learning to proactively identify "**Suspicious Adults**".³⁸ Suspicious Adults are automatically detected using more than 60 different **Signals**, which are behavioural elements (for example, the visual elements of media represented (i) on the users' account subject line information (*i.e.*, username, profile picture, and biography), and (ii) in the content shared by users themselves) used, alone or in combination with other behavioural elements, to determine potential **Suspicious Behaviour**. Based on the likelihood of maliciousness, Meta can impose both hard actions (*i.e.*, disabling an account) and soft actions (*e.g.*, limiting the discoverability of content by bad actors) against users with high MaCSA scores. MPIL stresses that the use of these frameworks to directly impose a hard action against users is limited to the highest severity violations, such as child safety.

Meta's user reporting

³⁵ Control ID CIDPA-003.

³⁶ Further details on 'How Meta prioritises content for review', updated 26 January 2022, available at: [[HYPERLINK "https://transparency.fb.com/en-gb/policies/improving/prioritizing-content-review/" \h](https://transparency.fb.com/en-gb/policies/improving/prioritizing-content-review/)].

³⁷ Control ID CIDPA-288. The MaCSA Framework is described further in response to question 9 below.

³⁸ Also sometimes referred to within Meta as Risky Adults.

32. At the outset, MPIL notes that user reports can be inaccurate and are difficult to review at scale. Meta is continuing to assess how to improve the accuracy of its review of user reports related to CSAM and other types of misconduct. Nevertheless, in addition to its proactive detection tools, Meta also identifies CSAM through user reports. Meta encourages user reporting, not only giving agency to potential victims but it can provide helpful context to take action against bad actors. On Instagram and Facebook, people can report content or conduct they believe violates Meta's policies and flag for review. Meta has built systems and review processes to prioritise and appropriately action violating content or accounts, and report it to NCMEC or law enforcement if appropriate.³⁹ And importantly, after consultations with child safety experts and organisations, Meta has made it easier to report CSAM, by taking the option "involves a child" from within the "nudity" option in the user reporting flow on Facebook and displaying it more prominently as a self-standing reporting option. Further, when a report is made by a teen user they will then be shown an entrypoint to Take It Down (Take It Down is discussed further below in paragraph 34).
33. Meta has also developed tools specifically for teens to report if something makes them feel uncomfortable while using Meta's platforms, as well as notifications that encourage them to use these tools.⁴⁰ For example, teens are prompted to report profiles after they block someone, and receive safety notices with information on how to navigate inappropriate messages from adults. In 2022, Meta had already made it easier for people to use our reporting tools and, as a result, saw more than a 70 percent increase in reports sent by teens in Q1 2022 versus the previous quarter.⁴¹

Meta's industry-wide approach to fight child exploitation

34. Meta closely collaborates on an ongoing basis with NCMEC to protect children against abuse and exploitation. Meta's policies do not allow content that sexually exploits or endangers children. When becoming aware of apparent child sexual exploitation through Meta platforms, Meta reports it to NCMEC, who coordinates with law enforcement authorities from around the world. This partnership enables constant, open collaboration between child exploitation experts (NCMEC) and Meta, enabling proactive coordination with law enforcement and the industry. In 2023, Meta contributed financial support for the development and became a founding member of Take It Down,⁴² a new platform by NCMEC that helps prevent minor's intimate images from being disseminated online. Take It Down allows minors to submit a hash (rather than content itself) to NCMEC. Once a user submits the hash, participating platforms like Instagram and Facebook proactively search those hashes to find any copies of the image, take them down, and help prevent the content from being reposted in the future. Meta has also funded and helped rebuild NCMEC's case management tool to ensure investigators identify and respond to the most high priority reports.
35. Meta also collaborates with other professionals to better address child safety by improving its services. Meta works regularly with child safety professionals to better understand evolutions in coded language and to identify new and evolving terms, phrases, slang, and emojis that could

³⁹ Control ID CIDPA-011.

⁴⁰ Meta has created close to thirty tools to support teens on its platforms, further information available at: [[HYPERLINK "https://about.fb.com/news/2023/01/providing-safe-experiences-for-teens/"](https://about.fb.com/news/2023/01/providing-safe-experiences-for-teens/) \h].

⁴¹ See <https://about.fb.com/news/2022/11/protecting-teens-and-their-privacy-on-facebook-and-instagram/>.

⁴² Further information on "Take it Down" available at: [[HYPERLINK "https://takeitdown.ncmec.org/"](https://takeitdown.ncmec.org/) \h] and on Meta's involvement in "Take it Down" available at: [[HYPERLINK "https://about.fb.com/news/2023/02/helping-prevent-the-spread-of-young-peoples-intimate-images-online/"](https://about.fb.com/news/2023/02/helping-prevent-the-spread-of-young-peoples-intimate-images-online/) \h]. Control ID CIDPA-004.

be used in an attempt to evade its detection systems and bypass Meta's policies.⁴³ Meta also builds various interventions with these professionals and organisations, including search interventions, safety notices and safety education campaigns.⁴⁴ For instance, Meta works with public awareness experts on an ongoing basis to launch and maintain the Help Protect Children Campaign that prevents the sharing of CSAM and encourages anyone who sees potentially harmful videos and images of children to protect the victim by reporting it immediately to Meta. The campaign educates people on the harm CSAM causes, and the impact it has on victims.⁴⁵ Meta has also worked with child safety researchers to conduct collaborative research to improve child safety protections on its platforms.

36. Meta is also deeply committed to improving the ecosystem and has partnered with child safety nonprofit organisations and academic researchers to complete child safety research with a broader impact. Meta's efforts with these professionals include developing industry best practices, building and sharing technology to fight online child exploitation, and supporting victim services, among other things. Meta collaborates across the industry through organisations like the Technology Coalition, an industry association dedicated solely to eradicating the sexual exploitation of children online. In December 2022, the Technology Coalition increased its commitment to the Safe Online Research Fund, awarding eight new researchers funding to undertake innovative, actionable research for tackling online child exploitation and abuse with a focus on applications for the tech industry. And in November 2023, Meta, through the Technology Coalition, launched Project Lantern.⁴⁶ This project enables technology companies to share a variety of signals about accounts and behaviours that violate their child safety policies. The Lantern participants can then use this information to conduct investigations on their own platforms and take actions. Meta provided the Coalition with the technology infrastructure that sits behind the program and encouraged its partners to use it. This follows the launch of Project Protect in June 2020, in which Meta, Google, Microsoft and fifteen other technology companies renewed their commitment and investment to combat online child sexual abuse.⁴⁷ Meta also holds leadership positions in organisations such as WePROTECT Global Alliance (which is focused on ending child exploitation) and sits on the steering committee of Tech Against Trafficking (TAT). In 2022 the TAT Summit focused on both sex trafficking and labour trafficking.⁴⁸ Meta hosted the 2023 TAT Summit which focused on labour exploitation.
37. Finally, Meta also hosts annual child safety hackathons with their nonprofit partners.⁴⁹ This initiative brings together engineers, data scientists and designers from across the industry, who code and prototype projects focused on making the internet a safer place for children.
38. From its earliest days, Meta has developed and refined various metrics to measure its work against content that violates Meta's Community Standards and Community Guidelines, including Child Safety violations. Every quarter, Meta publishes a Community Standards Enforcement Report where it shares some of these metrics, including how much violating content has been actioned.⁵⁰ In 2022, Meta made over 5 million reports on Instagram and, in Q3 2023, it took action against 1.6 million pieces of child sexual exploitation content on Instagram, approx. 96% of this content was removed proactively by Meta before any user reported it.

⁴³ Control ID CIDE-008.

⁴⁴ Control ID CIDPA-005.

⁴⁵ Control ID CIDPA-021.

⁴⁶ See <https://about.fb.com/news/2023/11/lantern-program-protecting-children-online/>.

⁴⁷ See <https://about.fb.com/news/2020/06/fighting-child-exploitation-online/>.

⁴⁸ For more details, see <https://techagainsttrafficking.org/tech-against-trafficking-summit-leveraging-technology-to-eradicate-forced-labor/>.

⁴⁹ Control ID CIDPA-008.

⁵⁰ Control ID CIDE-111.

c) In its response of 29 August 2023, Meta states that the task force has largely fulfilled its overall goals. Please substantiate this statement and provide information on the improvement to Meta's internal processes to identify, monitor and react more quickly when dealing with child sexualisation content. Please include in your answer the specific changes that were made, the effect these changes had on the effectiveness of the detection and prevention of the dissemination of child sexualisation content, the increase in average time to respond to the detection of child sexualisation content as compared to the prior situation, and any qualitative and/or quantitative evidence, including internal benchmarks and KPIs, supporting these answers.

39. As mentioned in question 1(a), the CSTF was an immediate and targeted response that mainly focused on the key allegations of the WSJ and its work was determined by specific goals to be achieved. By the end of August 2023, the CSTF had mostly met these exit criteria and, as of September 2023, was formally wound down, with its work on assessing and mitigating CSAM being continued by Meta's Child Safety Team.
40. MPIL's Response to the EC of 29 August 2023 addressed the steps that Meta has taken through the CSTF. In the present response, MPIL provides additional detail and specificity around the CSTF's work and the improvements it brought to Meta's processes and systems. MPIL notes that the CSTF was primarily focused on reviewing Meta's processes and identifying improvements to the quality of reviews in terms of prioritisation, ranking processes, auto-close procedures, reviewer guidance, tooling, and user messaging. The CSTF was less focused on, and therefore did not track, the average time to respond to the detection of child sexualisation content, as this was not raised as a specific concern in the June report, nor is it an effective way to measure the impact of the improvements Meta introduced.
41. **For Theme 1 - Handling of user reports**, the CSTF focused on four sub-themes (i) Queue Management and Ranking; ii) Review Quality; iii) User Messaging for Reporting, and iv) Takedowns and Reporting Metrics) and made substantial progress in both Meta's content review guidance and content review process. In sum, Meta updated reviewer guidance and better prioritised user reports by incorporating additional signals into prioritisation algorithms, and identified areas for additional research in order to continue to improve enforcement from user reports. The reviewers were also provided with additional training.
42. The CSTF first identified and fixed a technical issue that was erroneously auto-closing some user reports related to child safety content.
43. Regarding the *Queue Management and ranking*, the CSTF reviewed and understood how child safety jobs were prioritised, which included checking for gaps in signals used in ranking. Meta in turn enhanced its prioritisation systems to better prioritise child sexual solicitation and sexualisation reports in its human review queues, by incorporating additional classifier signals into these prioritisation algorithms (e.g., for solicitation, CSE actor signals). Meta also improved its review tooling to assist reviewers in identifying these types of violations. These changes led to a significant increase in the number of reports reviewed that were found to violate its child safety policies.
44. Regarding the *Review Quality*, through in-person shadowing to establish root causes for reviewer error on profile and content queue, the CSTF identified some instances where human reviewers handled reports incorrectly. For instance, the CSTF conducted targeted enforcement upon CSAM using regional specific keywords and slang. This exercise found certain issues with the translation tools, which resulted in reviewers missing some CSAM solicitations. As a result of this exercise, updates were made to words known to relate to CSAM which has improved review quality.

45. Meta also launched updates across the CSEAN policies and protocols to address enforcement issues identified in the audit. These updates clarified Meta's existing policies by addressing policy ambiguity and adding more examples for reviewers to help ensure they better understand the policies and can remove more violating content as a result. Meta also provided reviewers with additional training. The training took place at the end of July 2023 and aimed to ensure the reviewers understand the latest behaviours and terms used by bad actors in this area in many different languages. Reviewers seemed to feel comfortable with the updates and were easily able to escalate CSAM and CSAM solicitation reports across all surfaces. Meta also strengthened processes to regularly update this guidance as Meta receives new information and signals as to how bad actors have adapted their behaviours to attempt to evade its systems.
46. *Reporting metrics.* Post-CSTF, Meta continues to monitor post-launch questions on these updates and document any further areas for improvement. Leveraging the WSJ's dataset and information from searches that Meta undertook, the CSTF collated a "**Golden Set**" of known CSE-violating entities. The Golden Set was then utilised to understand quality and improvement of human review. The Golden Set analysis highlighted overall improvements in recall at this scaled review stage, as well as substantial progress in (1) establishing a new integrity metric for scaled Instagram Profile reporting rep recall, (2) improving scaled quality performance of CSE content and profile review, and (3) identifying additional levers to continue driving recall improvements, but it was concluded that additional work needed to continue in some primary areas after the end of the CSTF. Work is ongoing on how to approach reporting metrics for the handling of user reports related to CS and CSx concerns by increasing the number of user reports correctly actioned for CSE violations, without exceeding an overall guardrail on enforcement precision (the rate of false positive enforcement actions). Importantly, MPIL notes that, notwithstanding the ongoing work to refine these metrics concerning the user reports, user reports are not the main method through which Meta captures child safety concerns. For example, of the content that was found to violate Meta's CSE policies, 96% was removed before any user reports were made.
47. Regarding the *User Messaging for Reporting and Takedowns*, the CSTF reviewed user messaging for reporting experiences to ensure it appropriately represents the action taken. There are situations where Meta appropriately takes action but does not notify those who initially made the report. This can happen for many reasons including when the content or account is removed for a different violation or if the account was removed by Meta's proactive technology instead of a report. As a result of the review, Meta implemented changes on Facebook and Instagram to make sure that people reporting content or accounts understand the action Meta took, whether their report prompted that action or not. Meta also updated its user messaging to provide more information to ensure that individuals receive clear, accurate and more supportive feedback when they submit a report.
48. **For Theme 2 - Enforcement against bad actors**, the CSTF focused on five sub-themes (i) Auditing Enforcement, (ii) Recidivism and hard-linked propagation, (iii) Bad actor Enforcement, (iv) Enforcement on violating profile bios and keywords, and (v) Enforcement on SG-CSAM accounts) and completed important, foundational work that resolved issues with reporting in app, SRT enforcement flows, enforcement actions applying incorrectly, logging issues, and tooling issues. In sum, Meta improved its ability to automatically disable bad actors in this space and closed gaps to limit bad actors from reappearing on its platforms as it expanded recidivism and hard-linked propagation enforcement to all Child Safety violations. Meta's strictest policies for both account linking and device disables relate to child safety violations.

49. Regarding the *Auditing Enforcement*, the CSTF undertook a review of Meta’s processes for enforcing against accounts for violations of its Community Standards and Community Guidelines related to child safety to identify potential areas for improvement.⁵¹ Through this review, it determined that the vast majority of Meta’s account-level enforcement process was working as it should, meaning that the systems correctly detected child safety-related content and then enforced against the violating account as intended. However, the CSTF also found areas for improvement, which Meta has now addressed in the other sub-themes as explained below in paragraphs 50-62.
50. Regarding the *Recidivism and hard-linked propagation*,⁵² Meta improved its policies and processes for disabling accounts on Facebook and Instagram that are linked to accounts that were previously disabled for violations of Meta’s child safety policies.
- a. The CSTF expanded the violation areas that trigger additional enforcement to include linked accounts across Facebook and Instagram as well as accounts that are newly-created by the actor in question after Meta has disabled the account.
 - b. The CSTF also undertook a review of device blocking effectiveness and identified some areas for improvement, including the impact of over-enforcement on benign accounts that are adjacent to the violating actor, (e.g., internet cafe users, shared family device). Device blocking⁵³ is one enforcement mechanism that Meta uses to prevent bad actors who have been removed from its platforms for particularly egregious policy violations from returning to the platform (“**Device Blocking**”). For these types of violations, [REDACTED]
[REDACTED] Meta only triggers device blocks where it has incredibly high precision for high severity harms. The effectiveness studies allowed the CSTF to get a better understanding of the logging and measurement to be implemented for improved Device Blocking reliability. The CSTF built a new dashboard that monitors the child safety enforcement actions as well as other device blocking metrics. The CSTF thus helped improve logging to better understand the volume and reliability of account-level actions like Device Blocking, which will help Meta monitor the efficacy of its interventions and identify potential gaps moving forward.
 - c. In parallel, the CSTF also reviewed and updated the Device Blocking coverage for child safety violations through new policies and more granular violation types which allow for more targeted, bespoke enforcement actions. Since the beginning of August 2023, Meta automatically blocked more than 250,000 devices on Instagram for violations of child safety policies, and improvements in device blocking have led to more than 10,000 additional enforcements on Instagram and Facebook per day.
 - d. The CSTF implemented a single, consolidated policy position and implementation across all child safety violations for recidivism and hard-linked propagation, which resulted in a significant increase in the number of accounts Meta disables globally every day. Since August 2023, more than 250,000 devices have been automatically blocked on Instagram for Child Sexual violations.

⁵¹ Control ID CIDE-124.

⁵² Control ID CIDE-121 and Control ID CIDE-127 respectively.

⁵³ Control ID CIDE-123.

51. Regarding the *Bad actor enforcement*, Meta made some adjustment to its MaCSA Framework⁵⁴ As a result of the CSTF's findings, The thresholds were lowered for disabling accounts on Facebook, and disables were introduced for the highest risk actors based on behavioural signals (*i.e.*, very high MaCSA score). For the quantitative evidence of the impact of these changes to the MaCSA Framework, see the response to question 9(a) below.
52. Regarding the *Enforcement on violating profile bios and keywords*, Meta deployed a text matching service ("TMS")⁵⁵ to identify and enforce on accounts which contain overtly violating subject elements (coded terms used by potential predators) across Instagram and Facebook (*e.g.*, in the account's name, username, or bio). These accounts may be automatically removed or sent for further analysis. It also launched a proactive classifier that uses combinations of signals to identify accounts with novel profile elements and content-sharing behaviour that is likely to be violating, which are then sent for human review. Human reviewers then perform a holistic review of the account and its content thanks to these new detection methods and can disable these accounts if they belong to bad actors. More than 20,000 accounts were identified and automatically removed in August 2023 as a result.
53. Regarding the *Enforcement on SG-CSAM accounts*, the CSTF planned to launch an automated enforcement targeting accounts linked to SG-CSAM. Two initial options were considered: (i) building a new classifier using the findings of Meta's investigating team (i3) for SG-CSAM advertising (depending on the precision of the results, these accounts would have been directly enforced against or enqueued for human-review) or (ii) building a new pipeline to identify minors who receive the most interactions from MaCSA users). Some initial work regarding a version of the MaCSA Framework that included teens was also conducted to better identify SG-CSAM accounts.⁵⁶ More specifically, the CSTF looked at ways to develop signals that would be appropriate for all ages (not just adults) in order to capture violations like SG-CSAM. An evaluation was launched based on signals for adults only and all ages to address the problems with lying about age. The age-agnostic MaCSA evaluation did not show a high coverage over SG-CSAM (partly due to bio elements but without many other violating signals). The CSTF tried to improve signals and measuring progress against approx. 200 known SG-CSAM accounts but it was decided to focus instead on other relevant foundational work as part of keywords, TMS and recommendations.
54. **For Theme 3 - Recommendation, searches and hashtags**, the CSTF focused on three sub-themes (i) Hashtags, type-ahead, search terms and search interventions, (ii) Recommendation filtering for accounts and content, and iii) Measuring success on an ongoing basis). In sum, after the CSTF's work, Meta now has a unified keyword list ("**Unified Child Safety Keyword List**") across platforms and interventions (*i.e.*, if Meta blocks a term it will be blocked across Meta's platforms).⁵⁷ Meta ensures it keeps this list up to date using multiple inputs. Moreover, Meta implemented extensive recommendation filtering across Instagram and Facebook places where it makes recommendations. Meta also improved the process for incorporating signals from bad actors Meta has removed from its platforms into the Unified Child Safety Keyword List and recommendation filtering.

54 For more details on the MaCSA Framework, see response to question 1(b) (paragraph 28) above and the response to question 9 below.

55 Control ID CIDE-056.

56 For more details on the MaCSA Framework, see response to question 1(b) (paragraph 28) above and the response to question 9 below.

57 The Unified Child Safety Keyword List is described in more detail in response to question 2 below.

55. The SIO and WSJ alleged in their findings that Instagram connects paedophiles and guides them to content sellers via recommendation systems that link bad actors. One of the CSTF's main goals was to break the connection between bad actors, which is difficult as they are often sophisticated bad actors who actively try to evade detection. The CSTF thus focused on disrupting the process of discovery and connection to the abusive community. As an immediate step, the CSTF restricted more than two thousand additional search terms and hashtags on Instagram and Facebook to help reduce discoverability of this type of content. It also made changes to its 'search interventions' – e.g., (a) changing the pop-up interstitials process⁵⁸ to warn users about the harms of child exploitation when searching for CSAM-related terms and hashtags, in order to prevent people from clicking through to the search results; and (b) blocking inappropriate suggestions for how to complete the rest of a search when a user starts typing.
56. Regarding *Hashtags, type-ahead, search terms and search interventions*, the CSTF also focused on longer term actions by reviewing those features within Meta's products that make recommendations to people to identify any gaps around child safety signals. The CSTF noted that there were multiple different keyword lists used to regulate hashtags, search terms and search interventions, including different lists across platforms and intervention types, which in some cases made it more difficult to regularly update the lists and led to inconsistent enforcement. To address this, Meta consolidated these keyword lists across platforms and interventions to a unified bank with tens of thousands of terms in multiple languages. Meta also updated this list based on signals Meta gathered from known bad actors that have been removed from the platform, and input from external experts.⁵⁹ By the end of August 2023, Meta had implemented this newly centralised and governed keyword solution across Facebook and Instagram. And to stay ahead of bad actors, the CSTF also developed a process for adding new keywords with inputs across 8-9 different sources.
57. Regarding *Recommendation filtering for accounts and content*, the CSTF also took steps to try to prevent bad actors from using Meta's platforms to find each other and connect with each other off-platform. Meta has implemented new recommendation and search filtering protocols to address 'bad actors to minor' and 'bad actor to bad actor' connections. As explained in MPIL's 26 May 2023 response to the EC's preliminary information-gathering exercise concerning protection of minors, Meta already had steps in place to make it more difficult for adults to find and follow teens (*i.e.*, the MaCSA Framework⁶⁰). Meta significantly expanded the types of restrictions it places on accounts identified as belonging to Suspicious Adults.⁶¹
58. Leveraging the MaCSA Framework, the CSTF also developed a new metric, "Risky Actor Connectedness", to measure connections and interactions between Suspicious Adults. This metric is defined as the percentage of virtual page views ("VPV") by Suspicious Adults that is of content created by other Suspicious Adults. Meta also launched 14 interventions that have had a substantial impact on preventing MaCSAs from connecting with each other.⁶² Specifically:
- a. Explore, Feed Recs, Reels: Content posted by MaCSA accounts is not recommendable to other MaCSA accounts;

⁵⁸ For more details, see the response to Question 8(a) (paragraph 129(b)) below.

⁵⁹ Further details with respect to the changes made to the enforcement of hashtags are provided in the response to question 2 and 3.

⁶⁰ For more details on the MaCSA Framework, please see response to question 1(b) (paragraph 21) above and the response to question 9 below.

⁶¹ Further details with respect to improvement in the recommendation filtering are provided in the response to question 4.

⁶² On the effectiveness of these measures, please see response to question 8(a) (paragraphs 130-131) below.

Private and Confidential – Contains Commercially Sensitive Information

- b. Accounts You May Follow (“AYMF”): MaCSA accounts are not recommended to follow for any other MaCSA accounts;
 - c. Comments: Comments from MaCSA accounts are auto-hidden from the view of other MaCSA accounts;
 - d. Search: A MaCSA account cannot discover another MaCSA account in the typehead or the Search Engine Results Page (“SERP”) for accounts unless they type an exact match for the MaCSA’s account name;
 - e. Following/Follower and Like Lists: MaCSA accounts are not shown in Following/Follower lists on any Instagram profile, in Like lists on any Instagram posts, when viewed by MaCSA accounts;
 - f. Profile Follow button: MaCSA Strict accounts cannot follow other MaCSA Strict accounts;
 - g. Growth Notifications: Growth notifications between MaCSA accounts are filtered;
 - h. Media SERP, Locations Pages, Hashtag Pages: Content posted by MaCSA accounts is not recommendable to other MaCSA accounts.
59. Regarding *Measuring success on an ongoing basis*, the CSTF made significant progress in terms of reducing the discoverability of problematic content and accounts on Meta’s platforms. Meta continues to evaluate and improve its measurement of discoverability
60. **For Theme 4 - Future Proofing and auditability**, the CSTF focused on five sub-themes (i) Document risk mitigation approach for minor sexual solicitation and sexualisation mitigations, ii) Design controls to validate core mitigation functionality, iii) Ensure metrics are sensitive to emerging risks, iv) Document or create human systems for validating and adapting Meta’s systems, and v) Second line of defence). In particular, the CSTF completed an analysis and concluded that, following improvements that the CSTF implemented, the human-led reviews have a stronger feedback loop to ensure stronger scaled enforcement. The CSTF also tracked the process of establishing a second line of defence for Child Safety, which has made significant progress in standing up its processes and testing relevant controls.
61. As mentioned in the Response to the EC of 29 August 2023, MPIL’s first systemic risk assessment was in the process of being conducted when the WSJ article was released. Importantly, the systemic risk assessment proceeded independently from the CSTF and correctly surfaced a number of the same issues that were identified by the WSJ – showing that MPIL’s DSA processes achieved what they were intended to. As such, the work of the CSTF took place in parallel with MPIL’s systemic risk assessment under the CSEAN Problem Area for the purpose of the DSA, but leveraged the same underlying mitigating controls.
62. While the CSTF focused on further improving the Child Safety system, Meta also focused on long term sustainability to make sure that the issues addressed by the CSTF stay addressed and that appropriate ways to monitor the follow-through on these mitigations over time exist. Meta takes the integrity of its platforms very seriously and continues to evolve its approach to identifying, managing and mitigating risks. In addition to the work of the CSTF, Meta has been working on process improvements for safety and integrity matters as a whole. These process improvements focus on: (i) improving Meta’s “first line of defence” measures (*i.e.*, the risk management processes that take place at an operational level); and (ii) strengthening Meta’s “second line of defence” measures (*i.e.*, provided by Meta’s IGRC Team, which provides operationally independent risk management and assurance). These process improvements were informed by, but not limited to, the measures taken pursuant to MPIL’s obligations under the DSA.

2. The 7 June report suggests that interested buyers could easily identify seller accounts based on text or hashtag search. In the 29 August reply by Meta, “recommendations, searches, and hashtags” were identified as a key issue that needed to be addressed to remedy the shortcomings identified in the Stanford study. In line with the process of restricting hashtags as outlined by Meta in its 29 August reply, the 21 September follow-up study notes that many hashtags that were used during the first report have been restricted on Instagram.

However, the researchers stretch that “the overall ecosystem remains active, with significant room for improvement in content enforcement”. The SG-CSAM network makes use of hashtags with one or few emojis, which are hard to ban on a general level. The follow-up study marks that a circumvention of forbidden hashtags takes place on Instagram and refer to alleged difficulties in hashtag-enforcement. In this respect:

a) Meta explains that it has restricted ‘more than two thousand additional search terms and hashtags’ on its services to restrict discoverability of SG-CSAM in its response of 29 August 2023. Please outline how these search terms and hashtags have been identified (e.g., whether they correspond (totally or partially) to the terms identified by the Stanford study or whether they are the result of an internal assessment of Meta).

63. The additional search terms and hashtags were identified as the result of the CSTF’s work, which included adding keywords based on accounts, content and hashtags associated with violating content, or where Meta had intelligence from external or internal experts that certain terms or emojis were being used in connection with CSAM solicitation. After the findings of the SIO and the WSJ, one of the main tasks of the CSTF was to improve the tools that disrupt the ability of bad actors to discover and connect to abusive communities.

64. The CSTF first audited the existing Child Safety-related hashtag and search enforcement to test their (expected) efficacy. The CSTF thus audited approximately 35 thousand terms for Facebook and 7 thousand terms for Instagram. Then, and as mentioned in paragraph 54, the CSTF consolidated the different keyword lists used to regulate hashtags, search terms and search interventions to a Unified Child Safety Keyword List, with tens of thousands of terms in multiple languages. The Unified Child Safety Keyword List is continuously being updated, and as new terms are added to the list, they are restricted across Facebook and Instagram simultaneously.

b) Please outline and explain any structural changes made to the enforcement of hashtags on Instagram, with a specific focus on SG-CSAM and CSAM, following the report of 7 June 2023 and/or the follow-up study on 21 September 2023.

65. Following the 7 June 2023 report, the CSTF consolidated the different keyword lists that existed across platforms and intervention types to a Unified Child Safety Keyword List, and created new processes for updating it (e.g., automated approaches to ingest new keywords). Further details on the changes brought by the CSTF are provided in paragraphs 54 and 64 above. At the same time, Meta launched four TMS banks aiming to proactively detect usage of inherently violating child safety keywords and phrases within Instagram profile elements and enhanced the existing MaCSA Framework signal that captures more violating keyword/hashtag engagement.

66. Meta continues to improve the enforcement of hashtags on Instagram, with a focus on SG-CSAM and CSAM. Meta has established protocols for ongoing maintenance of the Unified Child Safety Keyword List. Meta continues to evolve its approach to this beyond the CSTF and drive keywords improvements. In September 2023, Meta identified additional hashtags by looking at the nearest hashtags of new known violating hashtags. From these new hashtags, approx. 30 thousand hashtags surfaced that were frequently used/created/followed together, when processed by Meta's internal systems. Over 3,000 hashtags were prioritised for review and 265 keywords were added to the Unified Child Safety Keyword List.
67. Based on these 265 keywords, Meta re-ran the same exercise and surfaced an additional 308 keywords, with a secondary list of over 260 keywords considered for addition to the Unified Child Safety Keyword List. Notably, the secondary keyword list was of lower quality, indicating that Meta was exhausting some of the more obvious keywords proactively. Further, the action rates on the keywords showed significantly diminished returns as compared to when that exercise was undertaken at the start of the CSTF. This indicates that Meta by this stage had a better coverage of obviously CSE-related keywords. Meta also launched, in December 2023, an expanded "fuzzy" matching (*i.e.*, matching non-exact variations of terms on Meta's keywords lists) for hashtags and search terms using TMS.⁶³
- c) **Meta mentions that "Meta has also put in place processes to update this unified list regularly as the terminology that bad actors use changes and as Meta identifies new signals from bad actors that it removes from its platforms and from external input." Please indicate how frequently Meta updates the mentioned list of hashtags and the specific internal procedures, safeguards, and policies in place to ensure that the mentioned list of restricted hashtags is representative of the consistently changing practices within the network of SG-CSAM. Please include in your answer any documentation evidencing updates to the list, consultations with third-party experts, or other information that concerns the constitution of the list of prohibited hashtags concerning SG-CSAM.**
68. The CSTF put in place the following key processes that govern and maintain the Unified Child Safety Keyword List:
- a. **Inputs:** Meta maintains existing and receives new keywords and hashtags from different sources within Meta and external third-party experts.
- i. **Meta's investigation teams:** Coded language related to CSAM and CSAM trading/solicitation is identified by investigators during SNDs and investigations (i3 Team, Law Enforcement and Investigation team). 1 to 10+ new terms are typically identified per SND/investigation. For an example of this, please see the response to Question 2(b) above.
- ii. **I3 Child Safety Risk Review:** The I3 team's risk review process provides regular leads coming from a variety of sources (such as intel vendors, open source research, investigator red teaming, etc.) that are likely to contain or point towards trending/emerging CSAM-related keywords.
- iii. **Monthly sweeps:** Discovery Intelligence Ops ("DIOps") conduct targeted sweeps via manual on-platform reviews. Search terms will consist of hashtags and queries related to CSAM/CSE. DIOps will monitor the discoverability of violating content on Instagram Search by reviewing the top 25-30 results across tabs (For you, Accounts & Tags).
- iv. **In-house detection tooling:** Meta employs leverage tooling that takes an input (search term) to identify misspellings, spelling variations, search terms with

⁶³

Further details on the TMS are provided in response to Question 8(a).

common media results, and search terms users searched in the same search session.

- v. **Query backtracker tooling:** Meta employs leverage tooling that associates all accounts and media that have been deleted for violating CSAM policies and identifies all hashtags and search terms that return those violating accounts and media as a search result.⁶⁴
 - vi. **Markets inputs:** Meta will validate each month with Market subject-matter experts (“SMEs”) if any market-specific new terms are being surfaced by the team. These keywords will be flagged in the internal teams as marketised terminology that may warrant addition to the keywords list.
 - vii. **External inputs:** Meta also becomes aware of new keywords via trusted child safety partners including ingesting lists from third parties (e.g., Thorn⁶⁵), keywords shared *ad hoc* by trusted child safety partners, industry partners, or surfaced in external reports.
- b. **Validation:** Different teams within Meta (Safety Policy, Product Policy and Core Policy) review and approve new keywords to be added to the Unified Child Safety Keyword List. This step can also involve the removal of keywords subject to some criteria, such as [REDACTED]
- c. **Technical updates:** Once approved, Meta updates the internal repository of keywords lists.
- d. **Application:** Meta ensures that the updates to the Unified Child Safety Keyword List are applied to search/hashtags intercepts, account scores and review queues.
69. The following table sets out the different sources, committed cadence and expected volume of updates to the Unified Child Safety Keyword List, to ensure that the list remains representative of the latest terminology used by bad actors.

⁶⁴ The Query backtracker tooling is also discussed in paragraph 79 below.

⁶⁵ Thorn is a nonprofit focused on fighting child sex abuse, further information available at: [HYPERLINK "https://www.thorn.org/" \h]

Source	Cadence	Expected volume
I3 team: Regular SNDs and investigations	Monthly	From 1 to 10+ new terms per SNI
I3 team: Intel vendors	Possibly every Quarter	Approx. 1 to 10+ new terms
LEI: Regular investigations	Monthly	From 1 to 10+ new terms per SNI
DIOps: Monthly sweeps	Monthly	100 query per analyst
DIOps: In-house detection tooling	Monthly	Cap at top 10
DIOps: Query backtracker tooling	Quarterly	Cap at top 100 search term
GRO: Market inputs (TBD)	Possible coordinated Monthly	TBD: From 1 to 10+ new terms per SNI (pending first test batch)
Policy: External Stakeholders	Ad hoc	Variable
Product: ML Sourced	Ad hoc	Variable

70. Meta incorporates language tools such as XLM and XML-R, aimed at building classifiers that understand the same concept in different languages.⁶⁶
71. In response to the request for documentation evidencing updates to the Unified Child Safety Keyword List, consultations with third-party experts, or other information that concerns the constitution of this list, MPIL has identified the following documents that are responsive to these specific descriptions and relevant to its compliance with Articles 34 and 35 DSA:
- An internal post to the Workplace group of the Child Safety Task Force of 6 October 2023 entitled '*[Update] Keywords/Hashtags review on WSJ escalation*', Annex 3,
 - An internal post to the Workplace group of the Child Safety FYI of 2 January 2024 entitled '*[LAUNCH] Fuzzy matching for Child Safety Keywords*', Annex 4, and
 - A copy extracted on 13 January 2024 of an internal document '*Ongoing maintenance of child safety keywords*' which sets out the process to surface new keywords across Meta's cross-functional teams, covering input, validation, technical updates and monitoring, Annex 5.

⁶⁶ For a full description of these classifiers, please refer to Meta's blog page of 7 November 2019 '*XLM-R: State-of-the-art cross-lingual understanding through self-supervision*' available at <https://ai.meta.com/blog/xlm-r-state-of-the-art-cross-lingual-understanding-through-self-supervision/>.

d) In its response of 29 August 2023, Meta outlines that the hashtag-list for (SG-)CSAM content exists in ‘multiple languages’. Please indicate if this list exists in all official language of the European Union and explain, if this is not the case, why these languages have been chosen.

72.

[REDACTED] It also includes multiple languages that are not official languages of the European Union, such as Arabic and Kurdish. The languages on the list reflect feedback that Meta receives from external and internal experts, which informs the terms that appear on the list. As more feedback becomes available for additional languages, Meta may expand the terms to those languages.

73. MPIL notes that, more generally, Meta’s CSAM content review is language agnostic. First, Meta’s classifiers⁶⁷ and machine learning models run across content in all languages, and are supported by an AI team, which searches across all language models for violations. Second, Meta’s human review is language agnostic: English-language reviewers use translation tools when needed to enable Meta to scale content reviews across other language markets, and reviewers are encouraged to assess content with market context (e.g. specific terms used to refer to CSAM) .

3. The Stanford follow-up study suggests proactive monitoring and discovery through human investigators, including effective tracking and actioning of relevant hashtags and keywords, as well as changing iconography, as effective strategies to track the evolution of trends in the SG-CSAM network. The researchers mark that ‘assessing emojis that take on high salience within the SG-CSAM networks and flagging them for a comprehensive review would be a reasonable action without too much user impact.’

a) Please outline in detail the resources Meta has put in place, including the number of human investigators, to monitor SG-CSAM related hashtags specifically for Instagram.

74. As described in response to question 2 above, Meta has consolidated multiple different keyword lists used to regulate hashtags into a Unified Child Safety Keyword List, with tens of thousands of terms in multiple languages, and put in place processes to regularly update this list to reflect changes in terminology by bad actors. This draws on inputs from a range of sources, including Meta’s i3 and law enforcement and investigation teams, intel vendors, Meta’s specialist child safety teams, machine learning technology, non-profit organisations, and experts in online safety.

75. Meta’s i3 team includes ‘subject matter expert’ investigators, who review content, hashtags, keywords and iconography in English, as well as specialist intelligence analysts, and security engineering investigators. Together with the other Meta teams listed above, the i3 team has conducted two SNDs on the SG-CSAM network.⁶⁸ Meta is currently conducting additional investigatory work in other language markets in relation to this issue. The results of these investigations are shared with the Cross-Meta Integrity and Product teams.

⁶⁷ For more detail on classifiers, see the response to question 8(a) (paragraph (129(d)) below.

⁶⁸ For more information on SNDs, see Question 1(b) above and for more information on these two specific SNDs, please see Question 5(c) below.

76. As explained in detail in response to question 1(b), above, Meta relies on both automated technology and reports to take action on violating hashtags, account names, search terms, and emojis. Meta works to ensure that inherently violating terms, and other terms that are not inherently violating, but may be used by bad actors seeking inappropriate content, are not treated by Instagram as valid hashtags, and Meta works to ensure that it does not show search results for these terms. Because Meta recognises this is a constantly evolving area, it also works regularly with child safety professionals to understand evolutions in coded language and to identify new and evolving terms, phrases, slang, and emojis that could be used in an attempt to evade Meta's detection systems and bypass Meta's policies.

77. These ongoing efforts encompass emojis as well as text-based hashtags. Meta has identified and taken action on emojis that take on high salience within SG-CSAM networks. For example, the list of hashtags which have been blocked on Instagram includes not only hashtags which clearly relate to CSAM (such as [REDACTED] but also those which reflect the use of coded language within CSAM networks (such as [REDACTED] or [REDACTED]⁶⁹

b) To what extent does Meta facilitate signal sharing between Instagram's trust and safety teams and changes for recommendation systems and machine learning models to better detect keywords, symbols and emojis that are misused by the network, and to see whether they are still used?

78. MPIL refers to its response to Question 2 above in relation to the Unified Child Safety Keyword List. Meta uses the results of its automated technology and input from child safety professionals to train its technology so that it can proactively find new trends in behaviour and coded language and take action. Meta's teams also use technology to find misspellings and spelling variations of this language, as well as terms and phrases related to child exploitation, that it can input into its systems to proactively find and disrupt efforts by bad actors who evade Meta's protections and policies.

79. For example, Meta conducts regular 'Query backtracker tooling', whereby it takes accounts and media that have already been deleted for violating Meta's CSAM policies, and endeavours to identify all hashtags and search terms that, if searched for, would return those violating accounts and media as a search result. In this way, Meta is able to use technology, together with the findings of its trust and safety teams, to identify new hashtags and search terms that might otherwise be missed by human reviewers.

80. Meta continues to update its unified list of restricted hashtags on Instagram. For more details please see the response to Question 2 above.

4. The 7 June report states that certain platform features, such as recommender systems can lead to the discovery of small networks of SG-CSAM on Instagram.

a) Please provide information, including internal documentation, on the testing of Instagram's recommender system, relevant test processes implemented, metrics used in relation to SG-CSAM, and their most recent values.

81. As indicated above in response to question 1(b), Meta's policies – such as the Instagram Community Guidelines⁷⁰ and the Facebook Community Standards⁷¹ – do not specifically address the different types of CSAM (e.g., self-generated or "known"). Meta's general prohibition against CSAM applies equally to SG-CSAM.

⁶⁹ The term 'map' is sometimes used within CSAM networks to refer to a 'minor-attracted person'.

⁷⁰ Instagram Community Guidelines available at: <https://help.instagram.com/477434105621119>.

⁷¹ Facebook Community Standards available at: <https://transparency.fb.com/en-gb/policies/community-standards/>.

82. In response to the request for internal documentation on the testing of Instagram's recommender system, MPIL has identified an internal post to the Workplace group of the Child Safety XFN group of 13 December 2023 entitled 'Tracking CSE on Recommended surfaces', [Annex 6](#).
- b) How does Meta test its systems to avoid any potential future problems related to the offering and distribution of SG-CSAM? Please provide information, including internal documentation, showcasing Meta's actions related to the prevention of (offering and selling) SG-CSAM content on Instagram.**
83. Meta is continuously developing and updating its recommender systems and other tools to ensure that they are well-prepared to respond to new and evolving risks and bad actors. Although Meta already has numerous measures and systems in place to act as 'first line' defences aimed at mitigating child safety risks, it is further strengthening those defences by improving its mechanisms to adapt faster as new risks emerge.
84. As detailed in response to Question 1(a), Meta reviewed the work of the CSTF to test Meta's recommender systems and try to ensure that the systems are prepared for future threats, and ultimately concluded that the CSTF had largely fulfilled the exit criteria, in particular those measures addressing CSAM (including SG-CSAM). Moreover, as also detailed in response to Question 1(a), several controls were put in place specifically to detect CSAM and SG-CSAM as part of the CSTF's work, and insights from Meta's previous efforts to improve scaled detection of SG-CSAM helped inform CSTF strategy. Beyond the work of the task force, Meta continues to drive a range of initiatives to test its recommender systems and ensure that they are prepared for future threats. These include the following:
- a. **Designing controls to validate child safety mitigation functionality.** Meta teams are actively working to deliver new controls to help mitigate minor sexual solicitation and sexualization, and are also actively building directly analogous overall child safety mitigations.
 - b. **Ensuring metrics are sensitive to emerging risks.** Meta teams are identifying the optimal metrics for measuring emerging risks, and are iterating on existing metrics design to ensure sensitivity to these risks
 - c. **Creating human systems for validating and adapting Meta's risk mitigation tools.** Meta teams are manually investigating sensitive risk areas, updating internal processes for responding to high-risk incidents, and incorporating their learnings into existing mitigation strategies.
85. In response to the request for internal documentation showcasing Meta's actions related to the prevention of (offering and selling) SG-CSAM content on Instagram, MPIL provides an internal note 'Child Safety Network Disruption on Commercial CSAM Networks', provided at [Annex 7](#).
86. This note sets out the second SND⁷² on commercial CSAM networks conducted on 6 November 2023, which is referred to in response to question 5(c) below. Meta's investigation team used the 242 accounts identified by SIO, which included 161 accounts of suspected SG CSAM sellers; 11 accounts of suspected minor SG CSAM sellers and 70 accounts suspected SG CSAM buyers, to uncover a broader community of over 50,000 CSAM-related accounts on Instagram, over 35,000 of which were still active at the time of the investigation, involved in the selling, purchasing, and consumption of SG CSAM.

⁷² SNDs are takedowns of entire networks of bad actors, carried out by Meta's investigation teams based on information received off-platform, described in Question 1(b) above and 5(c) below.

87. The note sets out the enforcement action that was taken as part of the SND, with different actions calibrated according to the severity of the behaviours exhibited by the accounts, from disabling accounts and device blocking for Tier 1 (coordinating tier) accounts, through to educational messages for Tier 4 (passive observer tier).

c) Please outline specific preventive measures Meta takes for the non-consensual re-distribution of initially voluntarily disseminated SG-CSAM content.

88. As explained in response to Question 4(d), it would be difficult to target the redistribution of initially voluntarily disseminated SG-CSAM content, since it is difficult, if not impossible to ascertain whether the SG-CSAM was initially voluntarily distributed.

89. Rather, Meta's focus is on identifying and stopping the initial distribution of any CSAM (including SG-CSAM), and on quickly and effectively preventing and detecting any content that violates Meta's child safety policies.

90. With regard to preventing distribution of potential CSAM content, Meta works constantly to find, remove, and report policy-violating content, deactivate accounts of predators, and dismantle abusive networks. Meta also continually explores new ways to actively tackle child exploitation to ensure that it stays ahead of predators' evolving efforts to evade its detection and enforcement. Meta takes the following preventive measures with regard to child exploitation:

- a. **Proactively detecting and removing accounts that violate child safety and human exploitation policies.**⁷³ Meta has expanded upon its existing technology to detect Instagram accounts exhibiting potentially suspicious behaviour, find coded terms used by Instagram accounts belonging to potential bad actors, and disable Facebook and Instagram accounts if they exhibit certain signals Meta monitors for potentially suspicious behaviours.
- b. **Making sensitive content more difficult to find.** As explained above in relation to the Unified Child Safety Keyword List, Meta has restricted thousands of search terms on Instagram and Facebook, and has updated its processes for keeping these lists of terms up-to-date as their usage evolves.⁷⁴ Meta has also updated the pop-ups it developed to warn people about the harms of child exploitation when searching for child exploitation related terms and hashtags, ensuring that users cannot click through these pop-ups to see results.⁷⁵
- c. **Proactively stopping unwanted contact between teens and unknown adults.** Meta takes various measures to help ensure that teens are protected from unwanted contact with adults they do not know.⁷⁶ For example, Meta does not allow teens on Instagram to be followed, tagged or mentioned, by people they do not follow, or for their content to be included in Reels Remixes or Guides.

⁷³ Meta's human exploitation (HEX) policy includes minor sex trafficking, and other issues impacting minors such as labour trafficking, child soldiers, sales of children or illegal adoption, orphanage trafficking and orphanage volun-tourism.

⁷⁴ For details on the Unified Child Safety Keyword List, see the response to question 2 above.

⁷⁵ For details on updates to pop-ups, see the response to question 8(a) below.

⁷⁶ For details see the response to question 1(b) (paragraphs 28-31) above.

- d. **Identifying accounts belonging to users demonstrating potentially suspicious behaviour.** Through the MaCSA Framework,⁷⁷ Meta has developed technology to identify accounts demonstrating potentially suspicious behaviour and apply a number of restrictions to them, including limiting their ability to engage with teens and other MaCSA accounts, up to and including disabling the accounts.
 - e. **Prevent potentially suspicious accounts from being recommended.** Meta takes steps to prevent accounts that belong to Suspicious Adults from being recommended to anyone as an AYM. These accounts are not recommended to each other in Search or AYM, and are not able to follow each other, see comments by other Suspicious Adults, or see each other's accounts on surfaces where Meta recommends content, such as Explore and Reels.⁷⁸
 - f. **Providing content reviewers with new guidance and training to ensure they can remove more violating content and behaviour.** As well as training content reviewers,⁷⁹ Meta is also making updates to the tools that content reviewers use to review content to minimise errors on an iterative and ongoing basis with the aim of improving the overall goals of reducing violating content exposure, removing violating accounts and improving user report response.
- d) **Please outline the number of content moderation actions Meta has taken for non-consensual re-distribution of initially voluntarily disseminated SG- CSAM content as of 29 August 2023. Please break-down these numbers by actions based on a notice in accordance with Article 16 Regulation 2022/2065 and own-voluntarily content moderation.**
91. MPIL is unable to calculate the number of content moderation actions it has taken for non-consensual re-distribution of initially voluntarily disseminated SG-CSAM content, because it does not have the specific data points that this calculation would require. Specifically, MPIL is unable to determine whether SG-CSAM was initially voluntarily distributed. For metrics regarding content removed for violations of Meta's child safety policies, please see MPIL's transparency reports for Instagram and Facebook, published 27 October 2023,⁸⁰ which provide metrics for: (i) content moderation actions taken by type of violation; and (ii) content moderation actions taken in response to notices provided in accordance with Article 16 DSA.
- e) **Please indicate in which category of 'organic content policy violations' or 'type of reported illegality' the number under question 7(d) is included in Meta's transparency report for Instagram, published on 27 October 2023.**
92. MPIL understands from its meeting with the Commission of 2 December 2023 that the reference to Question 7(d) should in fact be to Question 4(d). As detailed in its response to Question 4(d), Meta is unable to calculate the number of content moderation actions it has taken for non-consensual re-distribution of initially voluntarily disseminated SG-CSAM content.
93. However, content moderation actions taken on the basis that the content is SG-CSAM are contained in the 'Child Endangerment - Child Sexual Exploitation' category of organic content policy violations.

⁷⁷ For more details on the MacCSA Framework, please see response to question 1(b) (paragraph 28) above and the response to question 9 below.

⁷⁸ For more details see response to questions 1(b) and (c) above, and question 9 below.

⁷⁹ For more details see response to question 1(c) above.

⁸⁰ See Instagram transparency report, published 27 October 2023, available at <https://transparency.fb.com/sr/dsa-transparency-report-oct2023-instagram/>, and Facebook transparency report, published 27 October 2023, available at <https://transparency.fb.com/sr/dsa-transparency-report-oct2023-facebook/>.

94. Any SG-CSAM content reported pursuant to Art. 16 of the DSA is first reviewed for policy violations, and may be actioned on that basis. In circumstances where Meta's CSAM policies align with what might be construed as illegal content in this space and potentially go further (for example, we remove child nudity to protect minors under our policies), most if not all illegal SG-CSAM content will be removed as policy violating in the first instance. In the unlikely event that any SG-CSAM content reported under Art. 16 is deemed to be non-policy violating, such content will undergo a review for illegality. Any SG-CSAM content that is actioned on the basis that it is illegal (but not policy violating) content would be recorded in Table 15.1.b.(1) under "Other illegal content".

95. The phrase "type of reported illegality" is not used in relation to content moderation activities taken in response to Art. 16 notices, or on an own-voluntary basis, in MPIL's transparency report for Instagram. Rather, it is used in categorising Member States' authorities' orders to provide information about a user, under Article 10 DSA. The phrase would also be used in categorising Member States' authorities' orders to act against illegal content on Instagram, but Instagram did not receive any Article 9 orders during the report period. For more details, please see Table 15.1.a.(2) in MPIL's Transparency Report for Instagram, published on 27 October 2023.

5. The 7 June report identified 405 accounts advertising the sale of SG-CSAM. Meta outlined in its 29 August response that "there were some weaknesses in Meta's enforcement that resulted in certain bad actors not being enforced on correctly." In the 21 September report, 162 seller accounts were identified.

a) Please explain whether and to what extent Meta acts against seller accounts of SG-CSAM. Please explain the policy underlying those actions and outline potential actions and the average time to action on seller accounts.

96. As explained in relation to Meta's policies and processes in response to Question 1(b) above, Meta prohibits users of Instagram and Facebook from selling, sharing and distributing CSAM and it works aggressively to fight child exploitation off its platforms. However, Meta's policies do not specifically address the different sub-types of CSAM (e.g., SG-CSAM). However, in the case of CSAM sellers, Meta has different actions depending on whether a user is selling CSAM or a minor is sharing their own SG-CSAM. For completeness, Meta presents the actions taken in both cases.

97. Accounts that sell CSAM are disabled either via the MaCSA Framework or upon Meta becoming aware of a single malicious share of CSAM. For the latter Meta automatically applies a one-strike disable (one-and-done violation type). Enforcement is then propagated to any linked accounts, associated devices are blocked, and users are added to recidivism monitoring for enforcement should they try to recreate an account.⁸¹ For certain high risk violations, Meta can enforce across multiple simultaneously-active linked accounts through "Hard-linked propagation" and take a device blocking action.⁸² Finally, the linked accounts are enforced under Meta's Recidivism Policy and corresponding tools (such as the Generic Unified Adversarial Recidivism Detection System). These actions relate to the following CSAM-related SRA controls: CIDE-124 Disabling Accounts, CIDE-125 Disabling Pages and Pages (if relevant), CIDE-123 Device Blocking, CIDE-121 Recidivism.

⁸¹ For more information on Recidivism, please see the response to Question 8(b) below.

⁸² For more information on Device blocking, please see the Response to Question 1(c) above.

98. Meta has a narrow exception for teens sharing their own SG-CSAM. Instead of applying a one-and-done disable, Meta will disable the account only on the third violation. Meta chose not to apply an automatic disable of a minor's account in order to provide an opportunity for education. However, this does not apply if (i) Meta suspects an adult is lying and posing as a child or (ii) there is a commercial element, as it will be actioned for CSAM Solicitation (which constitutes a one-and done violation type). In these cases, the actions (*i.e.*, one-strike disable, hard-linked propagation, device blocking and recidivism monitoring) described above for users selling CSAM apply. These actions relate to the following CSAM-related SRA controls: CIDE-133 Strike System, CIDE-124 Disabling Accounts, CIDE-125 Disabling Pages and Pages (if relevant), CIDE-123 Device Blocking, CIDE-121 Recidivism.
99. Regarding the average time to take enforcement action on seller accounts, this will depend on how the seller account is identified, and the nature of the violation. Content identified by classifiers will normally be auto-actioned immediately. However, user reports will be enqueued to specialist reviewers. When CSAM content is escalated to the Child Safety review queue, the human reviewers have an informal, self-imposed service-level agreement of 48 hours to review this SRT job. Moreover, since August 2023, enforcement can be immediate under Meta's Recidivism Policy. Once an account is disabled, any recreation, no matter how long after the initial disable, is violating Meta's policy.
- b) Please explain whether and to what extent Meta applies proactive and/or preventive filtering to detect seller accounts on Instagram. If any filtering measures are applied, please explain their functioning, effectiveness and error rates, including the internal benchmarks and KPIs.**
100. Meta's proactive detection tools for CSAM also apply for seller accounts on Instagram. Content on Instagram flagged by the proactive detection tools and user reports as potentially violating is sent as an SRT job to Meta's content review system. The SRT job is then prioritised based on likelihood of violation, virality, and violation severity. However, if content is identified as CSAM or CSAM solicitation, it is escalated for review by a ring-fenced human review team specialising in Child Safety. These systems concern the following CSAM-related SRA controls: CIDPA-228 MaCSA Framework, CIDE-52 Landing Page Detection Service, and CIDE-48 Detection AI / Machine Learning Models & Classifiers.
101. Another detection system for users selling CSAM or adults posing as a child when selling SG-CSAM is the MaCSA Framework⁸³; the tools to detect keywords and URLs⁸⁴ and the CSAM-related classifiers that enqueue content for human-review, trigger search interventions, and create internal signals into the prioritisation algorithms to proactively detect bad actors.⁸⁵ For instance, accounts with confirmed general CSE terms such as [REDACTED] are automatically enqueued in the Instagram Profile review queue. Meta has also put in place a CSAM solicitation content filter, which functions as the CSAM-related classifiers. It enqueues content for review within the SRT that is above a score threshold indicating a likelihood of being CSAM solicitation.⁸⁶ Further details on the functioning of the detection tools are provided in the responses to questions 1(b), 8 and 9.

⁸³ Control ID CIDPA-288

⁸⁴ Control ID CIDE-052; For more information on URLs, please see the response to Question 8(a).

⁸⁵ Control ID CIDE-048.

⁸⁶ Control ID CIDE-099.

102. As noted in response to question 1(b) above, MPIL’s risk scoring and control design effectiveness scoring for the relevant mitigating controls for the risks associated with CSAM are contained in the CSEAN Workbook.⁸⁷ As well as the specific L4 risk for CSAM, other related risks were considered. The effectiveness and proportionality of each control in relation to each of these risks was assessed (in the tab ‘Detailed Control Scoring’). In relation to each risk, MPIL identified the Inherent Risk (before the application of the mitigating controls) and then Residual Risk, applying a percentage reflecting the Design Effectiveness and Mitigation Effectiveness of the suite of relevant mitigating controls. The Residual Risk score is the Inherent Risk score, less the product of the Control Suite Effectiveness and the Inherent Risk (in the tab ‘Risk Register & Scoring’). The following table sets out the overview of the assessment of the design effectiveness of controls related to these risks on Instagram, extracted from the CSEAN Workbook provided.⁸⁸
103. Furthermore, MPIL notes that it is in the process of developing an additional parameter of ‘Operating Effectiveness’, on top of the Design Effectiveness and Mitigation Effectiveness described above, with a view to applying this in the SRA to be carried out in 2024. In this context, it has prioritised the testing of a number of controls, including controls relevant to the Problem Areas of CSEAN, and this exercise is ongoing.

Problem Area	L4 Risk	Inherent Risk	Control Suite Effectiveness ⁸⁹	Residual Risk ⁹⁰
CSEAN	Child Endangerment	15.0	68%	4.8
	Child Nudity	14.0	63%	5.2
	Child Sexual Abuse Materials	15.0	63%	5.6
	Inappropriate Interactions with Children	15.0	68%	4.8
	Sexualization of Children	15.0	63%	5.6
	Solicitation of Child Sexual Abuse Materials	15.0	63%	5.6

⁸⁷ See Annex 10 of the 15 December Response.

⁸⁸ See Annex 10 of the 15 December Response.

⁸⁹ See page 8, Control Effectiveness Rubric, SRA Methodology, Annex 4 of the 15 December Response.

⁹⁰ Tier 1 (0-4.9) ‘Extremely Low’; Tier 2 (5-9.9) ‘Low’; Tier 3 (10-14.9) ‘Moderate’; Tier 4 (15-19.9) ‘Elevated’; Tier 5 (20-25) ‘Extremely Elevated’; See page 11, Residual Risk: Tiering, SRA Methodology, Annex 4 of the 15 December Response.

c) Please explain when these 162 seller accounts were created and whether these accounts were newly created accounts since the June report. Where possible, please indicate to what extent the 405 identified accounts in June were actioned upon and which actions were taken.

104. On 28 April 2023, the SIO provided Meta with a first list of accounts. The SIO had identified 448 of these as involving SG CSAM transactions (391 as suspected sellers and 57 as suspected buyers).⁹¹ Meta's i3 team⁹² reviewed these accounts and disabled 397 of them (88.6%) for child safety violations.
105. In addition, Meta used all 448 accounts identified by the SIO as a seed set for a SND. Throughout the course of the investigation, Meta discovered a community of over 82,000 Instagram accounts involved in selling, purchasing and consuming SG CSAM. Of these accounts, around 61,000 were active at the time of the investigation and 216 were SG CSAM seller accounts that had not been originally flagged by the SIO. As a result, a total of 14,434 Instagram accounts were disrupted and 39,638⁹³ accounts received custom messaging. Other actions included reporting CSAM content to NCMEC⁹⁴ and adding it to the appropriate MMS bank; deleting any explicit CSx content and adding it to the appropriate MMS bank⁹⁵; reporting any CSAM solicitation posts and comments to NCMEC; and sharing new coded language and other investigative insights with Meta's Child Safety Task Force for additional targeted detection and enforcement.
106. On 21 August 2023, the SIO provided Meta with a second list of 242 accounts.⁹⁶ These were identified as suspected SG CSAM sellers (161 accounts), suspected minor SG CSAM sellers (11 accounts), and suspected SG CSAM buyers (70 accounts). Meta's i3 Child Safety team reviewed these accounts and took the following actions:
- a. Of the 161 suspected (adult) sellers accounts:
 - i. 64 were already disabled at the time of the review
 - ii. 97 were active at the time of the review
 - 1. 80 were disabled for child safety violations
 - 2. 13 were disabled for SSPx/ANSA⁹⁷ violations
 - 3. 4 did not involve child safety violations
 - b. Of the 11 suspected minor sellers:
 - i. 7 were already disabled at the time of the review
 - ii. 4 were active at the time of the review
 - 1. 2 were disabled for child safety violations

⁹¹ It is not clear why the SIO June report contained slightly different figures.

⁹² For more information on the i3 team, please see questions 1(b) and 5(a).

⁹³ A total of 46,758 accounts were originally identified as a suitable cohort to receive custom messaging. In the time between the investigation and the deployment of custom messaging, 7,120 accounts were disabled. Therefore, custom messaging was deployed to 39,638 accounts.

⁹⁴ For more information on Meta's co-operation with NCMEC see Question 1(b) above.

⁹⁵ For more information on the MMS bank, please see response to question 1(b) above.

⁹⁶ It is not clear why the SIO September report contained slightly different figures.

⁹⁷ SSPx - Adult Sexual Solicitation and Prostitution; ANSA - Adult Nudity and Sexual Activity

2. 2 were disabled for SSPx/ANSA violations

107. As regards the dates on which these accounts were created: the 161 accounts identified as suspected (adult) sellers were created between 17 September 2017 and 1 August 2023; the 11 accounts identified as suspected minor sellers between 25 April 2017 and 13 August 2023. However, MPIL notes that the date of account creation is not a meaningful metric, since it cannot be assumed that the accounts engaged in violating behaviour from the outset (and it is unclear when they did begin to engage in violating behaviour). Moreover, MPIL notes that none of the 172 suspected seller accounts had been flagged by the SIO in its June report or the list it provided in April 2023. (Despite this, 15 of the 161 suspected seller accounts and 1 of the 11 suspected minor seller accounts had been captured in the SND described above).
108. After taking the above actions, Meta used the 242 accounts identified by the SIO as potential seller and buyer accounts as the seed set for a further SND. Through this investigation, Meta identified a broader community of over 50,000 Instagram accounts, over 35,000 of which were still active at the time of the investigation, involved in the selling, purchasing, and consumption of SG CSAM. This included an additional 84 seller accounts not flagged by the SIO. As a result, a total of 8,057 Instagram accounts were disrupted and 25,446⁹⁸ accounts received custom messaging, in addition to the other actions described above.
- 6. In the 7 June report, 58 buyer accounts were identified of persons using their real names to buy SG-CSAM content on Instagram. One month after reporting these, the researchers claim that 28 buyer accounts were still active. In the 21 September report, the researchers identified 70 potential buyer accounts.**
- a) **Please explain whether and to what extent Meta acts against buyer accounts of SG-CSAM. Please explain the policy underlying those actions and outline potential actions and the average time to action on buyer accounts. .**
109. Meta prohibits solicitation of CSAM and it works aggressively to fight child exploitation off its platforms. Solicitation of CSAM is a violation of Meta's external policies – the Instagram Community Guidelines and the Facebook Community Standards – regardless of whether the interaction is between two adults or an adult and a minor or between two minors. These policies thus apply equally to users buying both CSAM or SG-CSAM content.
110. Profiles that have confirmed Child Sexual Solicitation keywords (e.g., [REDACTED]) are disabled and directly reported to NCMEC for CSAM solicitation.⁹⁹ Further, the actions described in the response to question 5(a) are deployed (i.e., one-strike disable, hard-linked propagation, device blocking and recidivism monitoring). These actions are reflected in the following CSAM-related SRA controls: CIDE-124 Disabling Accounts, CIDE-125 - Disabling Pages and Groups (if relevant), CIDE-123 Device Blocking, CIDE-121 Recidivism, CIDPA-003 Partnership with NCMEC, and CIDE-127 Cross-Platform Enforcement.
111. Regarding the average time to action regarding buyer accounts, this will depend on how the buyer account is identified, and the nature of the violation. Content identified by classifiers will normally be auto-actioned immediately. However, user reports will be enqueued to specialist reviewers.¹⁰⁰

⁹⁸ A total of 27,058 accounts were originally identified as a suitable cohort to receive custom messaging. Custom messaging was not deployed to the accounts that had been disabled in the time between the investigation and the deployment of custom messaging.

⁹⁹ For more information on Meta's co-operation with NCMEC see Question 1(b) above.

¹⁰⁰ For more information on the Specialist Review (i.e., human), please see the response to Question 5(a). For more information on classifiers, please see the response to Questions 1(b) and 8(a) above.

- b) More specifically, to what extent Meta has taken action against the 58 buyer accounts identified in June and the average time it took to take action on them and to what extent Meta has taken action against the 70 buyer accounts identified in September and the average time it took to take action on them. Please also indicate which measures were taken.
112. MPIL refers to its response to question 5(c) above for a description of its actions against the accounts identified by the SIO in April and August 2023 (which were the basis of the June and September reports, respectively). Specifically in relation to the suspected buyer accounts identified by SIO:
- a. Of the 57 accounts identified in April 2023, 55 were disabled (with account end dates between 2 May and 28 June 2023) and 2 were found not to be CSAM buyer accounts;
 - b. Of the 70 accounts identified in August 2023, 39 were disabled (with account end dates between 7 August¹⁰¹ and 6 November 2023) and 31 were found not to be CSAM buyer accounts.
- c) Please explain whether and to what extent Meta applies proactive and/or preventive filtering to detect buyer accounts on Instagram. If any filtering measures are applied, please explain their functioning, effectiveness and error rates.
113. Meta applies the same proactive detection tools to detect buyer accounts on Instagram as it does for sellers accounts on Instagram. For the explanation on their functioning, effectiveness and error rates, we refer to the response to question 5(b).
- d) Please provide information, including internal documentation, showcasing Meta's assessment of the increase in buyer accounts on Instagram between June and September.
114. Meta did not conduct any specific assessment of any increase in the number of buyer accounts on Instagram between June and September.
115. Rather, the number of active CSAM-related accounts identified in the two large SND investigations (described in response to Question 5(c) above) dropped from around 61,000, to around 35,000. Regarding the number of buyer accounts identified in this period following the reports of SIO, this dropped from 55 accounts to 39 accounts (see response to question 6(b) above).

¹⁰¹ Meta had taken action on certain accounts even before they were identified by SIO.

7. Meta marks in its Risk Assessment for Instagram that the category of risks marked as *Child Sexual Exploitation, Abuse and Nudity* and *Child* “potentially” includes self-generated CSAM, sexualisation of minors, and exploitative intimate imagery and sextortion of minors. The combination of risks identified in the category *Child Sexual Exploitation, Abuse and Nudity* and *Child* are marked as a Tier 2 risk.

a) Please substantiate why the risk of self-generated CSAM, sexualisation of minors, and exploitative intimate imagery and sextortion of minors is referred to as ‘potential’. Pursuant to Article 34(3), please provide all underlying documentation supporting the assessment of these risks.

116. Meta takes the issue of protection of minors very seriously. As is clear from its context, the use of the word “*potentially*” in the Instagram SRA Report should not be understood as casting doubt over the existence of this very real issue. The SRA Report uses the term “potential” in relation to risks across numerous Systemic Risk Areas, which simply reflects the requirements of Article 34(1) DSA that the risk assessment capture “*actual or foreseeable*” negative effects in relation to various issues.

117. For the Instagram and Facebook SRAs, the risks arising from SG-CSAM were addressed under the Problem Area of CSEAN. The risk assessment underlying the SRAs are recorded in the Workbook for CSEAN.¹⁰² This identified seven L4 risks associated with the specific risks identified in the question are: (i) Child Endangerment; (ii) Child Nudity; (iii) Child Sexual Abuse Materials (including SG-CSAM); (iv) Inappropriate Interactions with Children; (v) Non-Sexual Child Abuse; (vi) Sexualization of Children; and (vii) Solicitation of Child Sexual Abuse Materials. The risk assessment process is described in more detail in paragraphs 1-13 of the 15 December Response.

118. In response to the request for the underlying documentation supporting the risk assessment of Child Sexual Exploitation, Abuse & Nudity, MPIL refers to:

- a. The CSEAN Workbook;¹⁰³
- b. The Scoresheet for the CSEAN Problem Area, in the SRA Executive Summary;¹⁰⁴ and
- c. The SRA materials in Annexes 1-6 and 25-27 of the 15 December 2023 Response

b) Please provide the calculations for and the final calculated score of the *Inherent Risk Score*, as applied in Meta’s Risk Assessment for Instagram, for all the risks identified under (a).

119. The CSEAN Workbook¹⁰⁵ contains MPIL’s risk scoring and control design effectiveness scoring for the relevant mitigating controls.

120. To calculate the Inherent Risk Score, MPIL multiplied the risk’s Severity calculation by its Likelihood calculation. This is explained in more detail in page 7 ‘Inherent Risk Rubric’ in the SRA Methodology.¹⁰⁶

¹⁰² See Annex 10 of the 15 December Response.

¹⁰³ See Annex 10 of the 15 December Response.

¹⁰⁴ See Annex 29 to the 15 December Response.

¹⁰⁵ See Annex 10 of the 15 December Response.

¹⁰⁶ See Annex 4 of the 15 December Response.

121. The Severity score was calculated according to the Severity Rubric, which considers the severity of the impact on users and society, based on ‘Scale,’ ‘Nature of Impact,’ and ‘Impacted Demographic.’ Where a particular risk was deemed to have a potential or particular impact on minors, it received the highest possible tier rating (Tier 5) in the ‘Impacted Demographic’ factor of the Severity Rubric. Where a risk was deemed to have a potential/particular impact on the rights of the child, it also received the highest possible tier rating (Tier 5) in the ‘Nature of Impact’ factor.
122. The Likelihood score considered the likelihood of the risk materialising, based on ‘Scope’ and ‘Volume’ (explained in page 7 ‘Inherent Risk Rubric’ in the SRA methodology).
123. The following table sets out the Inherent Risk Score for the risks identified under (a) for Instagram, and for completeness the Residual Risk Scoring (after the deduction of the product of the Inherent Risk Score and the Control Suite Effectiveness), which resulted in the determination that the residual risk was low with a risk rating at Tier 2. This information is extracted from the CSEAN Workbook, in the tab ‘Risk Register & Scoring.’

L4 Risk	Severity [1-5]	Likelihood [1-5]	Inherent Risk Score [Severity x Likelihood]	Control Suite Effectiveness ¹⁰⁷	Residual Risk Score ¹⁰⁸
Child Sexual Abuse Materials	5	3	15	63%	5.6
Sexualization of Children	5	3	15	63%	5.6
Inappropriate Interactions with Children	5	3	15	68%	4.8
Solicitation of Child Sexual Abuse Materials	5	3	15	63%	5.6

c) For all the risks identified under (a), please provide the results of the calculations for the scores on *Severity*, including *Scale*, *Rights* and *Impacted Demographic*, and the score for *Likelihood*, including *Scope* and *Volume*.

124. The Severity score is the average of the scores attributed for ‘Scale,’ ‘Rights,’ and ‘Impacted Demographic’. These are scored from 1 to 5 (5 being the most severe), according to the ‘Inherent Risk Rubric’ in the SRA Methodology. The following table sets out the scores relating to Severity for each of the risks identified in 7(a). This information is extracted from the CSEAN Workbook,¹⁰⁹ in the tab ‘Risk Register & Scoring.’

L4 Risk	Severity Rating (Scale)	Severity Rating (Nature of Impact)	Severity Rating (Impacted Demographic)	Total Severity [average]
---------	----------------------------	---------------------------------------	---	-----------------------------

¹⁰⁷ See page 8, Control Effectiveness Rubric, SRA Methodology, Annex 4 of the 15 December Response.

¹⁰⁸ Tier 1 (0-4.9) ‘Extremely Low’; Tier 2 (5-9.9) ‘Low’; Tier 3 (10-14.9) ‘Moderate’; Tier 4 (15-19.9) ‘Elevated’; Tier 5 (20-25) ‘Extremely Elevated’; See page 11, Residual Risk: Tiering, SRA Methodology, Annex 4 of the 15 December Response.

¹⁰⁹ See Annex 10 of the 15 December Response.

Child Sexual Abuse Materials	5	5	5	5
Sexualization of Children	5	5	5	5
Inappropriate Interactions with Children	5	5	5	5
Solicitation of Child Sexual Abuse Materials	5	5	5	5

125. The Likelihood score is the average of the scores attributed for ‘Scope’ and ‘Volume’. These are scored from 1 to 5 (5 being the most severe), according to the ‘Inherent Risk Rubric’ in the SRA Methodology. The following table sets out the scores relating to Likelihood for each of the risks identified in (a). This information is extracted from the CSEAN Workbook,¹¹⁰ in the tab ‘Risk Register & Scoring’.

L4 Risk	Likelihood Rating (Scope)	Likelihood Rating (Volume)	Total Likelihood [average]
Child Sexual Abuse Materials	5	1	3
Sexualization of Children	5	1	3
Inappropriate Interactions with Children	5	1	3
Solicitation of Child Sexual Abuse Materials	5	1	3

d) According to the Risk Assessment, each of these scores is calculated using several inputs “such as risk assessment workshops, documentation, data, and external research.” Please provide all supporting documentation used for the calculation of the scores outlined in your answers under (b) and (c).

126. The supporting documentation used for the calculation of the scores outlined in the answers under (b) and (c) is the CSEAN Workbook.¹¹¹ The Workbook is the product of the various workshops carried out with subject-matter experts, familiar with the relevant data and research (as detailed in paragraphs 9 -10 of the 15 December Response).

¹¹⁰ See Annex 10 of the 15 December Response.

¹¹¹ See Annex 10 of the 15 December Response.

8. Meta marks in its Risk Assessment that there is a risk that on Instagram “(..) the behaviour of threat actors is currently evolving, potentially including through intentional manipulation by threat actors to persistently adapt to evade detection, including using implicit signals like new keywords and hashtags (e.g., “chicken soup”), piggyback benign links, moving conversations off the service to take advantage of minors, and returning to the service through new accounts despite being blocked.”

a) Please share all specific mitigation measures Meta has put in place for Instagram to prevent the intentional manipulation identified as a risk above beyond the steps taken by the task force on ‘the prevention of bad actors reaching out’ as outlined in Meta’s 29 August response. This should include all evidence collected on the effectiveness of the measure(s) to mitigate the risk and all underlying evidence concerning internal testing of these measures.

On the specific mitigation measures in place

127. Beyond the work of the CSTF described in the letter to the EC of 29 August 2023, Meta deploys various controls to mitigate the risks bad actors evading detection through implicit signals such as new keywords and hashtags,¹¹² benign links, or moving conversations off the service and returning to the service through new accounts despite being blocked. These controls reflect a two-pronged approach. On the one hand, these controls aim to inform and empower teens to protect themselves from engaging with potential bad actors in the first place, thereby reducing their exposure to bad actors both on and off-platform. On the other hand, Meta deploys extensive measures to mitigate the risk of intentional manipulation by bad actors, including attempts to evade detection (e.g. through the use of new keywords, or returning to Instagram having been blocked).
128. Controls that focus on empowering teens not to engage with bad actors are:
- a. **Unwanted interactions.**¹¹³ Meta seeks to limit teen users from receiving unwanted contact from adult users through the MaCSA Framework.¹¹⁴
 - b. Meta has also developed tools that teens can report to if something makes them feel uncomfortable while using Instagram and notifications that encourage them to use these tools. For example, when a teen user blocks an account, Meta prompts them to report the account. Meta also makes it more difficult for adults to find and follow teens on Instagram.¹¹⁵ Meta has made technical changes to try to verify certain user reports are not unexpectedly or automatically closed, which includes improvements to our systems to prioritise these reports for content reviewers.

¹¹² See question 2 for more details.

¹¹³ Control ID CIDE-195.

¹¹⁴ For more details on the MaCSA Framework, see response to question 1(b) (paragraph 28) above and the response to question 9 below.

¹¹⁵ For more details, see response to question 1(b) above.

- c. **Encouraging the use of privacy tools.**¹¹⁶ In 2022, Meta launched a new way to encourage teens to update their safety and privacy settings on Instagram, by prompting teens to review their settings including: controlling who can reshare their content, who can message and contact them, what content they can see and how they can manage their time spent on Instagram. For Instagram users, Meta also offers a guide specifically aimed at teens, dedicated to staying safe online and creating a positive experience. Further, as the Commission is aware, on 9 January 2023 Meta announced that it will be implementing new measures to encourage teens to use the privacy tools. To help make sure teens are regularly checking their safety and privacy settings on Instagram, and are aware of the more private settings available, Meta will send new notifications encouraging them to update their settings to a more private experience with a single tap. If teens choose to “Turn on recommended settings”, Meta will automatically change their settings to restrict who can repost their content, tag or mention them, or include their content in Reels Remixes. Meta will also ensure only their followers can message them and help hide offensive comments.¹¹⁷

129. Controls that focus on the risk of intentional manipulation by bad actors are:

- a. **Instagram Search Interventions.**¹¹⁸ MPIL refers to its response to Question 2 above in relation to the Unified Child Safety Keyword List. Currently, tens of thousands of terms and hashtags are blocked on Instagram and return no results. The block can be temporary if hashtags and keywords are potentially risky. Alternatively, Meta offers generic search friction and does not allow users to view search results. This is referred to as an “empty inform” module or treatment. Meta also blocks certain terms from appearing in Typeahead suggestions.
- b. **Key Word Interstitial.**¹¹⁹ Meta deploys a pop-up that is shown to users who search for terms that have been identified as potentially CSAM-related and redirects them to educational resources and prevents users from clicking through to view search results. The interstitial is triggered by terms that are borderline, in that they are risky and may suggest a violation but also may have legitimate, non-violating uses.
- c. **TMS.**¹²⁰ Meta deploys a tool that allows for detection of keywords and phrases across various sources of text content on Instagram. Meta has used TMS to proactively detect usage of inherently violating child safety keywords and phrases within Instagram profile elements (e.g., usernames, names, bios, or media captions), in order to increase enforcement against profiles containing child safety-violating text. Profiles found through the TMS may be disabled; reported to the NCMEC; or enqueued for human review, depending on the keywords or terms detected. The list of keywords and terms used is regularly updated, based on input from internal Meta sources and external third-party experts. Besides identifying and disabling personal accounts,¹²¹ Meta also ensures that TMS identifies overtly violating subject elements in Groups, Pages, Profiles +, and Additional Personal Profiles.¹²²

¹¹⁶ Control ID CIDE-082.

¹¹⁷ See announcement ‘New Protections to Give Teens More Age-Appropriate Experiences on Our Apps’ of 9 January 2023, available at <https://about.fb.com/news/2024/01/teen-protections-age-appropriate-experiences-on-our-apps/>

¹¹⁸ Control ID CIDE-089.

¹¹⁹ Control ID CIDE-194.

¹²⁰ Control ID CIDE-056. For a description of the results of the deployment of this mitigation measure, please refer to paragraph 52.

¹²¹ As explained in the Response to EC of 29 August 2023.

¹²² See here for more information: <https://about.fb.com/news/2023/09/you-can-now-have-multiple-personal-profiles-on-facebook/>

- d. **Classifiers.**¹²³ Meta deploys child safety classifiers and systems, such as UnIS¹²⁴ and Google's Child API,¹²⁵ to enforce against interactions between teens and Suspicious Adults. There are a variety of enforcement actions, including soft actions (e.g., blocking People You May Know (PYMK),¹²⁶ etc.). Furthermore, Meta deploys a child safety keyword text classifier, trained on keywords from Thorn (a nonprofit focused on fighting child sex abuse) and data labelled by Product Data Ops, including Instagram hashtags, and search queries. This classifier produces a score indicating sharing of CSAM or child sex material, which is used to help trigger search help modules and search suggestion filtering and as an internal signal for understanding the intent of child safety bad actors. Classifiers are an essential tool to combat the risk of intentional manipulation posed by bad actors as they recognise patterns and enable predictions of potentially violating content being shared or searched. Classifiers are updated on an ongoing basis based on feedback from reviewers working child safety jobs. The frequency with which classifiers are calibrated depends on the type of classifier. Further information on the functioning of the Classifiers used within the context of bad actors is provided in response to Question 9, with respect to the MaCSA Framework.
- e. **MaCSA Framework.**¹²⁷ As described above, Meta leverages the MaCSA Framework in its recommendation and search filtering protocols to prevent Suspicious Adults from connecting with each other or with teens. In particular, recent efforts to tackle MaCSA-to-MaCSA contacts resulted in a reduction of Instagram 'Risky Actor Connectedness'¹²⁸ as evidenced in Annex 8. This is an internal Workplace post of 21 December 2023 on the completion of the MaCSA-to-MaCSA interventions, that also sets out the backtest results. In addition, an intervention was launched in October 2023, whereby content posted by Suspicious Adults is not recommendable to any Instagram user (teen or adult), following testing evidenced in Annex 9. This is an internal Workplace post of 22 September 2023 entitled '[Experiment Results] Filter MaCSA Strict for Gen-pop on AYM/Reels/explore/Feed Rec' that describes the tests in relation to filtering Suspicious Adults' accounts and contents from recommendation surfaces.
- f. **URL Detection.**¹²⁹ Meta deploys three main tools to detect URLs leading to off-platform websites with violating child safety content:
 1. A proactive URL Reputation system that runs a URL trunk classifier, capturing a limited preview of a website's landing page and assessing whether there is violating or suspicious content there. The system classifies the content on the URL landing page to assess whether the content is likely to violate Meta's policies. This mechanism is particularly effective in identifying piggyback benign links as, regardless of the text of the hyperlink, Meta can capture the substance of the content linked.
 2. The second proactive URL detection system is an Account Clusters Deserve Classification classifier which first identifies accounts that are likely to be malicious and then detects suspicious URLs shared by those accounts.

¹²³ Control ID CIDE-048.

¹²⁴ Control ID CIDE-195.

¹²⁵ Control ID CIDPA-287.

¹²⁶ Control ID CIDPA-014.

¹²⁷ Control ID CIDPA-288. For more details on the MaCSA Framework, see response to question 1(b) (paragraph 28) above and the response to question 9 below.

¹²⁸ For more details, see response to question 1(c) (paragraph 58) above.

¹²⁹ Control ID CIDE-052.

3. Finally, Meta has in place a reactive URL detection system, which first identifies host domains that have a bad reputation and then develops “weak signals” around the domain to identify potentially malicious URLs.

On the evidence of the effectiveness and internal testing of the mitigation measures

130. In addition to the testing described and evidenced above, as regards all those controls that were taken into account in the SRA process, their design effectiveness in relation to each of the identified risks was assessed in the workbook tab ‘Detailed Control Scoring’ of the CSEAN Workbook, and described in more detail in response to question 7 above. Specifically in relation to the L4 risk of “Inappropriate Interactions with Children”, the Inherent Risk score was 15.0 ‘Elevated’ but after applying the various mitigating controls (listed in the tab “Detailed Control Scoring”) the Residual Risk was 4.8 ‘Extremely Low’ (see the tab “Risk Register & Scoring”).¹³⁰ This is evidenced in the CSEAN Workbook.¹³¹
131. As regards the effectiveness of the various mitigating controls, MPIL notes that, as part of the SRA process, it is in the process of developing an additional parameter of ‘Operating Effectiveness’, on top of the Design Effectiveness and Mitigation Effectiveness described above, with a view to applying this in the SRA to be carried out in 2024. As part of this work, MPIL has prioritised the effectiveness testing of a number of controls that are relevant to the Problem Areas of CSEAN (as well as Human Exploitation). This exercise is ongoing, so the reports of this testing are not yet available.
 - b) Please provide evidence collected on the effectiveness of policies, including benchmarks and KPIs, and specific measures in place mitigating the following identified risks:**
 - i) Adaptations by threat actors to evade detection, including through the use of new keywords and hashtags.**
 132. The mitigation controls deployed to mitigate the risk of bad actors evading detection, including through the use of new keywords and hashtag, are detailed in response to question 8(a) above. Specific reference is made to the Unified Child Safety Keyword List and the MaCSA Framework, described above.
 133. As regards the effectiveness of these measures, reference is made to the response to question 8(a) above.
 - ii) Threat actors returning to Instagram through new accounts despite being blocked.**
 134. MPIL leverages specific controls developed and deployed by Meta to prevent bad actors from returning to Instagram through new accounts despite having been blocked. These controls are described below:

¹³⁰ Tier 1 (0-4.9) ‘Extremely Low’; Tier 2 (5-9.9) ‘Low’; Tier 3 (10-14.9) ‘Moderate’; Tier 4 (15-19.9) ‘Elevated’; Tier 5 (20-25) ‘Extremely Elevated’; See page 11, Residual Risk: Tiering, SRA Methodology, Annex 4 of the 15 December Response.

¹³¹ See Annex 10 of the 15 December Response.

- a. **Multi-block function.**¹³² Meta employs a “soft block” function that is automatically and silently initiated when a user blocks a single account. Because a single person can create multiple accounts, Multi-block uses a variety of signals to identify and block other accounts that are highly likely to be associated with the person who was blocked. [REDACTED] Multi-block silently blocks the associated accounts and aims to block accounts that will ever be created on that device in the future. Meta does not tell users which associated/secondary accounts are soft-blocked. This means that the blocked user will not be able to see the blocker’s profile or content from their secondary accounts; however the blocker will still be able to find and engage with the secondary accounts. The implementation of this mitigation measure makes it more difficult for accounts that have been blocked to return on the platform with the aim of endangering teen users.
- b. **Device blocking.**¹³³ For more details on Device Blocking, see the response to question 1(c) above (paragraph 50(b)).
- c. **Recidivism.**¹³⁴ Meta employs controls based on indicators that help signal whether a new entity is meant to side-step Meta’s enforcement on a previous entity. [REDACTED] Meta is not only trying to remove/disable nearly identical replicas of the accounts/entities removed, but is also taking steps to remove all accounts/entities that are functionally equivalent to the removed entity on the basis of the review of the content and of the purpose of these new entities through a proactive classifier.
- d. **Account propagation.**¹³⁵ When Meta disables an account for a violation connected to CSAM activity, it also disables explicitly linked accounts (where the user has linked their Facebook and Instagram profiles) and high-confidence linked accounts (where there is high likelihood that the same user is using multiple accounts).
- c) **Please include specific information on any policies or specific measures in place mitigating the risks included in your answer to Question 7(a), including any evidence collected on the effectiveness of these policies and/or measures.**
135. Question 7(a) refers to the risk of SG-CSAM, sexualisation of minors, and exploitative intimate imagery and sextortion of minors. These risks, and the mitigating controls, and MPIL’s assessment of their design effectiveness, are addressed in the Workbook for CSEAN described in its response to Question 7(a).

¹³² Control ID CIDE-089.
¹³³ Control ID CIDE-123.
¹³⁴ Control ID CIDE-121.
¹³⁵ Control ID CIDE-127.

Private and Confidential – Contains Commercially Sensitive Information

136. With regard to its policies, as explained in response to Question 1(a), Meta prohibits any content or activity that sexually exploits or endangers children. Meta's CSEAN policy available on its Transparency Center specifically prohibits sexualisation of minors, exploitative intimate imagery, and sextortion, and generally prohibits all CSAM in a manner that applies to specific harms including SG-CSAM.¹³⁶ Meta also works with external experts, including the Facebook Safety Advisory Board, to discuss and improve its policies and enforcement around online safety issues involving children.
137. With regard to specific mitigation measures in place, Meta has enacted various controls in all of these areas.

SG-CSAM

138. Meta notes that it has broad strategies in place to prevent the dissemination of all CSAM, which also help prevent the dissemination of SG-CSAM. Meta devotes substantial financial and human resources to detecting and enforcing against all CSAM violations, as explained in response to Question 1 above, and in June 2023 mobilised a Child Safety Task Force to, *inter alia*, specifically address the dissemination of CSAM during a period where such an additional measure became necessary. These efforts are detailed in response to questions 1 and 2 above. With regard to the effectiveness of Meta's efforts to prevent the dissemination of SG-CSAM, Meta concluded that the CSTF largely fulfilled its exit criteria.¹³⁷

Sexualisation of Minors

139. Meta also identified sexualisation of minors as a key focus area and has taken steps to bolster its capacity to detect and enforce against child sexualisation content and to better prioritise child sexualisation reports in its human review queues. Meta took action to increase enforcement against profiles containing text suggesting child sexualisation by proactively detecting usage of inherently violating child safety keywords and phrases within Instagram profile elements (e.g., usernames, names, bios, or media captions). Profiles that have confirmed CSx keywords are disabled and reported to NCMEC for CSAM solicitation,¹³⁸ and profiles with confirmed Child Sexualisation keywords are disabled. As part of the CSTF's conclusion, the Unified Child Safety Keyword List was updated using multiple inputs.¹³⁹ Within the first 48 hours of launching these improved enforcements, 5,500 profiles were enqueued in an Instagram profile review queue, and around 2,500 (18.7%) were disabled for violating Meta's Child Safety policy. Another 1,300 Instagram profiles were directly disabled for a high severity violation.
140. Moreover, recognizing the nuance inherent in determining whether sexualisation of minors is taking place, Meta is continuing to develop its strategies for addressing explicit and implicit sexualisation. For example, since August 2023, Meta has expanded its Operational Guidelines with clarifications and examples, as well as text and visual signals for implicit sexualization.

¹³⁶ See Meta's CSEAN Policy on its Transparency Center, at <https://transparency.fb.com/en-gb/policies/community-standards/child-sexual-exploitation-abuse-nudity/>.

¹³⁷ See response to Question 1(a) above.

¹³⁸ For additional details on Meta's partnership with NCMEC, and Meta's provision of CSAM hashes to NCMEC, please see the response to question 1(b).

¹³⁹ For more information on the Unified Child Safety Keyword List, please see the response to Question 2.

Exploitative Intimate Imagery and Sextortion

141. Specifically as concerns the designated platforms,¹⁴⁰ Meta has made extensive investments to further understand and counteract user behaviour and violations relating to exploitative intimate imagery ('EII') and sextortion. In 2023, Meta made significant strides responding quickly and accurately to user reports of sextortion and driving targeted enforcement against risky accounts with known violating behaviours for sextortion.
142. Meta investigative teams have also determined that financial sextortion actors often conduct a wide array of scams using a large number of fake or compromised accounts. As a result, Meta took additional steps to identify and enforce on device-linked accounts. It also carries out investigations into financial sextortion accounts, potentially resulting in SNDs.
143. Further, as described above in response to Question 1(a) above, Meta participates in Take it Down, a platform that helps prevent minors' intimate images from being disseminated online. Other support to victims is provided through StopNCII (a platform that empowers users to thwart the non-consensual sharing of their intimate images on various platforms), by directing potential victims to counselling support, through providing information on Meta's Safety Center, and by working with Thorn to expand its education and prevention campaign.

9. Meta mentions that "Meta has built and launched a malicious actor framework which is a flexible, adaptable tool which uses technologies like machine learning to identify and address potentially malicious activity, using behavioural signals to proactively identify and take actions against bad actor accounts." and furthermore that "Because Meta has high confidence that these accounts are likely to violate Meta's policies based on the likelihood of maliciousness, Meta has lowered the thresholds for disabling such accounts, which has resulted in Meta disabling many more accounts than before."

a) Please provide quantitative evidence to substantiate the claim that the malicious actor framework has resulted in disabling more accounts than before.

144. The claim that the MaCSA Framework resulted in disabling more accounts than before is supported by reference to the following statistics.
145. MaCSA signals-based enforcements were launched for Facebook and Instagram between the second and third quarters of 2023. In the 6-month period from June to November 2023, Meta disabled a total of over 1,400 Facebook user accounts in the EU, for violating policies against CSAM. This represented an increase of 169% as compared to the 500+ Facebook user accounts disabled between December 2022 to May 2023 in the EU. Of the over 1,400 user accounts disabled from June to November 2023 on Facebook (EU), more than 46% can be attributed directly to MaCSA signals.
146. In the 6-month period from June to November 2023, MPIL disabled a total of over 25,100 Instagram user accounts in the EU for violating policies against CSAM. This represented an increase of 1,102% compared to the over 2,000 Instagram user accounts disabled between December 2022 to May 2023 in the EU. Out of the over 25,100 user accounts disabled from June to November 2023 on Instagram (EU), more than 49% can be attributed directly to MaCSA signals.
147. The above numbers reflect all accounts that were disabled when the relevant threshold for MaCSA signals was met, as well as linked accounts that were disabled via propagation. They do not include other enforcements that may be undertaken through the MaCSA Framework, such as limitations on the provision of services.

¹⁴⁰ As recognised in recital 7 of the Commission Decision of 25 April 2023 designating Instagram, private messaging services are not covered by the VLOP designation.

b) To what extent does Meta register malicious actors internally? In case there is a list, what safeguards are in place to allow persons included as a malicious actor to challenge that decision and the number of people included on the list.

148. Meta does not maintain a register or ‘list’ of confirmed bad actors *per se*. Rather, as explained in response to question 1(a) (paragraph 28), it uses the MaCSA Framework to determine Suspicious Adults and take appropriate action. The consequences of a Suspicious Adult designation will depend on the level of risk resulting from the determination of Suspicious Behaviours - *i.e.*, Meta determines which responsive measure(s) are appropriate to address such risk - and it will depend on the Signals detected and the gravity of the action (*i.e.*, risk for teens resulting from the action). Measures that Meta may take include account limitations or in certain cases, the Suspicious Behaviours may be deemed so severe as to warrant the account of the Suspicious Adult being disabled. These measures are automatically implemented on the Suspicious Adult’s account.
149. Users who are classified as Suspicious Adults are not informed that they have been qualified as such or that any restrictions have been applied to their accounts (unless their account has been disabled, as explained in response to question 9(c) below). However, all users are informed about the Instagram Terms of Use or Facebook Terms of Service (which they are contractually required to agree to) and Instagram Community Guidelines and Facebook Community Standards (which link through to the [[HYPERLINK "https://transparency.fb.com/policies/community-standards/child-sexual-exploitation-abuse-nudity/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fchild_nudity_sexual_exploitation" \h](https://transparency.fb.com/policies/community-standards/child-sexual-exploitation-abuse-nudity/?source=https%3A%2F%2Fwww.facebook.com%2Fcommunitystandards%2Fchild_nudity_sexual_exploitation)] and explains what content, behaviour and actions are unacceptable) when they join the platform.
150. The reason for not informing all users who are classified as Suspicious Adults (other than those whose accounts are disabled, as explained in response to question 9(c) below) and not providing the specifics of how Meta detects Suspicious Behaviour and designates an Adult as a Suspicious Adult is that to do so would risk Suspicious Adults being able to circumvent the account limitations adopted by Meta to protect Teens. For example, Suspicious Adults might modify their behaviour online trying to avoid being qualified again as Suspicious Adults or might create a new account in an attempt to elude the determination process. This would result in the creation of new serious threats for Teens.
151. That said, Suspicious Adults are not categorised as “suspicious” permanently. Their status depends on their ongoing behavioural patterns over time. Meta has implemented a process to regularly verify the determination of Suspicious Adults: the determination of a Suspicious Adult is reviewed daily by Meta to identify whether that user might still be a threat for Teens based on: (i) the users’ current behaviour and activity and (ii) the current set of Signals. Following the daily assessment, the adult’s designation as a Suspicious Adult is updated to reflect any changes since the last daily update, including removing the designation as appropriate.
152. Specifically, each signal used to detect whether a user is manifesting Suspicious Behaviour is assigned a specific time period called ‘lookback period’ for which it is taken into account (*i.e.*, processed to automatically detect whether that adult’s actions determine Suspicious Behaviour). The lookback period for each signal will vary depending on the gravity of the action and recurrency of such actions during a period of time. This period can range from one day to over a year. Once the lookback period has expired, that particular signal will be deleted and no longer taken into account for the assessment of the suspicious behaviour of the user unless the same signal is collected again which would re-trigger the lookback period.

c) In situations where no violation has yet taken place, but Meta does disable access to a user's account, is the user notified of the reasoning (i.e., qualification of the user as a 'malicious actor' and potential likelihood of a violation) underpinning the disabling of access to an account? Are users informed of any subsequent consequences when they would want to return to the platform in the future?

153. Meta notifies Suspicious Adults where their Instagram accounts are disabled. The Suspicious Adult of the disabled account is: (i) notified that their Instagram account has been disabled; (ii) provided with brief information about why; and (iii) provided with the aforementioned rationale and a link to view the detailed policies that have been violated the ("Notice Process"). In order to protect the efficacy of Meta's enforcement processes, the Suspicious Adult is not provided with explanations of the Signals and how they have been processed to determine potential Suspicious Behaviour, as this information could be used by, and shared with, bad actors to circumvent detection by Meta, thus undermining the efforts made by Meta to improve the protection of Teens and creating a serious threat for Teens' safety.
154. Further, and as demonstrated to the Commission's DSA team on 18 July 2023, Meta has an appeals process for disabled accounts to provide an effective recourse to users affected by a disabled accounts decision. As part of the notice process described above, the Suspicious Adult will be given an immediate, user-friendly option to appeal Meta's decision. If the appeal is successful, Meta's decision will be reversed, and the disabled account will be reactivated.

B. In relation to Instagram's recommender system

10. Research commissioned by Meta suggests that 'video contain[ing] borderline nudity [...] may be considered problematic especially after repeated exposure' and that 'echo chambers related to specific item features can create a negative user experience. Explicitly mitigating such effects can have an overall positive impact.'¹⁴¹ Recent research also points at amplification of certain types of content by Instagram's recommender system.¹⁴²
- a) Please explain whether Meta has considered any measures related to its recommender system to avoid that child sexualisation content or other type of 'borderline, problematic content' is pushed to users interested in such type of content, on the basis of their previous preferences. If so, please include all information regarding the inputs, signals and predictive models used within their recommender systems to prevent the distribution of harmful content on their services, including details on how their effectiveness is measured.
155. MPIL notes that Instagram does not have one 'recommender system' that impacts what users do or do not see on the platform. MPIL respectfully refers the Commission to paragraphs 80-102 of the 15 December Response for a detailed explanation of Meta's approach to recommendations, and how, amongst other things, Instagram uses integrity-related predictions and rules to demote or reduce distribution of certain potentially problematic or low-quality content, such as content that may be sensitive for some users, including borderline content.
156. In relation to Articles 34 and 35 DSA, recommender systems were considered as an 'influencing factor' which MPIL took into consideration when assessing the systemic risks identified in its SRA process (see the Instagram SRA Results Report, p. 22 section 5.2.1 and 6.1.2¹⁴³).

¹⁴¹ D. Kalimeris, S. Kalyanaraman, S. Bhagat, U. Weinsberg, 'Preference Amplification in Recommender Systems', [HYPERLINK "https://dl.acm.org/doi/pdf/10.1145/3447548.3467298" \h].

¹⁴² J. Horwitz & K Blunt, 'Instagram's Algorithm Delivers Toxic Video Mix to Adults Who Follow Children', The Wall Street Journal, [HYPERLINK "https://www.wsj.com/tech/meta-instagram-video-algorithm-children-adult-sexual-content-72874155?mod=Searchresults_pos4&page=1" \h].

¹⁴³ See Annex 1 of the 15 December Response.

157. This analysis fed into the control effectiveness measurement efforts as part of the “Measure” phase of the SRA process to help determine the residual risk for the various risks associated with each of the 19 Problem Areas. The procedures through which MPIL evaluated the effectiveness of these measures are set out in paragraphs 11-12 of the 15 December Response.
- b) Please explain whether the Meta experiment referenced in section 4.3 of the report on ‘Preference Amplification in Recommender Systems’ relates to both Facebook and Instagram. If not, could you please provide information about whether similar experiments took place for the recommender system of Instagram and if so, share all relevant documents and data.**
158. The paper on the ‘Preference Amplification in Recommender Systems’ was a joint research project by representatives from Meta (then, Facebook) and Harvard University. The specific experiment referenced in Section 4.3 of the paper was conducted in relation to Facebook only. This research was published in August 2021, and therefore materially pre-dates, and was not prepared pursuant to (or in contemplation of), MPIL’s compliance with Articles 34 and 35 of the DSA. No similar experiments were conducted in relation to Instagram as part of MPIL’s evaluation to date of systemic risks and related mitigations under Articles 34 and 35 DSA.
- c) Please provide us with any research, analysis, internal testing such as A/B tests, or studies Meta undertook or is aware of related to potential amplification of content which may be harmful to underaged users’ mental well-being by Instagram’s recommendation algorithm. Please explain whether and to what extent Meta has made changes to the recommendation algorithm and/or any related training, testing and improvement processes in light of this information and the reasons for the changes (or lack thereof).**
159. At the outset, MPIL refers to its response to questions 7 and 8 in the 15 December Response for a detailed description of how it evaluated potential risks to the physical and mental well-being of its users as part of the SRA process. As explained there, the potential for harm in these areas was considered through the lens of certain specific risks (namely, ‘Disordered Eating’, ‘Mental Health’, ‘Personal Health and Appearance Ads’ and ‘Suicide & Self-Injury’) under the Suicide and Self Injury Problem Area. Paragraph 78(a) - (v) of the 15 December Response details controls MPIL identified in the SRA to mitigate these risks on Instagram.
160. In addition, paragraphs 80-102 of the 15 December Response provide a detailed explanation of how Instagram’s recommender systems work, and how they have been designed to intervene at both the content and account (or ‘actor’) level to reduce the risk of users encountering policy violating or other sensitive or borderline content, with even stricter measures applying to teen users. This includes, for example, Meta’s Account Quality Value Model (“AQVM”), which marks an account as non-recommendable based on a statistical probability (typically of over 75%) on whether the actor’s feed or profile violates the Recommendation Guidelines. As a result, all of their content is filtered out from recommendation surfaces (e.g., Reels, Explore) and their account from account recommendation surfaces (Search, Accounts You Might Follow). For teen users, an even stricter version of AQVM applies (employing lower thresholds for classifying an account as non-recommendable), which prevents an even larger population of potentially sensitive accounts from being included in their recommendations.
161. Beyond MPIL’s assessment of systemic risks and related mitigations under Articles 34 and 35 DSA, MPIL notes that as part of its day-to-day product development and design efforts, it also conducts a wide variety of internal testing, research and analysis across a large number of internal teams for varying business purposes, not all of which, in and of itself, will be directly relevant to the matters required to be considered under Articles 34 and 35 DSA, and therefore will not have been directly assessed as part of MPIL’s systemic risk assessment and mitigation process.

162. Accordingly, and without prejudice to MPIL's position as to the relevance of this category of information for the purposes of evaluating its compliance with Article 34 and 35 DSA, MPIL has, in the limited time available to respond to this RFI, identified the following materials documenting internal testing, research and analysis that were finalised after 29 August 2023 and that considered the potential effects of the amplification of content on teen mental well-being:
- a. 'SSI/ED External Advisor Feedback Synthesis' of September 2023, giving the synthesised results from feedback sessions with external advisors on Suicide and Self-Injury and Eating Disorders. This relates to the controls Partnerships and Consultation (CIDE-006) and Partnering with Experts (CIDPA-272), Annex 10; and
 - b. 'Appearance Comparison Strategy Refresh' of October 2023, which refers to analysis related to appearance comparison, Annex 11.
163. As these materials relate to research that is recent, MPIL notes that the research remains under review and therefore is not necessarily a basis upon which to inform specific changes. To the extent that this research may in the future inform MPIL's assessment of systemic risks under Articles 34 DSA or its ongoing review and implementation of mitigations under Article 35 DSA, then it will be reflected in MPIL's SRA documents in accordance with those DSA requirements.
164. Meta is also making a number of changes, as recently announced on 9 January 2024, to further enhance the experience of minors on its services¹⁴⁴:
- a. **New Content Policies for Teens:** Meta already aims not to recommend age-inappropriate content to teens in recommendation surfaces like Reels and Explore. Meta will no longer show such content to teens in Feed and Stories, even if shared by someone they follow.
 - b. **Updates to Instagram's and Facebook's Content Recommendation Settings for Teens:** Meta is automatically placing teens into the most restrictive content control setting on Instagram and Facebook. Meta already applies this setting for new teens when they join Instagram and Facebook, and is now expanding it to teens who are already using these apps. Meta's content recommendation controls – known as "Sensitive Content Control" on Instagram and "Reduce" on Facebook – make it more difficult for people to come across potentially sensitive content or accounts in places like Search and Explore.
 - c. **Hiding More Results in Instagram Search Related to Suicide, Self-Harm and Eating Disorders:** While Meta allows people to share content discussing their own struggles with suicide, self-harm and eating disorders, Meta's policy is not to recommend this content and Meta has been focused on ways to make it harder to find. Now, when people search for terms related to suicide, self-harm and eating disorders, Meta will start hiding these related results and will direct them to expert resources for help. Meta already hide results for suicide and self harm search terms that inherently break our rules and Meta is extending this protection to include more terms.
 - d. **Prompting Teens to Easily Update Their Privacy Settings:** This is described in more detail in response to questions 1(b) and 8(a).

¹⁴⁴ Meta press release of 9 January 2024 'New Protections to Give Teens More Age-Appropriate Experiences on Our Apps' [[HYPERLINK "https://about.fb.com/news/2024/01/teen-protections-age-appropriate-experiences-on-our-apps/"](https://about.fb.com/news/2024/01/teen-protections-age-appropriate-experiences-on-our-apps/) \h]

165. MPIL notes the Commission's questions of 10 January in relation to this announcement and will respond to them separately.

d) Please set out which additional risk mitigation measures Meta put in place following the information outlined in question 10(c), including whether Meta considered to change its recommendation algorithms for users under the age of 18. Please include in your answer the reasons for choosing the specific mitigations measures, supported by internal or external evidence and underlying documents. Please include the necessary (references to) internal studies and assessments, as well as exchanges between staff on this topic.

166. In relation to the risk mitigation measures put in place in relation to recommendation algorithms and improvement, as well as related studies and assessments, MPIL refers to the response to question 10(c) above.

C. In relation to the age rating of the Instagram service

11. According to Instagram's Terms of Use a user must be at least 13 years old to set up an account.¹⁴⁵ However, it appears that the Instagram app is rated as "age 12+" on some app stores, indicating to potential users that also minors aged 12 are eligible to use the app.

a) Please provide us with all information, documenting the exchanges you had with providers of app stores concerning the minimum age of users about who should be able to download the Instagram app from the respective app store?

167. At the outset, MPIL notes that Articles 34 and 35 DSA are limited to the systemic risks stemming from the design or functioning of the services and related systems of the VLOP providers.¹⁴⁶ As such, MPIL considers that its obligations under Article 34 and 35 DSA do not extend to the ratings applied by third party app stores, over which it has no control and no responsibility. These ratings are not part of MPIL's risk assessment or risk mitigation under Articles 34 and 35 DSA.

168. In any event, and without prejudice to the relevance of this question to MPIL's compliance with Articles 34 and 35 DSA, MPIL notes that Instagram's Terms of Use set the minimum age to have an Instagram account at 13 years old, except in limited jurisdictions that have different age requirements, including Spain.¹⁴⁷ As explained in the 15 December Response, the Instagram SRA did not identify the potential for teen sign-in to Instagram as a systemic risk in and of itself. Rather, MPIL: (i) identified specific risks arising under various Problem Areas; and (ii) uses age-assurance measures as a 'control' to mitigate those risks in relation to users under the minimum age set in the Terms of Use (in addition to many other controls that mitigate those risks in relation to the broader population of users, including those aged over 13 years old).

169. These controls include Age Collection (CIDE-073), which prevents users under 13 from registering through a neutral registration screen and blocks the creation of an account if the age inserted is below 13. Additionally, after registration, MPIL has a broad set of mitigation and controls that apply, which include User self-reporting (Age Verification (CIDE-074)), User Initiated Reports (CIDE-058), and Age Prediction Modeling (CIDE-075), among others. For a

¹⁴⁵ [https://help.instagram.com/581066165581870/?helpref=hc_fnav] HYPERLINK

¹⁴⁶ Article 34(1) provides that "*Providers of very large online platforms and of very large online search engines shall diligently identify, analyse and assess any systemic risks in the Union stemming from the design or functioning of their service and its related systems, including algorithmic systems, or from the use made of their services.*" (emphasis added).

¹⁴⁷ The minimum age to use Instagram in Spain is 14 years old.

more detailed description of these controls, as well as an explanation of their design effectiveness, please refer to paragraphs 36-70 of the 15 December Response.

170. In relation to the age ratings of app stores, these do not equate with a minimum age policy. The age rating for an app is based on information provided by the app owner about the level of frequency with which a certain type of content (e.g., violence, gambling, nudity, etc.) appears on the app. On the basis of this information, the app store owner automatically generates a rating for the app.¹⁴⁸ While it is possible for an app owner to edit the age rating, the choice is limited to pre-defined age ranges set by the app store. There is no “13+” age threshold: for the Apple App Store, the available options are 4+, 9+, 12+, 17+ or no rating (adults only);¹⁴⁹ the Google Play store uses the age ratings of the Pan-European Game Information (PEGI), which are 3, 7, 12, 16 and 18.¹⁵⁰
171. Additionally, and as explained in the 15 December Response, there are inherent, industry-wide challenges of accurately assessing age at the app level, as opposed to the app store level. If parents can verify the age of their teen at the app store or device level, this will tackle at source the problem of minors under the age of 13 potentially being exposed to inappropriate content on social platforms.

¹⁴⁸ The process is described in more detail at this link for Apple “App Store”: [[HYPERLINK "https://developer.apple.com/help/app-store-connect/manage-app-information/set-an-app-age-rating" \h](https://developer.apple.com/help/app-store-connect/manage-app-information/set-an-app-age-rating)], and at this link for Google “Play Store”: [[HYPERLINK "https://support.google.com/googleplay/answer/6209544?hl=en" \l](https://support.google.com/googleplay/answer/6209544?hl=en)]
"zippy=%2Ceurope-middle-east" \h].

¹⁴⁹ See <https://developer.apple.com/help/app-store-connect/reference/age-ratings/>.

¹⁵⁰ See ‘Apps & Games content ratings on Google Play’ in the Google Play Help centre, [[HYPERLINK "https://support.google.com/googleplay/answer/6209544?hl=en" \h](https://support.google.com/googleplay/answer/6209544?hl=en)].

D. In relation to the default settings for content sharing between Instagram and Facebook

12. In Meta’s risk assessment report for Instagram, Meta states that for users under the age of 16, Instagram defaults private accounts.

a) Could you please clarify whether this measure also means that content posted on Instagram is, by default, not shared on Facebook, even if both platforms are in the child's accounts centre?

172. At the outset, MPIL would like to correct the noted statement by confirming that Instagram defaults private account and location settings when a user under 18 (not only under 16) years of age signs up for an account in the EU. As to question 12(a), MPIL confirms that content that users post on Instagram is not, by default, shared to Facebook.

b) If that is not the case, could you please elaborate which measures Meta has put in place to protect minors from “accidentally” sharing content on Facebook without being aware.

173. As stated above, content shared on Instagram is not shared by default on Facebook.

E. In relation to addictive behaviour and design features

13. Publicly available research¹⁵¹ as well as information provided in Meta's risk assessment report for Instagram¹⁵² seem to suggest that Instagram's interface, design and features might lead to addictive behaviour. Instagram's services seem to be, to a significant degree, based on a number of live features, the content of which may no longer be accessible, if not accessed or watched within a certain period after release, potentially creating a "fear of missing out". Underaged users appear to be particularly susceptible to such a "fear of missing out".¹⁵³

a) Please elaborate how you have assessed the risk of such features for the mental well-being of children and which mitigation measures you have put in place, if you identified a risk in this respect? In this respect, please provide us with any internal documents (communication between staff, assessments, internal studies) that look at this issue and how your service may affect underaged users in this respect?

174. At the outset, MPIL recalls its understanding that the term "*underaged users*" is intended to refer to Teens (*i.e.*, users aged of 13¹⁵⁴ - 17 years old).
175. MPIL refers to its response to question 11 of the 15 December Response, which also concerns the suggestion that Instagram's interface, design and features might lead to 'addictive behaviour'.

¹⁵¹ Sayed Ali Sharifi Fard et al, 'Basic Psychological Needs and Psychological Well-being: The Mediating Role of Instagram Addiction', 2023, [[HYPERLINK "https://link.springer.com/article/10.1007/s41347-023-00313-6"](https://link.springer.com/article/10.1007/s41347-023-00313-6) \h]; Purohit, Aditya Kumar, et al., 'Designing for Digital Wellbeing on a Smartphone: Co-creation of Digital Nudges to Mitigate Instagram Overuse', 2023; Rozgonjuk et al, 'Fear of Missing Out (FoMO) and social media's impact on daily-life and productivity at work: Do WhatsApp, Facebook, Instagram, and Snapchat Use Disorders mediate that association?', 2020, [[HYPERLINK "https://www.sciencedirect.com/science/article/pii/S0306460320306171"](https://www.sciencedirect.com/science/article/pii/S0306460320306171) \h]; Ponnusamy et al, 'Drivers and outcomes of Instagram Addiction: Psychological well-being as moderator', 2020, [[HYPERLINK "https://www.sciencedirect.com/science/article/pii/S0747563220300492"](https://www.sciencedirect.com/science/article/pii/S0747563220300492) \h]; Foroughi et al, 'Associations Between Instagram Addiction, Academic Performance, Social Anxiety, Depression, and Life Satisfaction Among University Students', 2021, [[HYPERLINK "https://link.springer.com/article/10.1007/s11469-021-00510-5"](https://link.springer.com/article/10.1007/s11469-021-00510-5) \h]; Respita Trias Ardiana, Raja Oloan Tumanggor, 'Social Media Instagram Addiction and Self-Esteem in High School Students, 2020, [[HYPERLINK "https://www.atlantispress.com/proceedings/ticash-20/125948100"](https://www.atlantispress.com/proceedings/ticash-20/125948100) \h]; Simon et al, 'Worried about being imperfect? The mediating effect of physical appearance perfectionism between Instagram addiction and body esteem', 2022, [[HYPERLINK "https://www.sciencedirect.com/science/article/pii/S019188692100725X"](https://www.sciencedirect.com/science/article/pii/S019188692100725X) \h].

¹⁵² In Meta's risk assessment report (p. 21), mitigation measures, 'such as parental controls and guides, default privacy settings and teen setting shortcuts' are stated to be used as preventative tools to mitigate risk to minors using the platform. In particular, the [[HYPERLINK "https://about.instagram.com/community/parents"](https://about.instagram.com/community/parents) \h] to Instagram daily time limits and breaks are mentioned. This suggests that Meta considered that Instagram might stimulate addictive behaviour as a risk to be mitigated.

¹⁵³ See, for example, V. Franchina et al, 'Fear of Missing Out as a Predictor of Problematic Social Media Use and Phubbing Behavior among Flemish Adolescents, [[HYPERLINK "https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6211134/"](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6211134/) \h]

¹⁵⁴ The minimum age to use Instagram in Spain is 14 years old.

176. MPIL rejects any suggestion that Instagram’s user empowerment tools such as daily limits, Quiet Mode and Take a Break mean that it considers that the platform might lead to ‘addictive behaviour’, or that there is sufficient evidence to reach that conclusion. In this regard, MPIL reserves its position as to the contents of the publicly-available research referred to in the question.
177. On the contrary, it is in Meta’s long-term commercial interests for its teen users to have a safe and positive experience on its platforms, and to ensure that the time users spend on the platforms is meaningful. Feelings of loss of control run contrary to that objective, and Meta has designed and implemented tools on Instagram such as daily time limits, Quiet Mode and Take a Break to empower users, including minors, and to provide them with a greater sense of control over their social media consumption.
178. MPIL did not, for the purpose of its compliance with Article 34 and 35 DSA, specifically assess the potential risk of live features creating a “fear of missing out” amongst teen users, or otherwise, on Instagram. Rather, and as described in response to question 11(b) of the 15 December Response, MPIL assessed the potential for ‘problematic use’ of Instagram in the Instagram SRA, as reflected in its assessment of the L4 Risk of ‘Mental Health’, contained in the Suicide and Self Injury Workbook.¹⁵⁵
179. For ease of reference, the risk assessment is summarised below:
- a. In row 6 of the tab ‘Risk Register & Scoring’, this risk was assigned a severity rating of 4.7 (including the highest score of 5 for ‘Impacted Demographic’ because of the specific exposure of minors, and a score of 4 for ‘Nature of Impact’ due to the physical and psychological impact) and a likelihood rating of 5, resulting in an Inherent Risk score of 23.5 (which, absent controls, would be Extremely Elevated) (see Residual Risk: Tiering in the Methodology¹⁵⁶).
 - b. In the SRA process, 70 controls were identified that mitigate this risk, set out in the Tab ‘Detailed Control Scoring’, including four Problem Area-specific controls.¹⁵⁷ Each control was assigned a figure to reflect its Design Effectiveness and Mitigation Effectiveness in respect of each risk.
180. A percentage reflecting the Design Effectiveness and Mitigation Effectiveness of the suite of relevant mitigating controls was applied to the Inherent Risk score, resulting in a Residual Risk score of 8.6, which is Low (see Residual Risk: Tiering in the Methodology¹⁵⁸). For more information on identifying and scoring risks, please see Question 7 above.
- b) In relation to any live features and features such as “Stories”, please provide the number and percentage of users of these features who are underaged, where possible broken down by age groups.**
181. MPIL notes that the RFI is limited to MPIL’s risk assessment and mitigation measures taken after the date of application of Articles 34 and 35 DSA. These did not take account of the number and percentage of teen users of Instagram live features. Notwithstanding the above, MPIL provides data on teen users of Instagram Live and Stories at [Annex 12](#).

¹⁵⁵ See Annex 22 of the 15 December Response.

¹⁵⁶ See Annex 4 of the 15 December Response.

¹⁵⁷ These were Mental Health Crisis Support (CIDPA-119), Well-being Guides (CIDPA-120), User Mental Wellbeing Support (CIDPA-129), and CIS Review Process / Classifiers (CIDPA-302).

¹⁵⁸ See Annex 4 of the 15 December Response.

c) Please set out your policy and actions in relation to users, who have disabled notifications by Instagram on their smartphones. In this respect, please provide further information on whether you remind users unprompted to turn on notifications again, your reasons for doing so, and the frequency of such notifications.

182. As a preliminary matter, it is important to note that Meta has no control or responsibility over the functioning of notification settings on user devices or operating systems (e.g. iOS/Android). As such, these settings are outside the scope of MPIL's systemic risk assessment and mitigation obligations under Articles 34 and 35 DSA, which expressly relate only to the services provided by MPIL.
183. Notwithstanding the above, MPIL confirms that users who disable the relevant app notifications in their device settings (e.g., iOS/Android) will not receive push notifications from Instagram when they are not using the app. When users are using the app, they may still see notifications in the app itself. However, these may be easily disabled by adjusting the settings in the Instagram app. Clear, user friendly instructions for how to do this on different types of devices (e.g., iOS/Android) are available on the Instagram Help Center.¹⁵⁹
184. As described in response to question 7(b) (paragraph 78(u) and footnote 72) of the 15 December Response, Instagram users can also silence push notifications for a period of up to 12 hours each day using the Quiet Mode feature. Further, in November 2023, Meta launched 'Bedtime Reminders'. Users with Quiet Mode enabled receive an in-app notification triggered 1-10 minutes before Quiet Mode is due to start, informing them that it is about to start. When Quiet Mode starts, users receive a dismissable blocking screen informing them that they are now in Quiet Mode.

d) Please state whether you have undertaken an assessment of the effects of the infinite scroll system on the screen time of users, where possible broken down by underaged and adult users. If so, please share any internal documentation, research, exchanges between staff documenting such assessment and remedies proposed.

185. MPIL did not undertake an assessment of any potential effects of "infinite scroll" functionality on the screen time of users for the purpose of its compliance with Article 34 and 35 DSA. Rather, as described in the response to question 13(a), MPIL rejects any suggestion that the time spent on its platforms by users (including teens) is harmful in and of itself. MPIL notes that the time management tools discussed in response to question 7(b) (paragraph 78(u) and footnote 72) of the 15 December Response have been developed to provide users with a greater sense of control over their social media consumption, which can also help, for example, interrupt prolonged periods of scrolling.
186. In addition, Meta is preparing to launch "nighttime nudges" on Instagram to encourage teen users to consider leaving the app during night hours. These will appear when teens have spent several minutes on a particular surface (e.g., Reels) between 10pm - 4am. This is an additional measure aimed at ensuring that teens leave Instagram feeling like the time they spent was meaningful and intentional.

¹⁵⁹ See: "How to turn push notifications on or off on Instagram", [[HYPERLINK "https://help.instagram.com/546541825361643/?helpref=uf_share" \h](https://help.instagram.com/546541825361643/?helpref=uf_share)].

e) Please confirm whether the time underaged users spent on Instagram is a data point you track (including daily average use and the number of sessions for daily users). If this data is tracked, please include all data and figures stemming from this tracking. In this respect please also indicate to what extent you actively take any measures to alter this performance indicator, either by attempting to increase or decrease the amount of time underaged users spent on Instagram. In this respect, please provide us with any internal documents (communication between staff, assessments, internal studies) that concern this issue and outline how your service may affect underaged users in this respect.

187. Independently of MPIL's systemic risk assessment and mitigation process under Articles 34 and 35 DSA, various Instagram teams track certain key data related to the time spent by teen (and other) users for a wide variety of business purposes, including, but not limited to, monitoring progress towards key business objectives (such as increased engagement or market share across different age groups). The Instagram Ecosystems Data Science team maintains a central dashboard that tracks certain key Instagram usage metrics for different age groups, including teens (13 - 17 year olds), as well as 18 - 24 year olds, 25 - 29 year olds, and over 30 year olds (30+).¹⁶⁰ In addition to providing an estimate of daily time spent on platform, this dashboard tracks data on daily active users ("DAU"), monthly active users ("MAU"), an internal metric to track the number of "Core Sessions" per DAU,¹⁶¹ and other engagement metrics such as data to estimate total sharing on platform for different Instagram products (e.g., Reels). In addition to this centrally maintained dashboard, different product teams within Instagram (e.g., for Reels, Live, Stories, etc.) may also separately track similar (or potentially more specific or purpose-built) data, in order to monitor their progress against relevant team and/or business objectives.
188. Since MPIL considers that time spent on-platform is not harmful in and of itself, it did not need to rely on these metrics as part of its systemic risk assessment and mitigation exercise under Article 34 and 35 DSA. Rather, and as described in response to question 11(b) of the 15 December Response, MPIL assessed the potential for 'problematic use' of Instagram, as reflected in its assessment of the L4 Risk of 'Mental Health' contained in the Suicide and Self Injury Workbook. To mitigate risks related to the potential for such problematic use, Meta offers various time management tools to help teens (and families) better understand and manage *how* they spend time on the platform and create healthy digital habits. These tools may also have the effect of decreasing overall time spent on platform, for example, as discussed at paragraph 150 of the 15 December Response in relation to Quiet Mode reducing late night time¹⁶² spent amongst teens who have adopted it.
189. As regards the request for "*all data and figures stemming from this tracking*", MPIL notes the extreme breadth of this request and the unclear connection that this information might have to MPIL's DSA compliance and reserves its position as regards the necessity and proportionality of this request. Specifically, given that these categories of usage data did not form part of MPIL's systemic risk assessment and mitigation exercise under Articles 34 and 35 DSA, it is not clear what specific data (or related internal documents) the Commission is requesting for the purposes of evaluating MPIL's compliance with these requirements. To the extent that the Commission requires certain specific data relating to time spent by teen users on Instagram, in addition to what has already been provided in response to questions 13(b), above, and questions 11(d) and (e) of the 15 December Response, MPIL respectfully invites the Commission to identify which data is requested, and to clarify the relevance of any such request to MPIL's compliance with the DSA.

¹⁶⁰ The dashboard is based on a 1% sampled user data and backfilled age data, and relies on Meta's Age Prediction Modelling (described in response to question 4 of the 15 December 2023 Response) to estimate a user's age.

¹⁶¹ For these purposes, a "Core Session" is defined as a user that opens Instagram and keeps it open for at least 3 seconds.

¹⁶² 10pm to 4am.

f) Please provide us with any internal data you may have, including studies and internal reports, whether collected as part of your risk assessment process or in your possession otherwise, on your strategies to attract underaged users to join and use Instagram.

190. As indicated above, MPIL understands from its meeting with the Commission on 19 December 2023 that the term “*underaged users*” is intended to refer to Teens.
191. MPIL does not consider that the fact of teens joining and using Instagram is harmful, or presents a systemic risk, in and of itself. As a result, it did not consider the existence (or not) of any specific strategies to attract teen users to join and use Instagram as part of its systemic risk assessment and related mitigation exercise pursuant to Articles 34 and 35 DSA. Rather, MPIL considered in detail and across a number of ‘Problem Areas’¹⁶³ the various specific systemic risks that could arise from teens being exposed to inappropriate content or other potentially harmful behaviour once they are on the platform. This is discussed in detail elsewhere in this response and in the 15 December Response.
192. In any event, MPIL has not identified any internal data, studies or internal reports relevant to its systemic risk assessment and mitigations exercise under Articles 34 and 35 DSA that pertains to any strategy to specifically attract teen users to join and use Instagram.

F. In relation to access of minors to potentially harmful content

14. Publicly available research¹⁶⁴ suggests that exposure to certain Instagram content and photo manipulation may negatively impact young people’s body image and how they think about themselves.

a) Please explain whether, to what extent and in what ways Meta analysed and assessed whether, and how, the use of Instagram may negatively impact people’s body image and the way they think about themselves, distinguishing between minors and other users. Please, submit, on the basis of Article 34(3), any supporting documents in relation to the identification and assessment of this (potential) systemic risk.

193. MPIL refers to its response to question 9(a) in its 15 December Response, which detailed how the Instagram SRA assessed and responded to the issue of body image, particularly as those efforts relate to minors. As indicated there, this was considered under the Problem Area of Suicide & Self-Injury, and specifically the L4 risks of (i) Disordered Eating, (ii) Mental Health, and (iii) Personal Health Ads and Appearance Ads.
194. In response to the request to submit the supporting documents in relation to the identification and assessment of risks related to negative body image, MPIL submits, in addition to the SRA materials with its 15 December Response, the Suicide & Self-Injury Workbook.¹⁶⁵

¹⁶³ See in more detail paragraph 15 of the 15 December Response.

¹⁶⁴ See for example: Alfonso-Fuertes et al, ‘Time Spent on Instagram and Body Image, Self-esteem, and Physical Comparison Among Young Adults in Spain: Observational Study’, 2023, [HYPERLINK "<https://pubmed.ncbi.nlm.nih.gov/37027197/>" \h].

¹⁶⁵ See Annex 10 of the 15 December Response.

- b) Please provide us with all your internal documents, including studies, internal reports or exchanges between staff members, on the potential risks of social comparison and whether this practice can cause or exacerbate mental health issues such as but not limited to distorted body image, eating disorders, anxiety, loneliness, depression, envy, online aggression, and passive use.
195. For the purpose of MPIL's compliance with Articles 34 and 35 DSA, its assessment of the potential risks of social comparison and whether this can cause or exacerbate mental health issues, particularly as those risks relate to minors, the relevant internal document is the Suicide & Self-Injury Workbook, referred to in response to question 14(a) above and discussed in more detail in the 15 December Response.
- c) Please indicate and explain the measures Meta has put in place to prevent minors from accessing content that may cause users to feel worse about their body or appearance ("negative appearance comparison"), i.e., through the use of filters that suggest minor "improvements" that would be feasible with cosmetic procedures. Please explicitly elaborate in this respect on any assessment you may have undertaken and any internal documents you have on the effect of filters on girls below the age of 18.
196. At the outset, MPIL understands that Articles 34 and 35 DSA do not impose an obligation on MPIL to generally prevent teens from accessing content that may cause them to feel worse about their body or appearance, including through the use of filters that suggest minor "improvements" that would be feasible with cosmetic procedures. Rather, Articles 34 and 35 DSA require MPIL to assess and mitigate systemic risks, notably those of Protection of Minors and Public Health, which MPIL did, assessing in particular the risks of Disordered Eating, Mental Health and Personal Health Ads and Appearance Ads, under the Problem Area of Suicide and Self-Injury. MPIL's assessment and mitigations of these risks is explained in response to question 14(a) above and question 9 in the 15 December Response.
197. That said, and separate to MPIL's risk assessment and mitigation measures under the DSA, Meta recognises that negative social comparison can be difficult, especially for teens. This is why Meta bans effects that directly promote or encourage dangerous cosmetic surgery procedures (for example, effects that depict cosmetic procedures through surgery lines), changes in skin colour or extreme weight loss). Meta has also designed and implemented tools and other product solutions on Instagram that may reduce negative comparison for all users, including teens.
198. These measures include:
- a. Alternative Topic Nudge (see paragraph 78(d) of the 15 December 2023 Response),
 - b. Sensitive Content Controls (see paragraph 78(t) of the 15 December 2023 Response),
 - c. Hiding like counts (see paragraph 78(j) of the 15 December 2023 Response),
 - d. Limits and Mute (see paragraph 78(a) of the 15 December 2023 Response),
 - e. 'Not Interested' (see paragraph 96 of the 15 December 2023 Response),
 - f. Parental supervision (see paragraph 78(m) of the 15 December 2023 Response), and
 - g. Time management tools (see paragraph 78(u) of the 15 December 2023 Response).
199. With respect to the potential effect of filters on teen girls, MPIL has not identified any assessments of this issue relevant to its systemic risk assessment and mitigations exercise under Articles 34 and 35 DSA.

d) Publicly available research suggests that self-harm and suicide content is both highly prevalent and discoverable for minors on Instagram, particularly on Reels.¹⁶⁶ The findings of the report include that problematic hashtags can still be accessed without links to support sources, that sensitivity screens meant to blur graphic images were used in less than 1% of posts identified in the research, and the algorithm promotes large amounts of content or accounts that promoted suicide or self-harm behaviours.¹⁶⁷ Please set out which measures Meta has put in place to prevent the amplification of harmful content, for example in relation to suicide, self-injury, eating disorders, body image issues. If measures have been put in place, please include any information and/or evidence collected on the effectiveness of the measure to mitigate the risk.

200. For the purpose of responding to this RFI, MPIL reserves its position in relation to the contents of the publicly-available research referred to in the RFI.

201. We understand this question to be repeating in substance question 10 of the 15 December Response, which asked MPIL to describe in detail how the content moderation processes regarding pro-suicide, glorification of self-harm and eating disorders content function, including relevant performance metrics. We therefore refer to MPIL's response to that question in the 15 December Response.

e) Please provide us with any research, analysis or studies Meta undertook or is aware of related to potential amplification of content which may be harmful to underaged users' mental well-being by Instagram's recommendation algorithm. Please outline and explain to what extent Meta has made changes to the recommendation algorithm in light of this information and your reasons for changing it (or your reasons for not changing it).

202. Please see above in response to question 10(c).

f) Please set out which additional risk mitigation measures Meta put in place following the information outlined in question 14(e), including whether Meta considered to change its recommendation algorithms for users under the age of 18. Please include in your answer the reasons for choosing the specific mitigations measures, supported by internal or external evidence. Please include the necessary (references to) internal studies and assessments, as well as exchanges between staff on this topic.

203. Please see above in response to question 10(c).

¹⁶⁶ *Preventable yet pervasive: The prevalence and characteristics of harmful content, including suicide and self-harm material, on Instagram, TikTok and Pinterest*, Molly Rose Foundation, in partnership with The Bright Initiative, November 2023, [[HYPERLINK "https://mollyrosefoundation.org/november-2023-new-research-exposes-tech-giants-amplification-of-content-promoting-suicide-and-self-harm/"](https://mollyrosefoundation.org/november-2023-new-research-exposes-tech-giants-amplification-of-content-promoting-suicide-and-self-harm/)]. As part of the research, in an analysis of 100 algorithmically recommended videos watched through the autoplay function through an account set up with the identity of a fifteen-year-old girl, 99% of the material either "promoted or glorified suicide or self-harm, referenced suicide or self-harm ideation, or otherwise featured highly intensive themes of depression, hopelessness or misery, see p. 20.

¹⁶⁷ *Ibid*, p.23-25.

Glossary of terms

ANSA	Adult Nudity and Sexual Activity
AQVM	Account Quality Value Model
CSAM	Child Sexual Abuse Materials
CSE	Child Sexual Exploitation
CSEAN	Child Sexual Exploitation, Abuse & Nudity
CSTF	Child Safety Task Force
CSx	Child Sexualisation
DAU	Daily Active Users
DIOps	Discovery Intelligence Ops
DSA	Digital Services Act
EII	Exploitative Intimate Imagery
I3 Team	Meta’s Integrity, Investigations, & Intelligence Child Safety team
IGRC Team	Meta’s Integrity Governance Risk & Compliance team
MaCSA	Malicious Child Safety Actor
MAU	Monthly Active Users
MMS	Media Match Service
MPIL	Meta Platforms Ireland Limited
NCMEC	United States National Center for Missing and Exploited Children
SIO	Stanford Internet Observatory
SG-CSAM	Self-generated CSAM
SMEs	Subject-Matter Experts
SND	Strategic Network Disruption
SRA	Systemic Risk Assessment
SRT	Single Review Tool
SSPx	Adult Sexual Solicitation and Prostitution
TMS	Text matching service

Private and Confidential – Contains Commercially Sensitive Information

UnIS	Unwanted Interaction Solutions
URL	Uniform Resource Locator
VLOP	Very Large Online Platform
WSJ	Wall Street Journal

Schedule of Annexes

Please note that some of the documents provided as Annexes and listed in this Schedule may be subject to changes and periodic updating. Further, all these documents are confidential and contain commercially sensitive information.

Number	Document	Relevant Question(s)
1	CSEAN Control Descriptions	Introduction, 1(b)
2	Meta's Response to EC of 29 August 2023	1(a)
3	[Update] Keywords/Hashtags review on WSJ escalation	2(c)
4	[LAUNCH] Fuzzy matching for Child Safety keywords	2(c)
5	Ongoing maintenance of child safety keywords	2(c)
6	Tracking CSE on Recommended surfaces	4(a)
7	Child Safety Network Disruption on Commercial CSAM Networks	4(b)
8	Teen Interaction Safety H1 Milestone Goal Complete: Deliver MaCSAto-MaCSA Interventions	8(a)
9	[Launch] Filter MaCSA Strict for Gen-pop on Reels/Explore/Feed Rec	8(a)
10	SSI/ED External Advisor Feedback Synthesis	10(c)
11	Appearance Comparison Strategy Refresh	10(c)
12	Data on teen users of Instagram Live and Stories	13(b)