# Regulation (EU) 2022/2065 Digital Services Act (DSA)

## Systemic Risk Assessment and Mitigation Report for Facebook

August 2024

# Table of Contents


Executive Summary 4
1. Introduction 8
1.1 Purpose 8
1.2 Scope 8
1.3 Approach 8
1.3.1 Enhancements to Risk Assessment Methodology 9
1.4 Limitations and Assumptions 9
2. An Overview of Facebook 10
3. A Balancing Act: Respecting Rights and Mitigating Risk 11
3.1 Meta's Commitment to Respecting Voice and Enhancing User Safety 12
3.2 Complaints and Appeals 12
4. Meta's DSA Systemic Risk Assessment Methodology 14
4.1 Risk Assessment Process 14
4.2 External Stakeholder Engagement 15
4.3 Emerging and Unknown Risks 16
5. Systemic Risk Landscape 17
5.1 Systemic Risk Areas 19
5.1.1 Deceptive and Misleading 19
5.1.2 Civic Discourse and Elections 19
5.1.3 Public Health 22
5.1.4 Public Security 22
5.1.5 Gender-Based Violence 23
5.1.6 Protection of Minors 24
5.1.7 Fundamental Rights 26
5.1.8 Illegal Content 27
5.2 Influencing Factors 27
5.2.1 Recommender Systems 27
5.2.2 Content Moderation Systems 29
5.2.3 Terms of Service and their Enforcement 31
5.2.4 Ads Systems 32
5.2.5 Data Related Practices 34
5.2.6 Intentional Manipulation 35
5.2.7 Generative Artificial Intelligence (AI) 36
6. Our Detailed DSA Systemic Risk Assessment Results 39
6.1 Risk Analysis 40
6.1.1 Risk Rating 40
6.1.2 Problem Area Analysis: Inherent Risk 40
6.1.3 Problem Area Analysis: From Inherent Risk to Residual Risk 41
6.1.4 Year-Over-Year (YoY) Results Comparison 42
6.1.5 Systemic Risk Area Analysis: Risk Ratings 44
6.2 Mitigating Measures Analysis 44
6.2.1 Meta's Ecosystem of Controls 45
6.2.1.1 Policies and Standards 45
6.2.1.2 Systems and Product Integrity 47

6.2.1.3 Detection 52
6.2.1.4 Enforcement 54
6.2.1.5 Response and Notification 56
6.2.1.6 User Rights and Recourse 58
6.2.1.7 External Awareness and Support Resources 60
6.2.1.8 Internal Training and Resources 62
6.2.1.9 Risk Assessment 62
6.2.1.10 Governance 63
6.2.2 Detailed Risk Observations and Mitigating Measures 65
6.2.2.1 Account Integrity and Authentic Identity 65
6.2.2.2 Adult Sexual Exploitation and Adult Nudity 66
6.2.2.3 Bullying and Harassment 67
6.2.2.4 Child Sexual Exploitation, Abuse and Nudity 69
6.2.2.5 Coordinating Harm and Promoting Crime 70
6.2.2.6 Dangerous Organisations and Individuals 71
6.2.2.7 Discrimination / Discriminatory Actions 72
6.2.2.8 Disinformation 73
6.2.2.9 Fraud and Deception 75
6.2.2.10 Hate Speech 75
6.2.2.11 Human Exploitation 76
6.2.2.12 Inauthentic Behaviour 78
6.2.2.13 Intellectual Property (IP) Infringement 79
6.2.2.14 Misinformation 80
6.2.2.15 Privacy and Security 81
6.2.2.16 Restricted Goods and Services 83
6.2.2.17 Suicide and Self-Injury 84
6.2.2.18 Violence and Incitement 85
6.2.2.19 Voice and Free Expression 86
7. Risk Mitigation Enhancements 88
8. Conclusion 89
9. Appendix 90
9.1 Meta's Integrity Risk Assessment Methodology: Rubrics 90
9.1.1 Inherent Risk Rubrics 90
9.1.1.1 Severity Rubrics 90
9.1.1.2 Likelihood Rubrics 91
9.1.2 Control Effectiveness Rubrics 91
9.1.2.1 Design Effectiveness 92
9.1.2.2 Operational Effectiveness 93
9.1.2.3 Mitigation Effectiveness 93
9.1.2.4 Control Suite Effectiveness Calculation 93
9.1.3 Residual Risk Calculation 94
9.2 Principles for ensuring Reasonable, Proportionate, and Effective Mitigation Measures 94

# Executive Summary

Meta's mission is to give people the power to build community and bring the world closer together. We help people discover and learn about what is going on in the world around them, enable people to share their experiences, ideas, photos and videos, and other activities with audiences ranging from their closest family members and friends to the public at large, and stay connected everywhere by accessing our products.[1] For the 6-month period ending 31 March 2024, we have approximately 260.7 million average monthly active users on Facebook in the European Union (EU)[2] who are reaping the benefits of connectedness.

Although Facebook has been used to build communities, raise awareness, and grow small businesses, the risk remains for our services to be, in some instances, abused and manipulated. We take a risk-based approach for implementing mitigation measures to combat problematic actors, behaviour, and content on Facebook. We refer to our collective work combating problematic actors, behaviours, and content on Facebook as "Integrity Ecosystem" efforts, and the specific mitigating measures we deploy as "controls".[3] The backbone of our Integrity Ecosystem is our suite of policies, specifically our Facebook Community Standards, Ad Standards and Commerce Policies which outline what is and is not allowed on Facebook. We bolster this ecosystem using a three-line of defence model to manage risk, compliance, and operational changes.

### DSA Systemic Risk Assessment

Meta Platforms Ireland Limited, as the provider of Facebook in the EU, has undertaken its annual DSA Systemic Risk Assessment of Facebook. This Report sets out the results of the risk assessment conducted between September 2023 and August 2024. We will release a public version of this Report.

It is important that readers note when interpreting or commenting on this Report that (i) this is a European regulation and (ii) our risk assessment methodology involves a phased approach to identify, qualify, assess, measure, validate, respond and mitigate, and report out on identified risks and mitigations. As a result, risks that surface and mitigations that are implemented after the relevant risk assessment phase has been completed may be captured through our Issue Management Programme and reflected in next year's EU DSA Systemic Risk Assessment (SRA) Report 2025. Additionally, the scope of this assessment is limited to the Systemic Risk Areas as defined in Article 34 of the DSA, as well as Deceptive and Misleading, which is a Systemic Risk Area we added during our Year 1 (Y1) SRA that covers behaviour and content that is designed to deceive, mislead, or defraud users usually for personal gain.

The eight Systemic Risk Areas analysed are as follows:


[1] Meta's Platform Inc, Form 10-K, 2023
[2] Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Facebook
[3] Controls are a combination of people, processes, policies, and tools that Meta has put in place to mitigate integrity risks and prevent, detect, or correct integrity issues. Controls include any system, process, policy, device, practice, or other actions which reduce the likelihood or impact of a given risk occurring.

Our content policies, including our Facebook Community Standards, Ad Standards, and Commerce Policies, cover the Systemic Risk Areas listed in the graphic above. However, we break them down into more granular categories we refer to as "Problem Areas".[4] See Section 5: Systemic Risk Landscape for further details.

### What Has Changed between Year 1 and Year 2?

Based on feedback from the European Commission, our lessons learnt from the DSA Systemic Risk Assessment Year 1 (Y1), and insights from our Year 2 (Y2) assessment, the following changes have been implemented and/or observed:

- **Risk Landscape:** In Y2, we identified 122 risks associated with the 19 Problem Areas and 8 Systemic Risk Areas. The change in the number of risks identified, from 120 risks in Y1, is a result of the maturation of our risk assessment process, which included adding, consolidating and/or remapping risks within Problem Areas and Systemic Risk Areas;
- **Influencing Factors:** When evaluating the identified risks each year, we considered the role of each Influencing Factor defined in Article 34 of the DSA and the impact on the systemic risks.[5] Like the Systemic Risk Areas outlined in Article 34 of the DSA, we found that the Influencing Factors defined in Article 34 were non-exhaustive. Last year, we added "Deceptive / Misleading" as a Systemic Risk Area, and similarly this year (Y2), we added **Generative Artificial Intelligence (generative AI)** as an influencing factor. The rapid expansion of generative AI creates unlimited opportunities, including opportunities for abuse. To manage any unintended consequences, Meta has in place mechanisms to evaluate the impact of generative AI on Problem Area risks and how generative AI can be used to mitigate Problem Area risks on the platform;
- **Operating Effectiveness:** Meta employs a sophisticated set of controls for mitigation of risks. During Y1, Meta measured design effectiveness to derive overall residual risk; this year we have matured our assessment process to include operational effectiveness of our controls to determine residual risk. To enable this, we expanded our effectiveness evaluation signal base to include input from assurance testing, issue management, and platform data;
- **Continuous Improvement:** As part of our journey of continuous improvement, we routinely evaluate our Integrity Ecosystem to identify enhancement opportunities. As a result, we have enhanced systems and practices since Y1 including, but not limited to, the following:
    - ***Child Safety Measures*** - we enhanced and expanded our child safety measures, including improving our ability to automatically disable threat actors with our malicious child safety actor model and limiting recidivism via strict account and device linking policies. We also developed a unified keyword list to block searching violating content, and reduced teens' exposure to potentially child safety violating content via extensive recommendation filtering;

[4] These Problem Areas map to our Facebook Community Standards and may vary in naming convention.

[5] Article 34 (2) of the Digital Services Act defines Influencing Factors as: (a) the design of their recommender systems and any other relevant algorithmic system; (b) their content moderation systems; (c) the applicable terms and conditions and their enforcement; (d) systems for selecting and presenting advertisements; (e) data related practices of the provider. It also includes how the risks in Article 34 (1) are influenced by intentional manipulation of their service, including by inauthentic use or automated exploitation of the service, as well as the amplification and potentially rapid and wide dissemination of illegal content and of information that is incompatible with their terms and conditions.

- ***Programmatic Capabilities*** - we advanced our Global Integrity Governance, Risk and Compliance (GRC) function by scaling our integrity control assurance testing capabilities and maturing our issues management programme to manage issues identified across our ecosystem;
- ***Election Integrity Measures*** - we expanded our overall approach to managing election integrity on our platforms for EU Parliamentary elections in collaboration with the European Commission, other platforms, and civil society partners, including the signatories of the EU Disinformation Code;[6] and
- ***Generative AI Countermeasures*** - implementing focused measures to counter the risk related to industry generative AI technologies, including tools to identify invisible markers and label AI-generated content, ability to tag AI-generated content identified by fact-checkers and platform users, and partnering with other companies and industry bodies (e.g. Partnership on AI (PAI)) to build common standards and guidelines to combat the spread of deceptive AI content.[7]

- **Single-Combined Report:** Last year, our mitigation measures were shared in a separate DSA Systemic Risk Mitigation Report 2023; this year (Y2), our risk mitigations are included in Section 6.2 Mitigating Measures Analysis.

### DSA Systemic Risk Assessment Results Overview

During this year's DSA Systemic Risk Assessment (Y2), the integrity risk landscape shifted and elevated due to (1) the various elections that were carried out across the European Union which introduced increased abusive online behaviour and content, including Violence and Incitement, Misinformation and Disinformation; (2) global events, including, but not limited to, conflicts in adjacent regions and the preparation for the 2024 Olympics which involved more movement of people, potentially increasing the risk of human exploitation; and (3) the rapid expansion of generative AI which can be used to manipulate media and impersonate individuals.

Some of the key highlights as it relates to the Y2 assessment results are as follows:

- **Inherent Risk:** The majority of Problem Areas were impacted by external events that increase the complexity of the risk landscape and potential for abusive online behaviour and policy-violating content; however only two Problem Areas changed Inherent Risk Tiers.
- **Residual Risk:** Year-over-Year (YoY) our Residual Risk Tiers remained constant for around 95% of our Problem Areas and around 5% changed from Tier 1 to 2.

### The Road Ahead

We remain committed to providing a platform that provides value to our users and society at large and protects people's rights including freedom of expression, while continuing to enable innovation. For this reason, we have been working hard since the DSA came into force in November 2022 to respond to these new rules and adapt the existing safety and integrity systems and processes we have in place in many of the areas regulated by the DSA. We assembled one of the largest cross-functional teams in our history, with over 1,000 people having worked on the DSA,[8] to develop solutions to meet its requirements.

[6] https://about.fb.com/news/2024/02/how-meta-is-preparing-for-the-eus-2024-parliament-elections/
[7] https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/
[8] https://about.fb.com/news/2023/08/new-features-and-additional-transparency-measures-as-the-digital-services-act-comes-into-effect/

We look forward to working with the European Board for Digital Services to incorporate feedback from this Y2 Systemic Risk Assessment and continue enhancing our integrity measures towards the shared objectives of minimising harm effectively, protecting and empowering people, and upholding their fundamental rights.

# 1. Introduction

Pursuant to Article 34 and Article 42 of Regulation (EU) 2022/2065 (DSA), Meta Platforms Ireland Limited ("Meta") is pleased to provide our annual DSA Systemic Risk Assessment Report (the "Report") for Facebook. We are committed to maintaining a safe, reliable, and trustworthy online environment across all of our services, including Facebook. Our apps and services are designed to give people a voice and support fundamental rights, which is aligned with the EU's goal of creating "...a safer digital space in which the fundamental rights of all users of digital services are protected."[9]

### 1.1 Purpose

The purpose of this Report is to detail the results of Meta's annual DSA Systemic Risk Assessment and the reasonable, proportionate, and effective mitigation measures in place to address systemic risks evaluated during this assessment in accordance with Articles 34, 35, 42(4)(a) and (b) of the DSA.

Meta has identified, analysed, and assessed the systemic risks in the EU that could stem from or be influenced by the following:

- Problematic actors, behaviour (e.g., harassment), or content (e.g., hate speech) that violates our Terms and/or may be considered illegal;
- App design or functionalities, including but not limited to algorithmic systems (e.g., recommender systems); or
- The use made of our services.

### 1.2 Scope

Per the requirements stated in Article 34 of the DSA, Meta conducted its annual DSA Systemic Risk Assessment of Facebook between **September 2023 and August 2024**. The analysis and results of this assessment are limited to this time period and capture both actual and foreseeable risks.

Pursuant to the European Commission Decision designating Facebook as a very large online platform (VLOP) in accordance with Article 33(4) of the DSA, the Facebook VLOP does not include private messaging services like Facebook Messenger which, based on its technical functionalities, do not meet the definition of an online platform. As such, Facebook Messenger is not in scope for this Report.

The focus of this assessment is the Systemic Risk Areas defined in Article 34 of the DSA and our interpretation of the associated systemic risks in the EU. Whilst our systems and policies are global in nature, this Systemic Risk Assessment reflects risks that could have an impact in the EU, as well as the reasonable, proportionate, and effective mitigation measures Meta has in place to address the aforementioned Systemic Risk Areas.

### 1.3 Approach

Meta continues to evolve its practices and respond to content-related regulations, including by establishing and continuously enhancing our Integrity, Security, Support, and Operations Governance, Risk and Compliance (ISSO GRC) Programme. This programme includes an Integrity Risk Management Process that pulls from ISO 31000: Risk Management as a leading practice, is tailored to meet the needs of our

[9] https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

environment, builds on the existing integrity measures we have had in place for years, and accounts for feedback provided by the European Commission.[10]

The Integrity Risk Assessment Process and the associated outputs were tailored to meet the scope and needs of the Systemic Risk Assessment required under Article 34 of the DSA.

#### 1.3.1 Enhancements to Risk Assessment Methodology

Over the last year, we enhanced our Risk Assessment Methodology to account for the operating effectiveness of our control environment in our Control Suite Effectiveness calculation. Along with signals from the assessment workshops, we also increased our signal base to further inform our evaluation of the risk and control environment, including the following signals:

- **ISSO GRC Integrity Issue Management Programme:** Relevant information, from the centralised issue management programme that identifies, manages and tracks remediation of integrity related deficiencies, has been integrated into the process for the purposes of calculating control operating effectiveness;
- **Control assurance results:** The data from the periodic control testing activities and identified deficiencies have been leveraged to further inform the control design and operating effectiveness; and
- **Metrics and data:** Where available, metrics and data points were used as signals to inform control effectiveness.

### 1.4 Limitations and Assumptions

There were a number of limitations and assumptions associated with carrying out this Systemic Risk Assessment, including:

- **Guidance and Standards:** Our approach in this Report continues to reflect our understanding of the risk assessment requirements based on the text of the DSA. Over the last 20 years, we have established our own standards and practices to identify, assess, and manage Problem Areas and the associated risks, which has informed the development of our Integrity Risk Assessment Process. Additionally, we have leveraged guidance and standards relevant to integrity risks in the EU, such as ISO-31000 and the United Nations Guiding Principles on Business and Human Rights, to inform the design and execution of our Integrity Risk Assessment Process. We look forward to receiving guidance from the European Board for Digital Services to understand the lessons learned from this assessment and our 2023 assessment, which we expect will contribute to the further development of guidance and standards for the industry.
- **Signal Parity:** To inform the evaluation of risks and controls, we have gathered internal signals from different sources, including our issue management programme, assurance results, workshops, surveys, and publicly available transparency reporting data. While the signals provide a combination of qualitative and quantitative insights, different risks and controls will have different signals. Meta is on a journey of continuous improvement, and our risks and controls signal quality and strength will further improve and mature as we enhance our programme.

[10] https://www.iso.org/iso-31000-risk-management.html

# 2. An Overview of Facebook

Meta builds technologies that help people connect, find communities, and grow businesses. Facebook helps give people the power to build community and bring the world closer together. It's a place for people to share life's moments and discuss what's happening, nurture and build relationships, discover and connect to interests, and create economic opportunity.[11]

With approximately 260.7 million average monthly active Facebook users in the EU, Facebook helps people make connections and navigate life's many milestones with its various features and services.[12] There are different types of content that enable our communities to interact, including:

- Organic content (e.g., content generated by users);
- Paid content (e.g., ads created by brands and businesses); and
- Commerce content (e.g., products listed by merchants on Facebook).

Facebook provides the tools for people to do more together and empowers people to support each other around the world. Our global community has now raised approximately $8 billion for non-profits and personal causes through fundraisers on Facebook and Instagram.[13] Examples of how people are using Facebook to support each other around the world and how Meta is enabling this include the following:

- **Creators and Monetisation:** As creators look for ways to connect safely and more widely with their communities, we wanted to make it easier for anyone to become a creator and earn money on Facebook through new forms of expression, professional tools and monetisation programmes. For example, we have evolved our payout model to pay creators based on how well their content performs on Facebook, simplifying how creators earn and expanding monetisation opportunities.[14]
- **Voter Empowerment and Election Integrity:** Facebook values civic engagement and empowers voter participation in elections through providing access to crucial information, fostering community involvement and enhancing electoral integrity. Drawing upon the lessons learnt from 200 elections around the world since 2016, we have invested more than $20 billion into safety and security - including but not limited to election integrity - and quadrupled the size of our global team working in this area to around 40,000 people.[15] We also had a dedicated Elections Operations Centre for the EU Parliamentary Elections, bringing together experts from across the company from our intelligence, data science, engineering, research, operations, compliance, content policy and legal teams, to identify potential threats and put mitigations in place across our technologies in real time.
- **Business Growth:** Facebook contributes significantly to the EU economy, and it is helping users discover new products and brands that are most relevant to them. Personalised advertising enables businesses of all sizes to find customers and grow their presence.

[11] ANNUAL REPORT PURSUANT TO SECTION 13 OR 15(d) OF THE SECURITIES EXCHANGE ACT OF 1934
[12] Regulation (EU) 2022/2065 Digital Services Act Transparency Report for Facebook
[13] https://www.facebook.com/about/social-impact/
[14] https://about.fb.com/news/2024/05/the-future-of-facebook/
[15] https://about.fb.com/news/2024/02/how-meta-is-preparing-for-the-eus-2024-parliament-elections/

# 3. A Balancing Act: Respecting Rights and Mitigating Risk

Meta is committed to respecting the fundamental rights of our users located in the EU as outlined in the DSA. Our second annual Human Rights Report demonstrates how we are living up to the commitments made in our Corporate Human Rights Policy. We seek to champion respect for human rights in every action we take and every product we build. However, as is the case offline, certain aspects of human rights can at times be in tension with one another, such as the need to balance freedom of expression with the need to prohibit hate speech. At Meta, we strive to strike the right balance in every action we take, which includes the following:

- Our Facebook Community Standards, which outline what content is and is not allowed on Facebook, have human rights principles embedded into them;
- We consistently use feedback from our community and the advice of experts in fields to inform our Facebook Community Standards. To enable this feedback and enhance transparency, we have a robust policy management engagement process in place, which includes an outreach strategy for connecting with global stakeholders who are most affected by the policy change, and who have relevant expertise and lived experience. We post a summary of this engagement alongside the revised policy language in our Transparency Centre whenever we change our policies;
- As a part of the policy development process, the Human Rights Team and the Civil Rights Team conduct separate rights-based analysis and due diligence of proposed policies, submit such analysis to policy leadership, and present their views to each Policy Forum. Human and civil rights impacts and mitigations are a consistent part of policy development at Meta;
- We strive to assess human rights potential impacts through human rights due diligence as laid out in our Corporate Human Rights Policy and in alignment with UNGPs 17 and 21, the International Bill of Rights and the EU's Charter of Fundamental Rights, among others;[16]
- We provide pathways for stakeholders to report potentially problematic content, for Meta to review such content, and for Meta to create remediation consistent with UNGP 31. We maintain multiple grievance pathways, identified in the Help Centre on platforms and apps, including an appeals process to the first-of-its-kind Oversight Board;[17]
- We undertake proactive measures to maintain the momentum for addressing human rights related risks. These include our Comprehensive Human Rights Salient Risk Assessment (CSRA); ongoing product counselling; integration of human rights risks into content risk forecasting; and processes to respect freedom of expression and privacy, as mandated by our membership in the Global Network Initiative (GNI). We also offer human rights training ("Bigger than Meta") for employees,[18] as well as customised training;
- We have strengthened our governance systems to advance our work toward respecting human rights across all our services. This includes continuing to empower the Oversight Board, which has issued 268 non-binding recommendations from January 2021 through 19 July 2024, across policy, enforcement and transparency.[19]

[16] https://humanrights.fb.com/wp-content/uploads/2023/09/2022-Meta-Human-Rights-Report.pdf
[17] https://humanrights.fb.com/wp-content/uploads/2023/09/2022-Meta-Human-Rights-Report.pdf
[18] https://humanrights.fb.com/wp-content/uploads/2023/09/2022-Meta-Human-Rights-Report.pdf
[19] https://transparency.meta.com/en-gb/oversight/overview

### 3.1 Meta's Commitment to Respecting Voice and Enhancing User Safety

At Meta, we believe and work to protect freedom of opinion and expression. While we remain committed to respecting our users' voices, we also need to balance safety. Several of Meta's policies include a freedom of expression element that is taken into consideration by our detection and enforcement mechanisms. For example, while we do not allow content that promotes or celebrates suicide, Meta allows content that depicts recovery from attempted suicide, self-injury or eating disorders, such as healed wounds, as described in our Suicide, Self-Injury and Eating Disorders Community Standards.

In our commitment towards balancing voice and safety, we developed our Facebook Community Standards that outline what is and what is not allowed on Facebook. These standards are based on feedback from people and the advice of experts in fields like digital rights, freedom of expression, public safety, journalism, elections, and human and civil rights. To ensure everyone's voice is valued, we take great care to create standards that include different views and beliefs, especially those from marginalised communities. To enable this, we developed an Inclusivity Framework to ensure that the views of our diverse stakeholders are considered in the development of our policies and Community Standards.[20]

We also provide context about what users see and provide warning screens for content that some may find sensitive, so users can make their own decisions on what to read, trust, and share.[21] Furthermore, we have a network of Trusted Partners comprising over 400 non-governmental, not-for-profit, national, and international organisations in 113 countries who report content, accounts, and behaviour that we review with the benefit of local context provided by the Partners.[22] We have also established mechanisms for reviewing reports of allegedly illegal content from "Trusted Flaggers" in compliance with the DSA.

In rare cases, we may allow content which would otherwise go against our Facebook Community Standards if it's newsworthy and if keeping it visible is in the public interest. We do this only after conducting a thorough review that weighs the public interest value against the risk of harm, and we look to international human rights standards as reflected in our Corporate Human Rights Policy, and trusted experts to make these judgments.[23]

### 3.2 Complaints and Appeals

As with any set of complex systems and processes, we recognise that it is not possible to always get it right. Our complaints and appeals mechanisms have long been in place and made available to reporters of content and users who are affected by decisions. For example, in the first quarter of 2024, out of 39.4 million pieces of content related to Adult Nudity and Sexual Activity we took action on globally, we received appeals for 2.5 million pieces of content of which 517,000 pieces of content were later restored.[24] If we change our decision, we'll let the reporter know, and we will implement our revised decision.[25]

If we have reviewed an appeal and the user still does not agree with our decision, they may be able to appeal to the Oversight Board.[26] The Oversight Board independently reviews some of the most difficult and significant content decisions we make across our global operations. Once reviewed, they inform us whether

[20] https://transparency.meta.com/policies/improving/stakeholders-help-us-develop-community-standards/
[21] https://about.meta.com/actions/promoting-safety-and-expression/
[22] https://transparency.meta.com/en-gb/policies/improving/bringing-local-context
[23] https://transparency.fb.com/en-gb/features/approach-to-newsworthy-content/
[24] https://transparency.meta.com/reports/community-standards-enforcement/adult-nudity-and-sexual-activity/facebook/
[25] https://www.facebook.com/help/2090856331203011/
[26] https://www.facebook.com/help/711867306096893

or not they agree with our content decisions. The board's decisions are binding and if they do not agree with our initial decision, we'll reverse it, unless doing so could violate the law. In the fourth quarter of 2023, Meta completed work on 15 recommendations made by the Oversight Board, bringing our annual total to 61 recommendations out of 122 recommendations completed by the end of 2023. The recommendations we undertook in 2023 spanned our operations, policies, and services, contributing to broad and meaningful improvements across the company and our global community.[27] As out-of-court dispute settlement bodies become certified and provide evidence of certification under Article 21 of the DSA, we will also take steps to engage in this process.

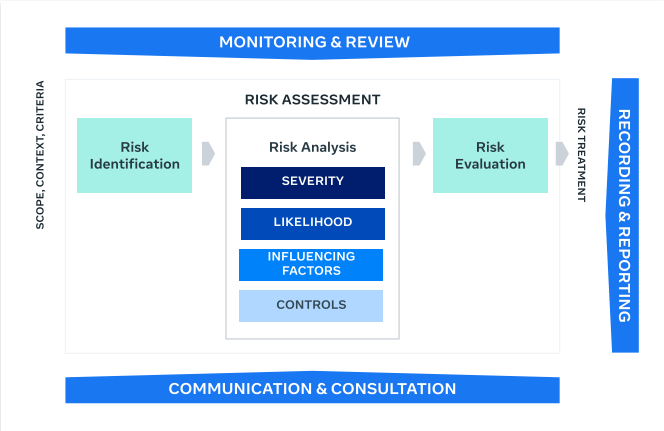
[27] https://transparency.meta.com/oversight/meta-quarterly-updates-on-the-oversight-board/

# 4. Meta's DSA Systemic Risk Assessment Methodology

Meta recognises the importance of identifying and addressing risks posed to our community of users and more widely. We have rigorous structures and processes in place to understand, identify, manage, and mitigate risks that surface on our platforms and within our underlying technology systems and processes.

Meta continuously evolves its practices to respond to the evolving risk landscape and regulatory environment. Meta's ISSO-GRC Programme and DSA Compliance Office provide ongoing risk governance and oversight of Meta's services, systems, and processes. One critical risk management capability within the ISSO-GRC Programme is the **Integrity Risk Management Process (Figure 1),** which was designed based on industry standards, specifically ISO 31000: Risk Management.

**Figure 1. Integrity Risk Management Process**




### 4.1 Risk Assessment Process

As part of Meta's Integrity Risk Management Framework, Meta has a Risk Assessment Process and Methodology designed to enable Meta to operationalise risk assessments for multiple integrity Problem Areas in a standardised and scalable manner. **The Risk Assessment Process (Figure 2)** detailed below explains the steps to conduct risk assessments at Meta, including the assessment of systemic risks that can materialise on Meta's services.

As part of our efforts to continuously evolve our Risk Assessment Methodology, we have made two key enhancements to our methodology in the past year, which includes the following:

- **Operational Effectiveness:** Along with control design effectiveness, we have incorporated operational effectiveness into our Control Suite Effectiveness calculation. Control operational effectiveness refers to the assessment of whether the design of a control is executed consistently in order to address the risk it was assigned to mitigate over a period of time. To enable this, we have expanded our signal base to include input from assurance testing, issue management, and other integrity data. Adding in operational effectiveness and increasing our signal base has provided more insight into our risk and control environment, which has impacted our residual risk scores. More information on our Year-Over-Year Results Comparison can be found in Section 6.1.4.

- **Order of our Process:** We have switched the 'Respond and Mitigate' Phase to come before the Report Phase to enable Meta to provide a combined risk assessment and mitigation report (see **Figure 2** below).

**Figure 2. Risk Assessment Process**


Our Risk Assessment Process consists of six steps and is used consistently to execute risk assessments, including our annual Systemic Risk Assessment. Outlined below is an overview of how this process was executed to carry out the annual Systemic Risk Assessment:

- **Identify and Qualify:** Meta leveraged a diverse set of signals and inputs to scope the risk assessment. These inputs were used to define the in-scope Problem Areas and the associated risks that collectively create a systemic risk to users in the EU.
- **Assess**: Meta issued surveys and conducted a series of interviews and workshops (50+) engaging with over 250 internal stakeholders to understand whether and how the overall risk landscape, including current and emerging risks and the control environment, changed over the last year using a standardised risk assessment framework.
- **Measure**: Using the results from the workshops and other signals, Meta finalised the list of relevant in-scope risks and calculated the inherent risk and effectiveness of the controls in place to mitigate the risks using Meta's Risk Measurement Framework. See Appendix 9.1 for more information on Meta's risk and control measurement approach.
- **Validate**: Meta documented the results and engaged with stakeholders to validate the findings.
- **Respond and Mitigate**: In conjunction with inputs from other risk management efforts, Meta worked cross-functionally (and does so on a routine basis) to determine mitigation priorities and determine what is reasonable, proportionate and effective to reduce residual risk on Facebook. See Appendix 9.2 for the Reasonable, Proportionate, and Effective Mitigation Principles.
- **Report**: Meta documented the findings and results in a detailed report.

### 4.2 External Stakeholder Engagement

Meta regularly engages with external stakeholders to gather input, knowledge, and insight on how integrity risks can manifest on social media services and the associated ramifications to users and society. For example, our approach to labelling AI-generated content and manipulated media is based on feedback from the Oversight Board and from consultations with over 120 stakeholders in 34 countries in every major region of the world.[28] Additionally, through our Policy Forum, we sought out input and looked at research from different perspectives, to assess our approach to use of the word "Zionist" under our Hate Speech Policy. As a result of this work, we now remove speech targeting "Zionists" in several areas where our process showed that the speech tends to be used to refer to Jews and Israelis with dehumanising comparisons, calls for harm, or denials of existence. For this revision, in total, we consulted with 145 stakeholders representing civil society and academia across the Middle East and Africa, Israel, North America, Europe, Latin America and

[28] Our Approach to Labelling AI-Generated Content and Manipulated Media | Meta (fb.com)

Asia, including political scientists, historians, legal scholars, digital and civil rights groups, freedom of expression advocates and human rights experts.[29] Additionally, Meta regularly surveys its users to understand what they perceive to be the negative experiences they most commonly encounter. This information helps inform the design, identification, evaluation, and scoring of risks and controls in risk assessments required under the DSA, as well as the prioritisation of risk management activities.

Furthermore, our internal stakeholder groups, namely the Content Policy, Human Rights, Civil Rights, and Social Impact User Experience Research (UXR) teams, were able to provide insights from third party consultations and co-design activities on the systemic risks through several existing initiatives Meta has underway. For example, Meta attended a European Rights and Risks: Stakeholder Engagement Forum, organised by the Digital Trust and Safety Partnership and the GNI on assessing systemic risks to fundamental rights as part of implementing the DSA. These types of engagements help inform the in-scope risks for the assessment, the challenges in managing these risks, the impact of these risks, and how they are managed. More information on how Meta engages with external stakeholders and the third parties consulted can be found here.

### 4.3 Emerging and Unknown Risks

Whilst the Risk Assessment Process is mainly focused on identifying and assessing known risks, we have established processes in place to assist us in identifying emerging risks and getting signals on unknown risks. These processes include, but are not limited to, our threat intelligence capabilities, engagement with external research institutions, advocacy groups and law enforcement, and industry information sharing partnerships where we share and ingest information on emerging risks and trends.

[29] https://transparency.meta.com/en-gb/hate-speech-update-july2024/)

# 5. Systemic Risk Landscape

A core part of our longstanding commitment to online safety is an in-depth understanding of potential Problem Areas that could arise on our platforms. Our policies, teams, systems, and processes are organised around these Problem Areas and we have dedicated internal experts and targeted approaches to addressing each of these Problem Areas.

The Systemic Risk Landscape depicts Problem Areas either mentioned in Article 34 of the DSA or understood by Meta to impact potential systemic risk in the EU. We used our deep knowledge of these Problem Areas and their associated potential risks to assess the DSA Systemic Risk Areas in Article 34 and define Facebook's Systemic Risk Landscape.

A visual representation of Meta's Systemic Risk Landscape and the Problem Areas aligned to each Systemic Risk Area is detailed in **Figure 3**. This landscape is meant to depict the most common mapping(s) between Problem Areas and the DSA Systemic Risk Areas based on explicit citations within the DSA, with the exception of Illegal Content. There are circumstances in which risks associated with the Problem Areas below could map to other Systemic Risk Areas.

**Figure 3. Meta's Systemic Risk Landscape**


**Our approach to Illegal Content**

Our risk landscape graphic above highlights a number of Problem Areas that some regulatory regimes and legal frameworks might deem as illegal content at a national or supranational level, but you will notice that they are not mapped to the Illegal Content Systemic Risk Area (per Art 34(1)(a) of the DSA). Our globally applicable Community Standards outline types of content or behaviour that are not allowed on Facebook. In addition, our Integrity Ecosystem supports and enforces these standards. In many cases, our Community Standards do indeed overlap with common areas of illegality (e.g., child exploitation), but they do not map to

specific laws, as laws vary significantly across countries and have many nuances. In many cases when we enforce against content, Pages, Groups or accounts, for example, for violating our policies, the content may also be illegal. Our policies also cover Problem Areas that would not commonly be considered to be illegal (e.g., bullying).

In addition to reporting options for content that might violate our Community Standards in the EU, we have had dedicated reporting tools for illegal content easily accessible from the relevant content and in our Help Centre.[30] We may receive court orders to restrict content on Facebook or reports from governments, regulators, as well from non-government entities and members of the public alleging content is unlawful. We review these requests in line with Article 16 of the DSA and our Corporate Human Rights Policy as well as with our commitments as a member of the GNI. For example, in our DSA Transparency Report for Facebook covering 1 October 2023 to 31 March 2024, we reported 2,089 Authority Orders to act against illegal content (including Article 9 orders) addressed to Meta. In that same Transparency Report, we reported 601,863 notices submitted in accordance with Article 16 of which 126,247 (or ~21%) led to content removal or restriction.[31]

**Our approach to Fundamental Rights**

Meta has well established and enforceable policies for each Problem Area mapped above with the exception of those solely mapped to Fundamental Rights. Those solely mapped to Fundamental Rights, such as "Voice and Free Expression" or "non-discrimination" are embedded and accounted for within all of our policies as we take a holistic approach to fundamental rights across all of our Problem Areas, including the protection of marginalised communities. Additionally, human rights inform and shape our tooling, technology development, and content moderation at scale.

**Our approach to Physical and Mental Well-Being**

Physical and Mental Well-Being is not listed as its own Systemic Risk Area in our Systemic Risk Landscape because it cuts across multiple Problem Areas and DSA Systemic Risk Areas. However, a number of the Systemic Risk Areas helped inform our methodology, including the *Scale, Nature of Impact* and *Impacted Demographic* categories in our Severity Rubric. As a result, the systemic risk of physical and mental well-being was directly incorporated into the Severity Rubric under the *Nature of Impact* category as "physical and psychological impact." Accordingly, elevated risk scores (Tier 4 out of 5) were assigned to risks deemed to have a potentially elevated impact on an individual's physical and psychological well-being. Given that physical and mental well-being was incorporated into the Risk Severity Rubric used to assess all in-scope risks, regardless of Problem Area or Systemic Risk mapping, the risk assessment considered the potential impacts on physical and mental well-being when assessing the severity of risks across all Problem Areas, and by extension all Systemic Risk Areas.

Furthermore, we identified specific risks related to physical and mental well-being in a number of Problem Areas. For example, under the Problem Areas of "Misinformation" and "Disinformation", we considered the risks of "Harmful Health Disinformation" and "Harmful Health Misinformation"; under the Problem Area of "Restricted Goods and Services", we considered the risks of "Alcohol, Tobacco, Prescription Products, Drugs, and Drug Paraphernalia" and "Medical and Healthcare Products"; and under the Problem Area of "Suicide and Self-Injury", we considered the risks of "Disordered Eating", "Mental Health", "Personal Health and Appearance Ads" and "Suicide and Self-Injury."

[30] https://transparency.fb.com/reports/content-restrictions/content-violating-local-law/
[31] https://transparency.meta.com/sr/dsa-transparency-report-apr2024-facebook

See Appendix 9.1 for more information on the Severity Rubrics which encompass well-being impacts and Section 6.2.2.17 for information on how we support users with digital wellness.

## 5.1 Systemic Risk Areas

The following sections detail how Meta has interpreted and approaches managing the DSA Systemic Risk Areas for Facebook including with respect to the EU. Please see Section 6.2 of this Report for an overview of each of the Problem Areas and risks associated with each DSA Systemic Risk Area.

### 5.1.1 Deceptive and Misleading

With the advent of new technologies and as the world becomes increasingly interconnected, threat actors find new ways and more vulnerable people that they can target with evolving deceptive and misleading tactics that may be fraudulent or seek to exploit others for money or property. Our goal is to detect and counter them, whilst updating our defences as adversarial actors change their behaviour.

Meta takes a multifaceted approach to reducing inauthentic behaviour and associated content. In line with our commitment to authenticity, we prohibit people misrepresenting themselves on Facebook, using fake accounts, artificially boosting the popularity of content or engaging in behaviours designed to enable other violations under our Community Standards. We have invested in a number of measures to address this risk, including on the product design front. For example, our deceptive identity model uses a variety of signals, such as information about account activity, frequency and type of posts, likes, comments, account profile information, name, profile picture, and bio to identify Facebook accounts that are likely to be fake or deceptive. Additionally, we built the largest fact-checking programme across the industry in Europe, with 29 partners across the EU covering 23 languages and further adding 3 new partners in Bulgaria, France, and Slovakia in 2024.[32] When a fact-checked label is placed on a post, 95% of people do not click through to view it.[33] We also have support resources, such as the Anti-Scam Hub, Scam Safety Centre, and our Media Literacy Campaigns. When we learn of coordinated inauthentic behaviour, we make considerable efforts to take down each known adversarial network of coordinated accounts and Pages as a whole, rather than removing them piecemeal. This makes it harder for malicious groups to come back and target people who use our apps.

The risk landscape can be subject to change based on internal and external factors, which is something we evaluated as part of this assessment. For our assessment of Deceptive and Misleading risks in 2024, we identified that the risk landscape could be impacted by the high number of elections in the EU, including the EU Parliamentary Elections, conflicts in adjacent regions, and the increasing adoption of generative AI technology. Meta has put in compensating controls to help manage this increased risk exposure, as detailed in Section 6.2.

### 5.1.2 Civic Discourse and Elections

Our approach to civic discourse and elections focuses on trying to prevent interference, increase transparency, prevent the spread of misinformation or disinformation, and empower people to vote. Over many years, Meta has invested in a comprehensive approach to managing risks related to elections on our platforms, not just during election periods but at all times. We continually review and update our election-related policies and take action if content violates our Community Standards. We use keyword

[32] https://www.facebook.com/business/help/997484867366026?id=673052479947730
[33] https://about.fb.com/news/2024/02/how-meta-is-preparing-for-the-eus-2024-parliament-elections/

detection to group content related to EU elections in one place to make it easier for fact-checkers to find. Additionally, our teams fight both foreign interference and domestic influence operations, and have taken down more than 200 malicious influence campaigns globally involved in what we call Coordinated Inauthentic Behaviour, something we publicly share as part of our Quarterly Threat Reports. We have also designated more than 700 hate groups around the world. Designated hate groups are not allowed to have a presence on our platforms and we remove glorification, support, and representation of these entities, their leaders, founders or prominent members, as well as unclear references to them. We assess these entities based on their behaviour both online and offline, and most significantly, their ties to violence. We continue to identify and assess new hate groups, particularly when they are tied to real-world violence.[34]

By the end of 2024, more than two billion people will head to the polls across some of the world's biggest democracies, including the EU.[35] As a result, Meta increased its investment in election-specific mitigation measures, including activating a dedicated team to develop a tailored approach to help prepare for the EU Parliamentary Elections. As a result, we have implemented the following election-specific risk mitigation measures:

- **EU-specific Elections Operations Centre:** We stood up an EU-specific Election Integrity Team dedicated to the EU Elections preparation work, bringing together experts from across the company from our intelligence, data science, engineering, research, operations, content policy and legal teams. These teams have been working together for more than a year to identify potential threats and put specific mitigations in place across our apps and technologies and ensure our elections readiness. Additionally, in the lead up to the election, we activated an EU-specific Elections Operations Centre to bring all these teams together in person and respond in real time to any new risks or time sensitive escalations.
- **External Engagement:** For the June 2024 EU Parliamentary Elections, Meta conducted outreach across the 27 member states, informing them of our approach to the elections and establishing a communication channel with national authorities. We proceeded to temporarily onboard national election authorities as well as other competent bodies to a dedicated reporting channel, allowing them to directly report content that may violate our policies or election laws for prompt review, and delivered a training session on this channel and on our elections-related policies. We have also conducted outreach and comprehensive training to formally appointed Digital Service Coordinators to help them navigate the "Single Point of Contact" Form for EU member states' authorities, the EU Commission, the EU Board for Digital Services, as well as the onboarding process, where required, in order to access the relevant contact forms. We also conducted training on paid and organic campaigning to EU Members of Parliaments, and to political parties at the member state level, and launched an EU Election Hub in all 24 EU official languages to support all our government partners.
- **Transparency:** We included a number of measures to enable transparency, such as requiring all political advertisers to verify their identity before buying ads; mandating "Paid for by" disclaimers to political and issue ads; maintaining an Ads Library for users to search through all political and social issue ads from the last seven years; and launching in-app Voter Information Unit and Election Day Reminders, where legally permitted, on both Facebook and Instagram on relevant election periods, reminding people of the day they can vote, and redirecting them to local authoritative sources on how and where to vote.

[34] https://about.fb.com/news/2023/11/how-meta-is-planning-for-elections-in-2024/
[35] https://about.fb.com/news/2023/11/how-meta-is-planning-for-elections-in-2024/

- **Rapid Response:** Meta is an active member of the EU Disinformation Code Taskforce's Election Working Group and is taking part in its newly formed Rapid Alert System. To this end, Meta set up both an email alias to flag trends and a standardised form to report content which poses serious or systemic concerns to the integrity of the electoral process and ensure its prompt review. We additionally attended weekly meetings with all the signatories of the EU Disinformation Code to discuss any new trends observed and provide feedback on any reports received. During the implementation period, we received five reports, all of which were reviewed, discussed with the Working Group, and closed. We will look to provide further insight on the Rapid Alert System in our upcoming EU Disinformation Code Report.

- **Proactive Measures:** We also invested in proactive threat detection and have expanded our policies to help address harassment against election officials and poll workers. Meta also sent people that face increased levels of election risk-related in-Feed notifications on Facebook and Instagram on how to protect themselves and their accounts, such as accounts from candidates that ran in the EU Parliamentary Elections.

- **Awareness of Generative AI Images and Media:** We introduced labelling of images that users post on our platforms when we can detect that they are AI-generated. We also added a feature for people to disclose when they share AI-generated video or audio so we can add a label to it and we may apply penalties if users fail to do so.[36] Additionally, AI-generated content is also eligible to be reviewed and rated by our independent fact-checking partners. In January 2024, we also announced that advertisers need to disclose whenever an ad about social issues, elections or politics contains a photorealistic image, video or realistic sounding audio that was digitally created or altered and such alteration is material. Failure to disclose digitally created or altered media could result in the ad being removed and penalties being applied on the account.[37]

- **Fact-checking Network:** We have invested in our fact-checking network, including conducting an online refresher training session on our policies with the fact-checking network, giving particular focus to our election-related policies and our approach to AI-generated content. We have started accepting European Fact-Checking Standards Network (EFCSN) certification as a prerequisite for consideration in the Meta fact-checking programme in Europe, in recognition of the strong standards it has established for the European fact-checking community. Additionally, we continuously engage with our fact-checkers as a means to exchange information.[38] For example, we met in person with fact-checkers in the lead up to the elections in Brussels and Warsaw to discuss misinformation trends and educate them about Meta's approach to the election. Additionally, Meta worked with the EFCSN to help improve the skills and capabilities of the European fact-checking community in debunking and countering AI-generated misinformation, facilitate common standards in addressing and fact-checking AI content, and raise the public's awareness on this type of misinformation through media literacy campaigns. This included a series of five workshops with experts giving training to over 200 individual fact-checkers across Europe. The media literacy campaign was published in 27 different languages across Europe.[39]

[36] https://about.fb.com/news/2024/02/labeling-ai-generated-images-on-facebook-instagram-and-threads/
[37] https://www.facebook.com/business/help/1486382031937045
[38] https://about.fb.com/news/2024/02/how-meta-is-preparing-for-the-eus-2024-parliament-elections/
[39] https://efcsn.com/news/2024-04-18_efcsns-new-project-for-identifying-ai-generated-and-digitally-altered-content/

The risk landscape can be subject to change based on internal and external factors, which is something we evaluate as part of this assessment. For our assessment of Civic Discourse and Election risks in 2024, we identified that the risk landscape could be impacted by several trends, including, but not limited to, the increasing adoption of generative AI technology, conflicts in adjacent regions, the high number of elections in the EU, including the EU Parliamentary Elections, and the assassination attempt on the Slovakian Prime Minister. Meta has put in compensating controls to help manage this increased risk exposure, as detailed in Section 6.2.

### 5.1.3 Public Health

Meta is committed to fostering an environment that supports public health globally, including within the EU, and strives to empower users by increasing access to credible health information, enabling people with similar health issues to connect with one another, and empowering them to make informed decisions about their health and well-being.

We have targeted measures to identify and manage public health risk on the platform. For example, we deploy interstitials to users searching for restricted goods and services on the platform to warn users and share resources to learn more information. For minors, Meta applies age gating restrictions to content related to diet products, cosmetic procedures, real money gambling, alcohol, and tobacco among others and leverages age enforcement infrastructure to reduce visibility of this type of content for minors. Additionally, over the last year, we have invested in managing public health on our platforms, including improving our classifiers and recommendations filtering related to suicide and self-injury and using AI tools to scale fact-checkers' work to detect false and misleading health information.

Although public health risks can potentially arise from the use of Facebook, the platform is also a vector to enable people to seek help and support through our resources, including our Crisis Support Resources, our Bullying and Harassment Safety Centre Resources, our Suicide Prevention Resources, our Emotional Health Hub, and our Family Digital Wellness Guides.

The risk landscape can be subject to change based on internal and external factors, which is something we evaluate as part of this assessment. For our assessment of Public Health risks in 2024, we identified that the risk landscape could be impacted by the increasing adoption of generative AI technology and new trends around the use of restricted goods and services, such as alcohol, tobacco, prescription products and drugs, and healthcare products. Meta has put in compensating controls to help manage this increased risk exposure, as detailed in Section 6.2.

### 5.1.4 Public Security

Meta is dedicated to securing our services and negating potential public security risks that could arise through the use of Facebook. We know that no single company can solve various global threats to public security alone.

We have targeted measures to identify and enforce against dangerous or violent actors on our platforms, including enhancing our automated detection technology and using intelligence to identify and remove actors and objects that are connected from a network with our account enforcement propagation efforts. We also leverage detection tools to highlight early warning signs of threats against public security, such as accounts with multiple recent strikes or an increase in violent content to craft proactive and reactive mitigations in response to these signals. We also monitor regions that may be the target of violence using our Temporary High-Risk Location (THRL) list. We maintain a Market-specific Implicit Threat Terms List which enables our classifiers to more effectively detect violent content on our platforms. We use keyword

interstitials to help prevent users from viewing problematic or violent content and share support resources. If we become aware of content on our platform relating to a credible threat of real world harm, we do not hesitate to notify the applicable authorities and provide relevant information in accordance with our Terms of Service and applicable laws. For example, shortly after the assassination attempt on Slovakia's Prime Minister, we took down the Facebook account of the alleged shooter, classified the incident as a violation of our Dangerous Organisations and Individuals Community Standards, and notified law enforcement in line with our crisis response approach.

Additionally, Meta invests in countering the misuse of our services by authoritarian governments, terrorist groups, or other threat actors who may attempt to surveil regime critics, opposition figures, and Human Rights Defenders (HRDs).[40] When we detect such activity on our platforms, we seek to block their domain infrastructure from being shared on our services and notify people who we believe were targeted by these malicious operations in accordance with our Terms of Service and applicable laws.

At times, we also share findings about threats we detect with our industry peers and security researchers to help our entire community better understand and counter internet-wide challenges, including threats to fundamental societal interests (e.g., threats to the functioning of essential public services or institutions and military interests) and large-scale threats to life (e.g., armed conflicts and acts of terrorism). Additionally, we have a robust process for reviewing and prioritising countries with the highest risk of offline harm and violence every six months.[41] To help support users who may be experiencing public security concerns on our platforms, Meta has developed safety tools, such as Crisis Support Resources in our Safety Centre, where users can get urgent expert local support.

The risk landscape can be subject to change based on internal and external factors, which is something we evaluate as part of this assessment. For our assessment of Public Security risks in 2024, we identified that the risk landscape could be impacted by the increasing conflict in adjacent regions, the high number of elections in the EU, including the EU Parliamentary Elections, and the increasing adoption of generative AI technology. Meta has put in compensating controls to help manage this increased risk exposure, as detailed in Section 6.2.

#### 5.1.5 Gender-Based Violence

Meta believes that women and people of all gender identities and expressions deserve equal access to the economic, educational, and social opportunities the internet provides. Facebook and other social networks enable marginalised communities to connect and often provide them with a platform to use their voice for change. However, we know that there are some challenges that can be faced in how these communities engage on Facebook, including for our lesbian, gay, bisexual, transgender, queer, intersex, and asexual (LGBTQIA+) community. We have worked to strengthen our relationships with the LGBTQIA+ community by increasing engagements with groups and representatives across the world and in the EU, on the impact of Meta's content policies on users, particularly regarding hate speech, bullying, and harassment. These included LGBTQIA+ HRDs, civil society organisations, academic scholars, and activists.[42]

[40] https://humanrights.fb.com/wp-content/uploads/2023/09/2022-Meta-Human-Rights-Report.pdf
[41] https://about.fb.com/news/2021/10/approach-to-countries-at-risk/
[42] https://humanrights.fb.com/wp-content/uploads/2023/09/2022-Meta-Human-Rights-Report.pdf

Our commitment to women's safety is longstanding. Over the years, we have sought the help of experts in the field to ensure our platforms are safe for women. Our five-pillar approach works to keep abuse off our platforms:[43]

1. **Policies:** We have developed strong policies to help protect women from online abuse, including rules against behaviours that disproportionately impact women, such as the sharing of non-consensual intimate imagery and rules against harassment. These policies have been developed in partnership with Meta's Global Women's Safety Expert Advisors;[44]
2. **Tools:** We have built tools to empower users to control their experience online, protect themselves against unwanted content and contact, and report violations. We have also launched technology to combat the sharing of non-consensual intimate images and made significant investments against sextortion on our platforms;
3. **Resources:** We offer 24/7 access to resources designed with the safety of women in mind. The Facebook Help Centre provides step-by-step guides to protect users against threatening or unsafe content. We also have our Women's Safety Hub as a centralised location for our online safety resources specifically geared towards women;
4. **Expert Engagement:** After working with over 400 women's safety organisations and experts across the world, we established a specific group of advisors that serve on our Global Women's Safety Expert Advisors Group. This Group includes security experts, academics, non-governmental organisations (NGOs), human rights activists and policymakers who provide their guidance in how we build our policies, tools and resources. These experts have contributed to furthering the safety of women both online and off and are distinguished in the field; and
5. **Community Engagement:** Facebook provides a shared community and knows user experience matters. We gather input from users to develop the policies, tools and resources that promote women's safety online.

The risk landscape can be subject to change based on internal and external factors, which is something we evaluate as part of this assessment. For this Systemic Risk Area in 2024, it was identified that the risk landscape could be impacted by global events, such as the preparation for the Olympic Games in Paris which involves more movement of people and could increase the risk of human exploitation, including human smuggling. Additionally, we see evolving trends that could impact the risk of violence against targeted or vulnerable communities, such as women and minors, the high number of elections in the EU, including the EU Parliamentary Elections, and the increase in anti-refugee sentiment. Meta has put in compensating controls to help manage this increased risk exposure, as detailed in Section 6.2.

### 5.1.6 Protection of Minors

At Meta, child protection is always a top priority. Among other risks, preventing child exploitation is one of the most important challenges facing our industry today. We have an extensive child protection and well-being ecosystem that works to safeguard minors on our platform, which ranges from enhanced default settings for minors, industry collaboration across various youth-focused initiatives, advanced and continually evolving technological interventions, safety tools to provide options to minors and their parents on what they can see and who can contact them, features that let minors manage their time on-platform and set reminders, help

[43] https://about.meta.com/actions/safety/audiences/women/
[44] https://about.meta.com/actions/safety/audiences/women/#partners

prevent unwanted interactions, among many other things. To implement this, we leverage a three-pronged, industry leading approach to protecting minors online;[45]

- **Detection:** First, we strive to find ways to help prevent harm from happening in the first place by developing preventative tools and resources, such as our parental controls and guides, age prediction and gating models, default privacy settings, and teen settings shortcuts. We also invest in technological interventions and deploy tools targeted at proactively finding, removing, or reducing policy-violating content, Groups, and Pages among other things, that violate our policies or may be problematic. For example, we deployed a search intervention aimed at reducing malicious searches for content. Additionally, we developed technology to identify potentially suspicious adults by reviewing more than 60 different signals, such as if a teen blocks or reports an adult, or if someone repeatedly searches for terms that may suggest suspicious behaviour.[46] Once we identify potentially suspicious adults on Facebook, we work to prevent them from discovering and connecting with teen accounts. Similarly, we improved our proactive detection of potentially suspicious Facebook Groups and updated our protocols and review tooling so our reviewers can remove more violating Groups. We also aim to detect networks of individuals engaging in behaviour that puts minors at risk; for example between 2020 and 2023, our teams disrupted 37 abusive networks globally and removed nearly 200,000 accounts associated with abusive networks.[47]
- **User Reporting:** Then, we encourage users to report potential harms as soon as they can and we respond to take action. Reporting can help prevent child sexual harassment content spreading and help protect children from harm. We have improved the systems we use to prioritise reports for content reviewers. For example, we are using technology designed to find child exploitative imagery to prioritise reports that may contain it.
- **Industry Collaboration:** Lastly, in addition to developing technology to tackle online safety of minors, we hire specialists dedicated to online child safety and we share information with our industry peers and law enforcement, including the National Centre for Missing and Exploited Children (NCMEC), law enforcement, and other industry partners. We have partnered with NCMEC to develop the Take It Down portal to help teens take back control of their intimate imagery. The Take It Down portal helps remove online nude, partially nude or sexually explicit photos and videos of users under the age of 18. We have also partnered with Thorn to update our Stop Sextortion Hub, offering new tips and resources for teens, parents and teachers on how to prevent and handle sextortion.[48] Our Meta EU Youth Privacy Forum established in 2022 continues to convene a broad range of experts from the privacy and safety communities to explore key policy issues relating to the protection of young people online through a multi-disciplinary and multi-faceted lens. We also announced our participation in Lantern, a new programme from the Tech Coalition that enables technology companies to share a variety of signals about accounts and behaviours that violate their child safety policies.[49] Furthermore, we conduct co-design sessions with parents, teens, guardians and experts through the Trust Transparency and Control Labs (TTC Labs), a cross-industry effort to put people in control of their privacy, and work with a number of advisory groups, including our Youth Advisors, to build safe, positive, and age appropriate experiences for teens and their families.

[45] https://about.meta.com/actions/safety/onlinechildprotection/
[46] https://about.fb.com/news/2023/12/combating-online-predators/
[47] https://about.fb.com/news/2023/12/combating-online-predators/
[48] https://about.fb.com/news/2024/02/helping-teens-avoid-sextortion-scams/
[49] https://about.fb.com/news/2023/12/combating-online-predators/

The risk landscape can be subject to change based on internal and external factors, which is something we evaluate as part of this assessment. For our assessment of Protection of Minors risks in 2024, we identified that the risk landscape could be impacted by the increased risk of bullying and harassment towards minors, global events, such as the preparation for the Olympic Games in Paris, and the increasing adoption of generative AI could impact the inherent risk of some of the risks within this Systemic Risk Area. Meta has put in compensating controls to help manage this increased risk exposure, as detailed in Section 6.2.

### 5.1.7 Fundamental Rights

Meta's mission is to give people the power to build community and bring the world closer together. Our services and apps help make it possible for grassroots movements to flourish and challenge established authority and orthodoxy.[50] As new challenges emerge, our human rights work evolves to address such issues and trends, including the increase of anti-semitism and islamophobia in the EU.

We have various processes and policies in place to help enshrine respect for users' fundamental rights, including having our teams review new features, services and policies and weigh in on potential impacts to freedom of expression and other rights. For example, we reviewed our Dangerous Organisations and Individuals Community Standards and refined it to allow more social and political discourse, including about elections, conflict resolution, and disaster and humanitarian relief. This and other policy developments are assessed with human and civil rights analysis as a key part of decision-making.

We manage our human rights work by training our Meta employees to have a human rights mindset in their day-to-day work and encouraging respect for human rights to the benefit of all who use our services. The majority of full-time employees are also required to take Meta's Civil Rights and Technology training. This training helps employees understand civil rights concepts and principles, how to identify issues and concerns in their work, and where to go for help with issues or questions. The Civil Rights Team enhances this training module as needed and also engages in internal workshops and analysis to help teams build with civil rights in mind. We also provide users with pathways to report concerns and appeal decisions made about their content and have enabled our Oversight Board to make fully independent content moderation decisions and recommendations about content policy, services, and operations. We stay committed to the GNI Principles of Freedom Expression and Privacy through various actions, such as undergoing an independent assessment of our implementation of the GNI principles every two to three years. Engagement with external stakeholders around the world helps us live up to our human rights responsibilities, creating an important lever for accountability and transparency, and strengthens our work. We recognise the importance of meaningful engagement with stakeholders from marginalised communities.[51] Additional information on how we approach human rights is outlined in Section 3: A Balancing Act: Respecting Rights and Mitigating Risk.

Meta recognises that the other Systemic Risk Areas identified in the DSA, such as Civic Discourse and Elections, Public Security, Public Health, Gender-based Violence, Illegal Content, or Protection of Minors can all have potential fundamental rights implications. Furthermore, the risk landscape can be subject to change based on internal and external factors, which is something we evaluate as part of this assessment. For our assessment of Fundamental Rights risks in 2024, we identified that the risk landscape could be impacted by the increasing conflict in adjacent regions, the high number of elections in the EU, including the EU Parliamentary Elections, which could increase the risk of discriminatory civic-related ads about social issues, elections, or politics; and impact of generative AI. Global events, such as the preparation for the Olympic

[50] https://humanrights.fb.com/wp-content/uploads/2023/09/2022-Meta-Human-Rights-Report.pdf
[51] https://humanrights.fb.com/wp-content/uploads/2023/09/2022-Meta-Human-Rights-Report.pdf

Games in Paris which involves more movement of people, could have an impact on fundamental rights. Meta has put in compensating controls to help manage this increased risk exposure, as detailed in Section 6.2.

#### 5.1.8 Illegal Content

We have a robust process for reviewing reports alleging that content on Facebook goes against local law. When we receive a report alleging that content is illegal under EU law, we first review it against the Facebook Community Standards. If we determine that the content goes against our policies, unless deemed newsworthy, we remove it. The Community Standards do address many areas of harm that in practice overlap with concepts of illegality (e.g., hate speech, child exploitation), as mentioned above in Section 5 and are applied globally.[52] Specifically related to Intellectual Property (IP) Infringement, rights holders can report different types of content they identify on our platforms, including individual posts, photos, videos or advertisements to an entire profile, account, Page, Group or event, which is then processed by our IP operations team.

In addition to reporting tools for content that may go against our Community Standards, we have dedicated, user-friendly reporting available for content alleged to be illegal, in keeping with Article 16 of the DSA. In line with Article 21 of the DSA, we will provide the ability for users across EU member states to refer a decision, which may include illegal content, to a designated Out-of-Court Dispute Settlement Body. We also have mechanisms to notify law enforcement of suspected criminal offences involving a threat to life or safety. Partnering with law enforcement agencies and being responsive to their requests is critical to our integrity efforts. In 2023 alone, we responded to over 116,000 government requests in the EU across our platforms, including Facebook.[53]

When something on Facebook is reported to us as violating local law within the EU but doesn't go against our Community Standards, such as blackout periods during elections, we may restrict the content's availability in the country where it is alleged to be unlawful. We undertake such analysis in line with our commitments in our Corporate Human Rights Policy and as a member of the GNI.

The risk landscape can be subject to change based on internal and external factors, which is something we evaluate as part of this assessment. For our assessment of Illegal Content Risks in 2024, we did not identify any new trends that could impact the risk landscape.

### 5.2 Influencing Factors

As detailed in **Figure 3**, we evaluated the impact of our Integrity Ecosystem and operations on the Systemic Risk Landscape and the associated risks. This section provides a thematic overview of the role of each Influencing Factor and the impact each can have on the systemic risks. **As part of this risk assessment, we evaluated the potential impact these Influencing Factors could have on each of the Problem Areas and in turn Systemic Risk Areas.** This section details the overall objective and scope of each factor, circumstances by which it induces or reduces risks, and key insights and learnings. Specific details on how we mitigate these factors using Meta's set of controls are provided in Section 6.2: Mitigating Measures Analysis.

#### 5.2.1 Recommender Systems

Facebook's recommender systems are designed to help the millions of people who use our services discover content that they will hopefully find useful, interesting, relevant, and valuable.[54] To determine what content is

[52] https://transparency.meta.com/reports/content-restrictions/content-violating-local-law/
[53] https://transparency.fb.com/reports/government-data-requests/data-types/
[54] https://ai.facebook.com/blog/how-ai-powers-experiences-facebook-instagram-system-cards/

eligible to appear in recommendations, we have Recommendation Guidelines.[55] These guidelines are published in the Facebook Help Centre to help people better understand the kinds of content we recommend, and provide context on why some types of content are not included in recommendations, and therefore may not be distributed as widely.[56]

Furthermore, our recommender systems are powered by multiple AI systems that work separately and, in some cases, together to identify content and predict how likely a person is to be interested in it or interact with it.[57] Our recommender systems are designed to try to prevent the recommendation, recirculation, or amplification of potentially policy-violating or otherwise problematic content. Our systems typically first produce an inventory of available content and then filter it to remove content that potentially violates our policies and standards and content that is not eligible for recommendation.[58] After this filtering, the inventory is pared down further to the items users are most likely to be interested in. Additionally, using our repository of signals, our algorithms may down-rank content for a variety of reasons and do not recommend certain accounts, or content, such as those related to repeat policy violators in feeds.[59] Furthermore, we want our users to feel empowered, so we provide information on how content is recommended to them and options to control or customise their experiences on Facebook. As part of our efforts towards DSA compliance and our ongoing commitment to transparency, we now have 15 recommender 'System Cards'. More information on how our recommender systems work can be found here.

While recommender systems are a core feature of our platform and we have an extensive ecosystem of controls in place to manage them, some potentially problematic content that does not violate our guidelines may be recommended before we can detect and remove it, and some violating accounts may evade our detection technologies and be recommended. This influencing factor could have an impact on all Systemic Risk Areas. This insight was derived through our risk assessment process where we evaluated recommender systems in the following manner:

- **Assess Phase:** We conducted workshops with experts from our Content Policy, Cross-Integrity Team, Global Operations, and Legal within expertise in recommender systems and asked a series of questions for each Problem Area to understand how recommender systems may influence Problem Areas and the associated risks. The questions focused on understanding the following types of matters:
    - Content ranking and recommender systems;
    - Potential amplification and distribution of policy-violating content; and
    - Potential limitations of mechanisms in place to manage recommender systems.
- **Measure Phase:** For each Problem Area, we evaluated the responses to the questions; took into consideration any issues and improvement areas identified through various signals, including assurance testing, where applicable; analysed data, where available; and evaluated compensating controls in place to address any limitations, such as our classifiers and monitoring controls.

As a result of the "Assess" and "Measure" phases of the assessment, we identified the following areas that should be considered for enhancement related to recommender systems:

[55] https://about.fb.com/news/2020/08/recommendation-guidelines/
[56] https://www.facebook.com/help/1257205004624246
[57] https://ai.facebook.com/blog/how-ai-powers-experiences-facebook-instagram-system-cards/
[58] Our Content Distribution Guidelines describe the types of content we think may either be problematic or low quality, so we reduce, or "demote", its distribution in Feed.
https://transparency.fb.com/features/approach-to-ranking/types-of-content-we-demote
[59] https://transparency.meta.com/features/approach-to-ranking/types-of-content-we-demote/

- **Cross-Functional Integration:** We are still in the process of enhancing our enforcement processes across all content categories, specifically improving our ability to ingest signals from cross-functional teams at scale to help inform enforcement against violating content in a consistent manner across Problem Areas. This is an evolving process as we integrate with more signal sources and improve our database of potentially violating content.
- **Abuse of Hashtag Feature:** Preventing manipulation of hashtags intended to distribute policy violating content remains a challenge, as threat actors continue to seek new ways to circumvent our systems. Meta is continuously working to improve and enhance our capabilities to address mitigations for Facebook.
- **Recommendation Features:** People You May Know (PYMK) and recommended friends features could connect threat actors to minors, which has an impact on risks in the Child Sexual Exploitation, Abuse and Nudity and Human Exploitation Problem Areas. Meta uses tooling to identify potentially suspicious actors and takes appropriate action, including removal from recommendation surfaces. Additionally, we continuously work to improve and enhance our capabilities to address mitigations needed on Facebook.
- **Content Ranking:** We routinely evaluate whether the signals we use to enable users to get relevant content could lead to exposure of problematic content. We reduce this risk by limiting the role of shares and comments in the distribution of sensitive topics.[60] We are continuously working to improve and enhance our capabilities to address mitigations needed on Facebook.

### 5.2.2 Content Moderation Systems

'Content moderation systems' at Meta are referred to as 'integrity systems'. Facebook's integrity systems and processes are designed to detect and review potentially violating content and accounts, including organic, paid content, and commerce. We utilise technology and user reports to identify potentially policy violating content, and use both technology and human review to confirm policy violations and take action on content and accounts that go against our policies. We invest extensively in continuously updating our integrity systems and processes to keep up with new behaviours and trends. Our ecosystem is made up of several components including, but not limited to, detection, enforcement, and appeals. More information on how our integrity systems work can be found here.

Of the violating content we take action on, our technology detects the vast majority of it before anyone reports it. However, like any complex system, there are limitations. Threat actors study how our detection and enforcement classifiers are designed and try to exploit them as quickly as we are able to discover and address these limitations. That's why we also rely on users to report content to us so we can identify and take appropriate action. In addition, our Trusted Partner Reporting Channel provides signals of emerging trends and potential policy violations to inform our proactive detection efforts. We may remove, reduce the distribution of, or inform users of problematic content based on our Community Standards. We also have our Strike System to hold users accountable for continuous violations of the Community Standards. For most violations, the first strike will result in a warning with no further restrictions. If we remove additional posts that go against our Community Standards, we may also apply additional strikes to the user's account, who may lose access to some features for longer periods of time. In addition, our actor recidivism process helps

[60] To help inform users about what they see and read, we include warning screens over potentially sensitive content on Facebook, such as: violent or graphic imagery; posts that contain descriptions of bullying or harassment, if shared to raise awareness; some forms of nudity; and posts related to suicide or suicide attempts.

prevent threat actors from creating new recidivist accounts to engage in continued abusive/violating behaviour.

Our content moderation systems enable policy violating content to be detected and removed systematically, helping to keep users safe online; at the same time, making mistakes in enforcement can impact user voice and freedom of expression on our platform. The latter directly impacts the Fundamental Rights Systemic Risk Area and is a delicate balance we must strike when managing risk and user rights. This insight was derived through the execution of our risk assessment where we evaluated and accounted for content moderation systems as an influencing factor in the following manner:

- **Assess Phase:** We conducted workshops with experts from our Content Policy, Cross-Integrity Team, Global Operations, and Legal and asked a series of questions for each Problem Area to understand how these systems influence Problem Areas and the associated risks. The questions focused on understanding the following types of matters:
    - Tools for managing EU languages, dialects, and cultural nuances;
    - Balancing fundamental rights and safety;
    - Scope and capacity of human reviewers;
    - Monitoring precision and recall; and
    - Potential limitations of content moderation systems.
- **Measure Phase:** For each Problem Area, we evaluated the responses to the questions; took into consideration any issues and improvement areas identified through various signals, including assurance testing, where applicable; analysed data, where available; and evaluated compensating controls in place to address any limitations, such as other classifiers, monitoring controls, user reporting, and appeals processes.

As a result of the “Assess” and “Measure” phases of the assessment, we identified the following areas that should be considered for enhancement related to content moderation systems:

- **Detection of Sporadic Content:** Our automated detection systems are trained and improved when a specific type of content occurs more regularly on our platform. In instances where policy-violating content is posted more sporadically, such as necrophilia, we rely more heavily on user reporting and human review. Meta is continuously working to improve and enhance our capabilities to implement further mitigations on Facebook.
- **Bullying and Harassment Classifiers:** We prohibit bullying and harassment on our service; however bullying and harassment has nuanced definitions because it is personal and can sometimes require context. It's difficult for technology to pick up on these subtleties, limiting Facebook’s ability to proactively capture all forms of bullying and harassment content and behaviour and enforce against it. In addition, there are emerging risks related to how different generations of users, such as Generation Z (Gen Z), use certain words or phrases that our classifiers are not yet able to detect. We are always working to further improve the capabilities of our technology in this space and we also provide and maintain tools and features for users to manage the risk as well, such as blocking and reporting other users and content.
- **Recidivism Prevention:** Recidivism continues to be a challenge that we manage across Problem Areas. To help combat this, Meta has built a classifier that can help detect if a bad actor that has previously been removed from the platform is behind the creation of new accounts in order to take down these accounts. We have also increased investment in enforcement of recidivism related to

sextortion. When we take enforcement against a user for sextortion, our systems aim to identify and block attempts from these users to create new accounts and come back on the platform.

- **Circumvention:** Threat actors continue to explore ways to avoid detection and enforcement by using coded language with emojis and slurs, avoiding certain phrases, or other strategies which can make it challenging for technology to detect potential violations. Meta has proactive teams and mechanisms to identify and subsequently integrate these patterns into the automation detection system.

### 5.2.3 Terms of Service and their Enforcement

Facebook is a global community, so the Facebook Community Standards apply equally to everyone, everywhere and to all types of content. Our Terms of Service and content policies, including the Community Standards, are designed to help define what is and isn't allowed on Facebook and manage systemic risks by providing clear guidelines on our approach to these issues in a way that is easy to understand for users. They form the foundational structure of our Integrity Ecosystem so that we can help keep users safe and maintain a trusted, equitable, and secure environment. We also provide further information on how those policies and other relevant procedures are applied. All of our policies can be accessed through the Help Centre and Transparency Centre, both of which are readily available to users on the platforms, and are available in more than 90 languages, including the official EU languages.

Our foundational policies that determine how we operate Facebook include, but are not limited to, the following:

- Terms of Service (ToS): These terms detail the services we provide, how our services are funded, user commitments to Facebook and our community, additional provisions, and other Terms and Policies applicable to users;
- Community Standards: These standards outline our approach to content that users post to Facebook and user activity on Facebook and other Meta services;
- Advertising Standards: These standards apply to partners who advertise across Meta's services and specify what types of ad content are allowed;
- Commerce Policies: These policies outline the policies that apply when users offer products or services for sale on Facebook, Instagram, and WhatsApp;
- Privacy Policy: This policy details how we collect, use, share, retain and transfer information, along with detailing user rights;
- Corporate Human Rights Policy: This policy commits to respecting human rights as set out in the United Nations Guiding Principles on Business and Human Rights; and
- Code of Conduct: This Code of Conduct defines the expectations we have for how we act and how we make decisions as a company.

As the world changes, so do our policies. We have processes in place to review, maintain, and validate our policies, standards, and terms to reflect the evolving world. Specifically for our Facebook Community Standards, our Content Policy Team includes subject matter experts in issues like hate speech, child safety and terrorism as well as people with experience in criminal prosecution, rape crisis counseling, academics, human and civil rights, law and education. The Content Policy Team consults with internal and external stakeholders from around the globe to discuss potential updates on a routine basis. These updates are communicated to our engineering teams and human review teams who will adjust our detection and enforcement systems. Additionally, our enforcement systems are routinely trained using data sets of human decisions. We review metrics to validate that our precision against our standards is accurate.

Sometimes it is not possible for us to always get the enforcement of our terms right (e.g., over enforcement) which can impact user's fundamental rights like freedom of speech. Like content moderation systems, Terms of Service, Standards, and other policies and their enforcement can have both a positive (e.g., establishing behavioural and content guardrails and enforcing them) and negative influence (e.g., over enforcement). As a result, this influencing factor could have an impact on all Systemic Risk Areas.

The following insights highlighted were derived through the execution of our risk assessment where we evaluated and accounted for Terms of Service, Standards, and other policies and their enforcement as an influencing factor in the following manner:

- **Assess Phase:** We conducted workshops with experts from our Content Policy, Cross-Integrity Team, Global Operations, and Legal and asked a series of questions for each Problem Area to understand how our policies and their enforcement influence Problem Areas and the associated risks. The questions focused on understanding the following types of matters:
    - Maintaining and updating our policies;
    - Managing under and over enforcement; and
    - Potential limitations of our policies and their enforcement.
- **Measure Phase:** For each Problem Area, we evaluated the responses to the questions; took into consideration any issues and improvement areas identified through various signals, including assurance testing, where applicable; analysed data, where available; and evaluated compensating controls in place to address any limitations, such as our policy management change management process, proactive measures taken to address human rights, and our appeals process.

As a result of the "Assess" and "Measure" phases of the assessment, we identified the following areas that should be considered for enhancement related to our Terms of Service, Standards, and other policies and their enforcement:

- **Policy Governance:** The integrity risk landscape is always evolving with the changing internal and external factors and emerging trends. This requires us to continuously review our policies to stay ahead of new trends and adversarial actors. Meta is working to improve our policy governance and policy update processes to maintain tight alignment with other cross-functional teams, provide transparency at all levels, and establish accountability across teams.
- **Change Management:** Any policy change has upstream and downstream impacts, such as related product design changes, legal approvals, user notifications, internal process changes, lags in system updates, and updates to Terms of Service. Meta is continuously working to improve and enhance our coordination across various cross-functional teams for seamless change management as we work towards addressing the various integrity systems risks.
- **Policy Harmonisation:** Over the years, Meta's ads, commerce, and organic policies, have been written and evolved independently based on the specific requirements and circumstances of applicable surfaces. However, there are focused efforts underway across Meta to harmonise policy lines between organic and ad content to better ensure consistent enforcement.

### 5.2.4 Ads Systems

Facebook's ad systems and processes are designed to help the millions of people who use our services discover ads that they will hopefully find interesting and help businesses build a community, increase online sales, drive in-store traffic, and find new customers. Our Advertising Standards provide policy detail and guidance on the types of ad content we allow, our ad review process, and ad transparency requirements for

EU users pursuant to Article 39 of the DSA. We want our users to feel empowered and provide options to all users to control or customise their ad preferences on Facebook, which may include, but are not limited to, the following: hide an ad; hide all ads from an advertiser; select "Why am I seeing this?" to get more context; manage their ad preferences; and restrictions around targeting ads to minors. Additionally, Meta strives to provide transparency around ad targeting through our Ad Library.

As part of our ads review process, all ads are automatically reviewed against our Advertising Standards before launching on Facebook. We also use human reviewers to improve and train our automated systems, and in some cases, to manually review ads. Ads remain subject to review and re-review at all times, and may be rejected or restricted for violation of our policies at any time. We continue to improve our existing enforcement systems by testing and implementing new approaches to ensure a fair and effective ad review process. More information on how our ad systems work and how user data is leveraged to provide personalised ads can be found here. As part of our DSA compliance efforts, Meta established advertiser self-disclosure of beneficiary/payer as part of the ad buying process. We also updated our ads systems to prohibit the use of sensitive categories of data for ads generally.

While ads systems are a core feature of our platform and we have an extensive ecosystem of controls in place to manage them, some problematic content that does not violate our guidelines and policy violating paid content may be disseminated before we can detect and remove it. In other instances, some violating accounts may evade our content moderation systems. This influencing factor could have an impact on all Systemic Risk Areas. This insight was derived through the execution of our risk assessment where we evaluated and accounted for ads systems as an influencing factor in the following manner:

- **Assess Phase:** We conducted workshops with experts from our Ads Team, Content Policy, Cross-Integrity Team, Global Operations, and Legal and asked a series of questions for each Problem Area to understand how these systems influence Problem Areas and the associated risks. The questions focused on understanding the following types of matters:
    - Managing minors and ads;
    - Managing policy-violating ads; and
    - Potential limitations of our ads detection and enforcement mechanisms.
- **Measure Phase:** For each Problem Area, we evaluated the responses to the questions; took into consideration any issues and improvement areas identified through various signals, including assurance testing, where applicable; and evaluated compensating controls in place to address any limitations, such as our ads removal and monitoring controls.

As a result of the "Assess" and "Measure" phases of the assessment, we identified the following areas that should be considered for enhancement related to ads systems:

- **Intentional Manipulation:** Threat actors seek to exploit our systems and change how they act on the service to evade our detection and enforcement mechanisms, such as encouraging creators to post branded content as organic to circumvent our ads safeguards. Meta is aware of this and mitigates this risk by limiting an account's ability to post by taking down an ad, removing the account altogether, and/or using our dedicated team to perform more thorough research to see if threat actors operate as part of a network. Meta is continuing to mature its ads approach, including increased action at an account level.
- **Ad Standards Enforcement Guidelines:** We have identified areas where we need to uplift our current guidelines to improve the ads enforcement processes, including enhancing ad restriction and removal

process documentation specific to gambling and developing guidelines for our personal health and appearance ads policy and for misinformation/ disinformation ads.

- **Ad Technology:** Facebook's ad systems and processes are constantly innovating and improving. For example, gambling has been a particularly challenging area to address. We are working to improve our ad technology by enhancing our ad enforcement capabilities, re-training our classifiers, and adjusting our policies to address new trends.

### 5.2.5 Data Related Practices

Privacy and the protection of personal information are fundamentally important values for Facebook. As expectations around privacy evolve, it's critical Meta continues investing in guardrails and processes to meet people's privacy needs and expectations. We work hard to safeguard users' personal identity and information, and we do not allow people to post certain types of personal or confidential information about themselves or others.[61] We have a robust Product Compliance and Privacy Programme in place, led by our Chief Privacy and Compliance Officer, with extensive controls to protect privacy and security across Meta's services in line with privacy regulations, including the EU's General Data Protection Regulation (GDPR), and our GNI commitments. Since 2019, we have overhauled privacy at Meta, investing $5.5 billion in a rigorous privacy programme that includes people, processes, and technology designed to identify and address privacy risks early and embed privacy into our services from the start. We have grown our product, engineering, and operations teams focused primarily on privacy across the company from a few hundred people at the end of 2019 to more than 3,000 people at the end of 2023.[62] Additionally, our Privacy and Data Policy Team leads our engagement in the global public discussion around privacy, including new regulatory frameworks, and ensures that feedback from governments and experts around the world is considered in our product design and data use practices. Our External Data Misuse Team consists of internal experts dedicated to detecting, investigating, and blocking patterns of behaviour associated with scraping. Additionally, we want our users to feel empowered and provide options to all users to control their privacy on Facebook, including, but not limited to, the following: easy access to manage user information, controlling user experiences across our services, our Privacy Checkup Tool, and Two Factor Authentication.[63] Our technology detects the vast majority of privacy-related risks, but like any complex infrastructure, there are limitations.

Maintaining good data practices is critical to upholding our users rights (e.g., data retention allows us to restore accounts and actioned content that has been successfully appealed) and protecting our most vulnerable population from harmful content (e.g., age gating content and deployment of controls that protect minors). As such, this influencing factor could have an impact on the Fundamental Rights and Protection of Minors Systemic Risk Areas. The following insights were derived through the execution of our risk assessment where we evaluated and accounted for data related practices as an influencing factor in the following manner:

- **Assess Phase:** We conducted workshops with experts from our Global Privacy Programme and Legal and asked a series of questions for each Problem Area to understand how data-related practices influence Problem Areas and the associated risks. The questions focused on understanding the following types of matters:
    - Managing data access requests;
    - Managing data retention processes; and

[61] https://transparency.fb.com/en-gb/policies/community-standards/privacy-violations-image-privacy-rights/
[62] https://about.meta.com/privacy-progress/
[63] https://about.meta.com/actions/protecting-privacy-and-security/#privacy-controls

    - Potential limitations with engineering and pipelining challenges.
- **Measure Phase:** For each Problem Area, we evaluated the responses to the questions, took into consideration any issues and improvement areas identified through various signals, including assurance testing, where applicable, and evaluated compensating controls in place to address any limitations, such as our third-party reporting channels.

As a result of the "Assess" and "Measure" phases of the assessment, we identified the following areas that should be considered for enhancement related to data practices:

- **Protection of Minors' Privacy and Security:** Facebook defaults to private accounts and location settings when a minor signs up for an account, though minors can choose to turn their account to public and/or share their locations. Given that minors can make this choice, Facebook has implemented mechanisms, including tooling and classifiers, to help protect minors from interacting with potentially suspicious adults. However, this becomes challenging when threat actors leverage private groups and pages to evade detection and enforcement.
- **Data Use Limitations:** Meta is unable to utilise user level data in the EU due to an EU privacy regulation that restricts Meta from collecting and analysing personal data to identify if a medical or life-threatening issue is at hand. This impacts our ability to provide resources and a greater level of support to many users as it relates to suicide and self-injury risks. Additionally, Meta does not collect data regarding certain protected characteristics or types of users, such as minors, to protect user privacy, making it difficult to determine if certain groups are being targeted disproportionately.

### 5.2.6 Intentional Manipulation

Intentional manipulation negatively influences experiences on our platform and manifests in a variety of ways including, but not limited to, manipulated media, inauthentic use, or automated exploitation of our services. Whilst we have an extensive ecosystem of controls in place to manage intentional manipulation, some bad actors can continue to evade our controls and safeguards which could result in policy-violating and illegal content and behaviour occurring on the platform. This influencing factor could have an impact on all of the Systemic Risk Areas. The following insight was derived through the execution of our risk assessment where we evaluated and accounted for intentional manipulation as an influencing factor in the following manner:

- **Assess Phase:** We conducted dedicated workshops with experts from our Inauthentic Behaviour Team who are responsible for identifying intentional manipulation across all of our Problem Areas, along with representatives from Content Policy, Cross-Integrity Team, Global Operations, and Legal. We assessed all risks associated with inauthentic behaviour and asked a series of questions to understand how intentional manipulation influences Problem Areas and the associated risks. The questions focused on understanding the following types of matters:
    - Circumvention of our controls;
    - Types of intentional manipulation as it relates to each Problem Area; and
    - Assessing for intentionality.
- **Measure Phase:** For each Problem Area, we evaluated the responses to the questions, took into consideration any issues and improvement areas identified through various signals, including assurance testing, where applicable, and evaluated compensating controls in place to address any limitations, such as our Threat Disruption Network.

As a result of the "Assess" and "Measure" phases of the assessment, we identified the following areas that should be considered for enhancement related to intentional manipulation:

- **Inauthentic Cluster prevention:** While we have seen public discourse ahead of the June 2024 EU Parliamentary Elections focus primarily on foreign threats, including from Doppelganger, we have also seen domestic activity focused on the EU, including coordinated inauthentic activity in Croatia and simpler inauthentic clusters we removed in recent months in Europe, including in France, Germany, Poland and Italy.[64]
- **Circumvention:** Threat actors test new strategies and behaviours to evade detection and enforcement requiring Meta to consistently update and strengthen systems and processes in place (e.g., use of emojis and hashtags, implicit threats, slurs, and coded language).
- **Spam:** Certain Problem Areas, such as Human Exploitation and Child Sexual Exploitation, Abuse, and Nudity, may be subject to adversarial spamming which may draw away resources from managing actual policy violating content. We manage adversarial spamming by launching strategic network disruptions, which allow us to disable multiple threat-related accounts at once and also help address coordinated attacks.
- **False User Reporting:** Adversarial actors may also seek to misuse user reporting, for example, using the reporting system to take down a political opponent for false reasons. Meta has made significant investments in enhancing reactive measures to identify inauthentic reports, including machine learning models to read signals and email verification in contact forms.
- **Bot Farms:** Coordinated attacks using automated programmes can be used by threat actors to scrape user data, fabricate and push initiatives, and disseminate disinformation and scams. Meta is aware of these types of attacks and implements controls to detect these types of manipulation, including our coordinated attack prevention controls.

### 5.2.7 Generative Artificial Intelligence (AI)

As a leader in the AI space, Meta has made several major investments and maintains its commitment to open innovation in our foundational AI technologies. Prior to deploying functionalities on Facebook in the EU that are likely to have a critical impact on systemic risks, Meta undertakes a Critical Impact Risk Assessment (CIRA), in line with Article 34, to determine if systemic risks may be impacted, and accordingly, which reasonable, proportionate, and effective mitigations we need to implement prior to launch. Any such new functionalities, including generative AI-related products, will undergo this process to the extent applicable.

We believe that an open ecosystem brings transparency, scrutiny, and trust to the development of AI and leads to innovations that everyone can benefit from that are built with safety and responsibility top of mind.[65] In late spring of 2023, we began reevaluating our policies to see if we needed a new approach to keep pace with rapid advances in generative AI technologies and usage. We completed consultations with over 120 stakeholders globally. We also conducted public opinion research with more than 23,000 respondents and engaged with dozens of experts all over the world to get their feedback on how we should approach AI-generated content on our platforms. Additionally, the independent Oversight Board provided us with recommendations based on consultations with civil-society organisations, academics, inter-governmental

[64] Doppelganger is a long running covert influence operation from Russia centred around a large network of websites spoofing legitimate news outlets. Meta Quarterly Adversarial Threat Report Q1 2024
[65] https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/

organisations and other experts. Overall, we heard broad support for labelling AI-generated and photo/audio realistic content and strong support for a more prominent label in high-risk scenarios.[66]

In response, we have set up a number of activities to monitor and respond to the use of generative AI by users. We perform risk assessments of the generative AI violations on our platforms as needed and conduct gap analyses to understand the comprehensive landscape of the use of generative AI on our platforms. We are also implementing a transparency labelling and self-disclosure flow to allow users to report AI generated content, utilising synthetic data[67] to train classifiers and improve their performance against AI content, and providing signals to human reviewers when something we know is created by generative AI. We also recently held a Community Forum on generative AI that included over 1,500 people in Brazil, Germany, Spain and the United States. The Forum was designed to solicit public feedback to complement the inputs we receive from experts, academics and other stakeholders through our policy development processes and we shared the results of the form publicly so more companies, researchers and governments can benefit from what participants shared.[68] We will continue to keep a pulse on the evolving use of generative AI and scale monitoring and response activities to address identified risks.

While we have an extensive ecosystem of controls in place to manage the evolving use of generative AI, some problematic content that does not violate our guidelines and policy violating content may be generated and disseminated before we can detect and remove it. This influencing factor could have an impact on all of the Systemic Risk Areas. This insight was derived through the execution of our risk assessment where we evaluated and accounted for generative AI as an influencing factor in the following manner:

- **Assess Phase:** We conducted a workshop with our generative AI risk Team and asked a series of questions for each Problem Area to understand how generative AI influences Problem Areas and the associated risks. The questions focused on understanding the following types of matters:
    - Problems Areas impacted by third party use of generative AI;
    - Meta's approach to managing generative AI, including third party use and Meta's use; and
    - Challenges in managing generative AI.
- **Measure Phase:** For each Problem Area, we evaluated the responses to the questions, took into consideration any issues and improvement areas identified, where applicable, and evaluated compensating controls in place to address any limitations, such as our classifiers and monitoring controls, and determined control effectiveness scores accordingly. Additionally, we also evaluated the impact of generative AI on inherent risk and increased the volume scores of risks, where applicable.

As a result of the "Assess" and "Measure" phases of the assessment, we identified the following areas that should be considered for enhancement related to generative AI:

- **Deepfakes detection:** Our processes are still evolving to detect content that violates our policies and this issue specifically gets more complicated with regards to deepfakes and any other form of altered content, including impersonation of celebrities and high-profile individuals. We are working on detection of AI generated content with visible and invisible markers to further mitigate this issue.
- **Scaled Content Volumes:** Generative AI makes possible the creation of large volumes of content, which could impact areas, such as Child Sexual Exploitation, Abuse and Nudity, Misinformation, Disinformation, Fraud and Deception and Inauthentic Behaviour. Through the formalised approach

[66] https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/
[67] Synthetic data is information that is artificially created rather than produced by a real-person.
[68] https://about.fb.com/news/2024/04/leading-the-way-in-governance-innovation-with-community-forums-on-ai/

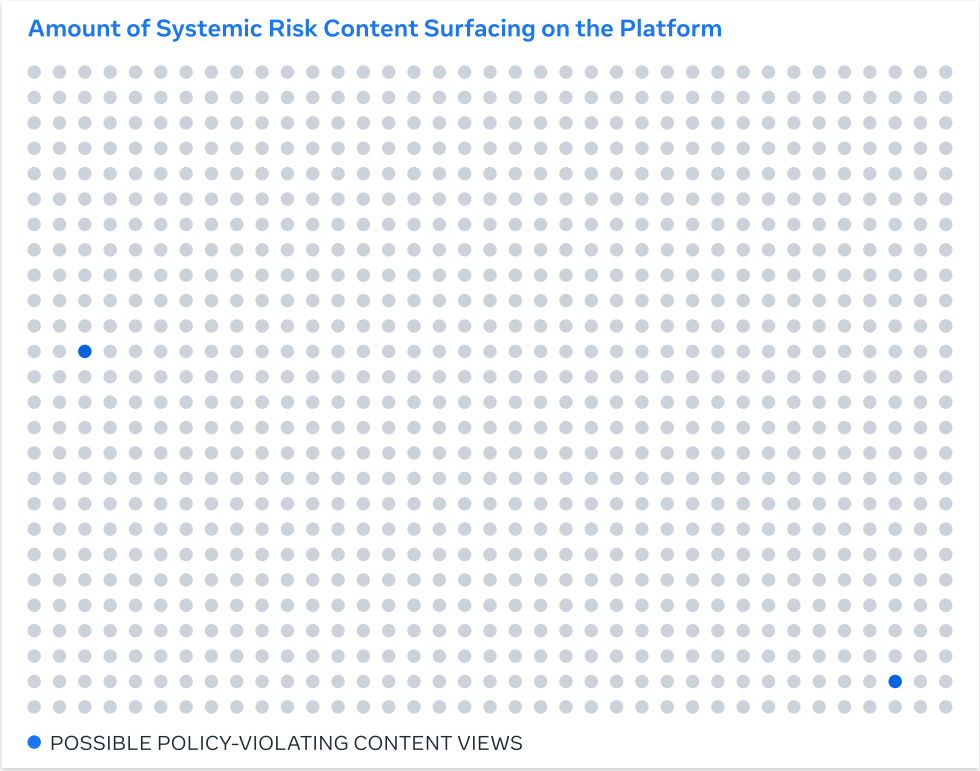
and programme Meta has stood up to manage generative AI risks, we have not identified a notable increase in volume to date due to generative AI for the majority of these areas. It is something that Meta is continuing to monitor while scaling its mitigation capabilities.

# 6. Our Detailed DSA Systemic Risk Assessment Results

Meta is committed to maintaining a safe and trusted environment on Facebook. Risks can arise on our service when users share policy violating content or engage in policy violating behaviour, and in some cases these risks can have an impact in the EU. We acknowledge that these risks can be influenced or exacerbated by certain 'Influencing Factors' (as described in Section 5.2 Influencing Factors) and that, despite our best efforts, we alone cannot identify and address every risk that may arise. To address this limitation, we empower our Facebook community with tools to help us identify potential risks. We also respond to lawful requests from law enforcement, regulators, and courts, and work closely with other external partners in an effort to learn and to help keep users in the EU safe on our service.

Furthermore, it is critical to note that policy-violating behaviour and content accounts for a very small portion of the content that the average user sees and interacts with on Facebook. As illustrated in **Figure 4**, when problematic content does arise, the percentage of users who come across this type of content is minimal.

**Figure 4. Policy-Violating Content on Facebook (Illustrative)**[69]




Meta is committed to identifying, assessing, and mitigating systemic risks associated with use of Facebook, using the DSA Systemic Risk Assessment as one of our key primary assessment instruments.

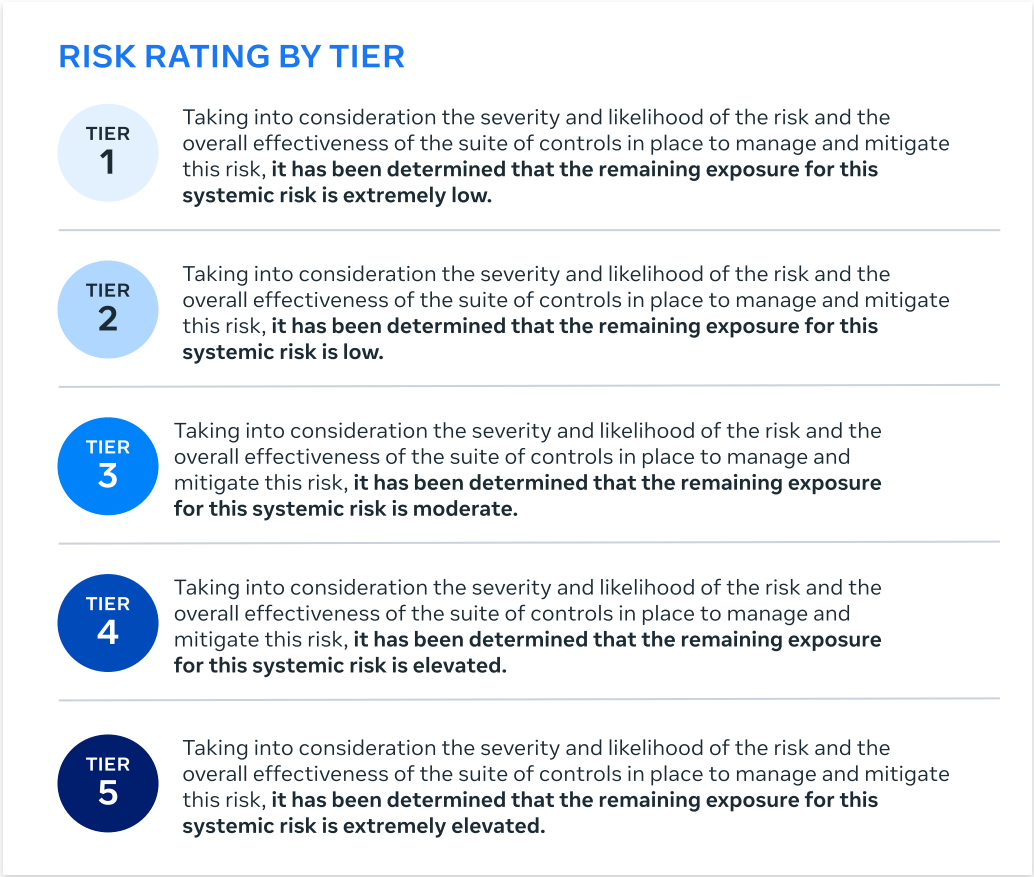
[69] https://transparency.meta.com/en-gb/policies/improving/prevalence-metric/

### 6.1 Risk Analysis

Meta evaluated the inherent risk of 122 risks that cut across the 19 Problem Areas and determined the effectiveness of controls for Facebook to determine the residual risk for each in-scope risk. These controls were evaluated individually to determine how effectively they mitigated each risk, as applicable.[70] All residual risk scores were ranked and rated using an ordinal tiering scale from Tier 1 to 5 that allows us to measure Problem Areas consistently.

#### 6.1.1 Risk Rating

These Tiers denote the significance of the risk to users in the EU and are described as follows:




#### 6.1.2 Problem Area Analysis: Inherent Risk

In order to measure this risk, referred to as inherent risk in the assessment, we estimate the significance of: 1) the potential negative impact of a systemic risk to users and society (Severity), and 2) the probability that the risk will materialise (Likelihood), absent any mitigation measures. As part of this evaluation, we consider how various Influencing Factors increase or reduce each of the Problem Areas and the associated risks.

During the assessment, we identified certain trends that could potentially increase the inherent risk exposure across the platform, including the relatively higher (compared to other years) number of elections in the EU,

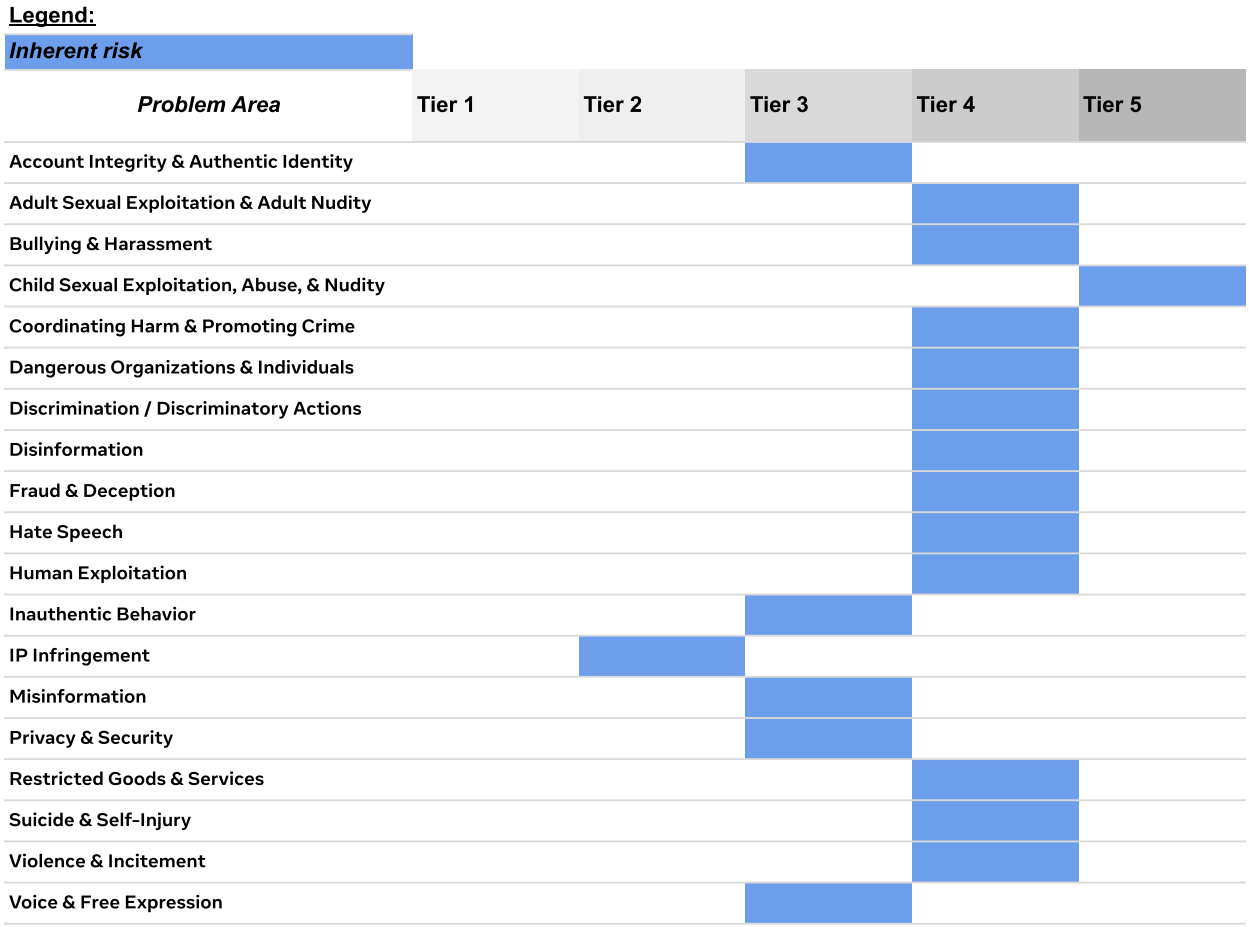
[70] More information on our Integrity Risk Assessment Methodology and Severity, Likelihood, and Control Effectiveness Rubrics can be found in Appendix 9.1.

increasing adoption of generative AI, crises in adjacent regions, and global events, such as the preparation for the Olympic Games in Paris. Other external factors, such as trends related to bullying and harassment of minors, rising popularity of gambling online, and targeting minors for sale of certain prohibited goods, also impacted the inherent risk posture across specific Problem Areas.

The Inherent Risk Results for each Problem Area are included in **Figure 5** and details on how inherent risk is calculated and broken down into Tiers can be found in Appendix 9.1.1.

As we have matured our methodology for Y2, we've made substantive improvements and we are therefore reporting our risk scores in the form of Tiers to better support YoY comparisons.

**Figure 5. Inherent Risk Results**




### 6.1.3 Problem Area Analysis: From Inherent Risk to Residual Risk

As illustrated in **Figure 5,** we determined that inherent risk exists for each Problem Area. When factoring in our controls designed to mitigate and control risk, the overall inherent risk is reduced across all Problem Areas, resulting in a measurement of the residual risk. The residual risk measurement equation can be found in Appendix 9.1.3.

The Problem Areas with the highest average inherent risk, which also have some of the most notable reductions in residual risk, thanks to the effectiveness of mitigations, are: **(1) Child Sexual Exploitation, Abuse**

**and Nudity, (2) Bullying and Harassment, and (3) Suicide and Self-Injury.** The actual and foreseeable risk context of these Problem Areas are further described in Section 6.2 along with controls in place to mitigate and control for their associated risks.

**Figure 6. Residual Risk Results**


As detailed in **Figure 6,** the evaluation we conducted reveals our controls meaningfully reduce the inherent risk across Problem Areas, shifting all to **Tier 1 and Tier 2.** This is further evidenced by our Community Standards Enforcement Report (here), which includes global data that tracks our progress and demonstrates our continued commitment to making Facebook safe.

### 6.1.4 Year-Over-Year (YoY) Results Comparison

Outlined below are several notable YoY assessment results headlines and insights on the risk landscape, inherent risk, control and mitigation measures, and residual risk.

**Risk Landscape: The risk landscape over the last year shifted and became more elevated and complex due to external events and extenuating factors, such as elections, geopolitical conflict, preparation for the 2024 Olympics and generative AI.**

During this year's DSA Systemic Risk Assessment (Y2), the integrity risk landscape shifted and elevated due to (1) the various elections that were carried out across the European Union, which introduced the increased potential for risks related to areas, such as Violence and Incitement, Misinformation, and Disinformation; (2) global events, including conflicts in adjacent regions and the preparation for the 2024 Olympics, which involved more movement of people and posed increased risk around areas like human exploitation; and (3) the advancement of generative AI, which can be abused to manipulate media and impersonate individuals.

Given this risk landscape shift and our efforts to mature our overall risk assessment process, we identified 122 risks associated with 19 Problem Areas and, in turn, the 8 Systemic Risk Areas. This represents an increase of two additional risks from Y1, which are as follows:

- "Live Non-Endangered Animals and Endangered Species"; and
- "Under Thirteen (U13) Year Olds on the Platform".

**Inherent Risk: The majority of Problem Areas were impacted by external factors and events that increased the complexity of the risk landscape and potential for abusive online behaviour and policy-violating content; however only two Problem Areas changed Inherent Risk Tiers.**

Due to external factors and events, our assessment of average inherent risk resulted in a more complex risk landscape across most Problem Areas, with the noticeable exception of Adult Sexual Exploitation and Adult Nudity, IP Infringement, Privacy and Security, and Suicide and Self-Injury, which remain the same as Y1. Despite this increased complexity, most Problem Areas remained in the same Inherent Risk Tier as in Y1 with the exception of Account Integrity and Authentic Identity and Child Sexual Exploitation, Abuse and Nudity, which in Y2 moved to a higher Inherent Risk Tier. This is mainly attributed to the higher number of elections YoY, conflicts in adjacent regions, generative AI and the discrete risk of having "Under Thirteen Year Olds on the Platform".

**Controls and Mitigation Measures: We worked to improve the overall effectiveness of our control environment.**

Last year, we shared our mitigation measures in our DSA Systemic Risk Mitigation Report 2023. This year (Y2), our risk mitigations are directly included in this Report and focus on observations identified during design and operating effectiveness evaluation. Since Y1, Meta implemented new controls and enhanced existing controls, which are described in detail in Section 6.2. Some enhancements that have been implemented since we finished our analysis of this assessment, can be found in Section 7. It should also be noted that, during Y1, we measured design and mitigation effectiveness to derive overall residual risk; this year we have matured our overall risk management process, and started to measure residual risk by evaluating both design and operational effectiveness of our controls and relying on a set of new signals.

Our approach to risk mitigation involves developing and executing mitigations that are reasonable, proportionate, and effective to reduce risk exposure while maintaining our commitment to respect the human rights of our users, including the fundamental rights recognised in the EU Charter.

**Residual Risk: YoY our residual risk Tiers remained constant for around 95% of our Problem Areas and around 5% changed from Tier 1 to 2.**

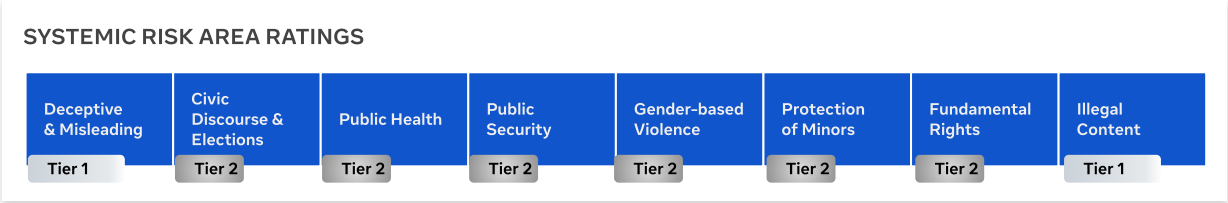
Problem Areas whose residual risk remained the same or decreased YoY had increased the strength of controls and deployed targeted investments to manage these risks and the impact of external factors. Despite the increase in external risk events, only one Problem Area (Human Exploitation) moved from Tier 1 to Tier 2 YoY, which is due to conflicts in adjacent regions and the preparation for the Olympics. In short, around 63% of all Problem Areas were assessed as being a Tier 2 level of residual risk and the remaining Problem Areas were determined to be a Tier 1 level of residual risk. We are able to manage and mitigate these risks due to our effective control environment and continued investment in critical areas, including, but not limited to, the following:

- Activating a dedicated team to develop a tailored approach to help prepare for the EU Parliamentary Elections;
- Tools, classifiers and training;
- Enhancing mechanisms to further help protect minors;
- Harmonising our policies across organic and paid content; and
- Cross-platform enforcement.

### 6.1.5 Systemic Risk Area Analysis: Risk Ratings

Once we assessed, evaluated, and measured all the identified Problem Area risks and accounted for mitigating measures in place, we then derived an overall risk rating for each Systemic Risk Area by combining the risk scores of each Problem Area associated with a Systemic Risk Area based on the Systemic Risk Landscape described in Section 5: Systemic Risk Landscape.

The image below depicts the Tier Ratings across all Systemic Risk Areas and how each Systemic Risk ranks among each other. These Tier Ratings remain the same as in 2023.




## 6.2 Mitigating Measures Analysis

Meta has a robust set of controls in place, including rigorous structures and processes to identify, mitigate, and manage risks in a variety of ways. We are committed to continuously improving our ecosystem of controls and the ISSO-GRC Programme, which includes capturing a set of controls to address the evolving risk landscape and regulatory requirements, including Article 35 of the DSA. We have a comprehensive **Integrity Common Control Framework** that groups our documented integrity measures we have established across our lines of defence including, but not limited to, controls implemented within our governance model, platform infrastructure, operating processes, Facebook services, and the ISSO-GRC Programme. We maintain an inventory of these controls, which groups controls into domains (see **Figure 7)** and is used to map controls to risks ahead of executing the assessment.

We have made significant improvements to our control documentation. For example, we further standardised control descriptions and included additional attributes, improving our ability to document controls in a clear, consistent, and programmatic manner. Additionally, we have expanded the Control Framework to include more specific controls which detail differences at the product, problem area, jurisdiction or content type level.

This allows us to maintain a single source of truth across all efforts which require an understanding of the hundreds of integrity controls Meta has in place, including assurance testing, other risk assessments, and audits.

**Figure 7. Meta's Integrity Common Control Framework: Control Domain**


### 6.2.1 Meta's Ecosystem of Controls

As seen in **Figure 7** above, Meta has a robust set of controls in place that are organised into control domains. Meta's ecosystem of controls include controls that are implemented across Problem Areas and controls that are implemented for specific Problem Areas. The control domains detailed in this subsection focus on how each of these control areas operate and detail the foundational controls Meta has in place for each domain. Information on some of the critical controls for each Problem Area is detailed in Section 6.2.2.

#### 6.2.1.1 Policies and Standards

At Meta, we are committed to giving people a voice and keeping them safe. To help with this commitment, we have a set of globally applicable policies and standards, including our Terms of Service, Facebook Community Standards, and Advertising Standards, which describe what is allowed and what is prohibited on our platforms. Our teams work together to develop our policies and enforce them. We engage externally with global experts in technology, public safety, human and civil rights, civil society organisations, activist groups, thought leaders and academics to create and update our policies. We take great care to include people from all communities, particularly marginalised communities, in the evolution of our standards. In accordance with these policies, we build features to keep our users safe and enforce our policies using technology and human review. We provide information about all our policies, their effect on enforcement, and transparency reports in our Transparency Centre. We make our policies and standards accessible by translating our Facebook Community Standards into more than 90 languages.

Over the past year, Meta has invested in the harmonisation of our policies across organic and paid content to unify approaches, systems and enforcement activities and mitigate any potential coverage gaps. This is a phased approach that Meta is working on at a Problem Area level to enable better cross-platform enforcement and consistency across platforms and surfaces.

The following table details the foundational controls that were assessed for this year's Systemic Risk Assessment as it relates to this control domain.

| Domain | Foundational Control Name | Foundational Control Description |
|---|---|---|
| Policies and Standards | Child Safety Playbooks | Meta maintains playbooks relating to child safety that are leveraged by teams across Meta to create standard operating procedures (SOPs) for managing content relating to child safety on Meta's platforms. These playbooks enable Meta to develop SOPs and inform the deployment of technology and enforcement activities related to child safety at the content or account/actor level, including detection technology, classification systems, machine learning (ML) algorithms, reactive user reporting, and root cause evaluations. |
| Policies and Standards | Commerce Policies | Meta publicly maintains Commerce Policies within the Other Policies section on the Transparency Centre page to provide guidance to sellers and buyers over products sold on Meta's platforms. The Commerce Policies can be accessed by users and non-users and include specific policies over prohibited and restricted content, as well as steps to take to appeal decisions, as needed. |
| Policies and Standards | Facebook Community Standards | Meta maintains public Facebook Community Standards that provide policy detail and guidance on what is and is not allowed on Facebook. The Facebook Community Standards detail the policies in place, rationale and aims of the policies, and behaviours that may result in a violation of the Community Standards. The following categories are used to define policy-violating and lead to actioning and enforcement of the Community Standards: Violence and Criminal Behaviour, Safety, Objectionable Content, Integrity and Authenticity, Respecting Intellectual Property, and Content- Related Requests and Decisions. |
| Policies and Standards | Implementation Standards | Meta's Implementation Standards provide detailed guidance for applying Meta's Community Standards. These standards are used to inform and define Operational Guidelines and system implementations. Meta's Implementation Standards are regularly updated, which includes reviewing the list of new policy updates from the Policy Launch Pipeline, adding them to the Implementation Standards Tracker, determining if a subsequent change(s) to the Implementation Standards is required, and updating as needed. |
| Policies and Standards | Internally Assigned Turnaround Time (TAT) on User Initiated Reports | Meta maintains a process to respond to user initiated reports within the appropriate internally assigned turnaround times. Meta's Problem Integrity Operations sets and executes the Service Level Agreements (SLAs). Operational teams monitor performance against SLAs and put interventions into place so reports that pose the highest risk to users and society are reviewed expeditiously. |
| Policies and Standards | Meta Advertising Standards | Meta maintains public Advertising Standards that provide policy detail and guidance on the types of ads allowed at Meta and the types of ad content that are prohibited. The Advertising Standards also detail the ads review process and advertiser behaviour that may result in ads restrictions. The standards are reviewed periodically and updated on an as-needed basis. |
| Policies and Standards | Meta Integrity Standards | Meta maintains the Meta Integrity Standards (MIS) which are a set of standards that inform integrity mitigations to products prior to launch. Meta Integrity Standards require product integration with Meta's applicable integrity systems and apply to product launches that impact user generated content and/or what users will be able to view on Meta's platforms. |
| Policies and Standards | Meta's Code of Conduct | Meta's Code of Conduct is translated into 16 languages and establishes Meta principles, mission and values, defining the desired behaviour expected from all Meta Personnel and leaders. Violations may result in disciplinary action, up to and including termination of employment or assignments. Meta has different channels in place to report a violation by phone, email or |

| | | |
|---|---|---|
| | | SpeakUp. Training on the Code of Conduct for all Personnel at Meta is provided annually, and additional training for teams prior to annual training launch can be requested to the Regional Ethics and Compliance Manager or emailing Legal to request a session. |
| Policies and Standards | Policy Change Management | Meta develops and maintains a policy change management process to enable governance over changes made to Meta's Integrity policies. The policy change management process provides guidance over how teams at Meta can draft, propose, and make changes to existing Integrity policies, including references to the relevant sign-offs and forums used throughout the process. This policy change management process also covers vendor change management that impacts integrity work. |
| Policies and Standards | Privacy Policy | Meta maintains and updates the Meta Privacy Policy, which documents how Meta collects, uses, shares, retains and transfers information for all in-scope products. The policy provides a variety of resources for additional context, support, and management of a user's privacy settings. |
| Policies and Standards | Product Policies | Meta's Transparency Centre maintains multiple policies that apply to Meta's platforms, aiming to protect users and society from potentially harmful content or behaviour. Policies housed within the Transparency Centre are updated on an ad hoc basis typically due to changes or evolutions in risks, Problem Area, users, within region, trends and the landscape. Updates needed for DSA purposes are managed by designated teams working with Legal. Meta's Transparency Centre contains the following information:<br>- Pages, Groups and Events Policies for Facebook;<br>- Branded Content Policies for Facebook and Instagram;<br>- Commerce Policies for Facebook, Instagram and WhatsApp;<br>- Recommendations Guidelines for Facebook; and<br>- Supplemental Facebook View Terms of Service. |
| Policies and Standards | Terms of Service (TOS) | Meta maintains publicly available terms and conditions inclusive of TOS for their platforms which includes information on the restrictions that may be imposed in relation to the use of their services and may make available through the TOS links to policies, procedures, measures, and tools used for content moderation and internal complaint handling. Where legally required, Meta notifies users in the event of a material change to the TOS and has mechanisms that enable this notification (such as jewel notifications, megaphone notifications, or email, as appropriate). |

#### 6.2.1.2 Systems and Product Integrity

In addition to our standards and detection and enforcement mechanisms, we take additional steps to embed safety by design to help users engage safely online and on our services, particularly for minors and marginalised communities. This includes age gating certain types of content, parent guides, and default privacy settings to help protect minors on our services. We work to remove inappropriate content from teens' experiences even if shared by someone they follow. For teens using our apps, we have expanded the automatic placement of teen users into more restrictive content control settings to make it less likely for them to come across potentially sensitive or problematic content.[71] Additionally, we implement several safety prompts and features, such as frictions, to inform users when their actions may violate or be close to violating our policies. We also have blocking and snoozing functions that enable users to limit unwanted content and interactions.

To enable Meta's technologies to continue to help users have positive online experiences, investment in new tools and routine training and evaluation of existing tools and policies are a priority to maintain system integrity. This includes routine detection and enforcement system training, monitoring, and maintenance to test the efficacy of these functionalities, monitoring of metrics and trigger mechanisms, coordinated attack prevention, and internal reviews of new or changing services and functionalities before they are launched.

[71] https://about.fb.com/news/2024/01/teen-protections-age-appropriate-experiences-on-our-apps/

Meta is continuing to invest in strengthening its age-related safety measures which includes improving the use and consistency of our age prediction and assurance models across our platforms and surfaces.

The following table details the foundational controls that were assessed for this year's Systemic Risk Assessment as it relates to this control domain.

| Domain | Foundational Control Name | Foundational Control Description |
|---|---|---|
| Systems and Product Integrity | Account Integrity and Verification | Meta maintains verification processes for official accounts related to elections, commerce, and paid content on Meta's platforms. Accounts may be reviewed by Meta prior to these accounts being able to post content and interact with users. |
| Systems and Product Integrity | Ad Targeting Restrictions for Minors | Meta has safeguards and restrictions around advertising, including teen-specific safeguards for users under the age of 18 (minors). This is done by limiting the types of ads that can be shown to teens and limiting the targeting choices available to advertisers to reach users under the age of 18 across Meta's platforms. |
| Systems and Product Integrity | Age Appropriate Content Experiences | Meta implements processes to provide users under 18 years with age appropriate content experiences on Meta's platforms. Meta leverages age-appropriate content detection and filtering methods to reduce teens' exposure to potentially harmful and inappropriate content. |
| Systems and Product Integrity | Age Identification and Verification | Meta has processes to identify and verify users' ages on its platforms allowing Meta to provide age-appropriate experiences. Depending on the platform, Meta requires users who may be misrepresenting their age and/or seeking to change their age from under 18 to over 18 to verify their age through various verification options, such as uploading an identification, providing their birthday, and/or recording a video selfie. |
| Systems and Product Integrity | Age-related Recommendation Restrictions | Meta implements age-related recommendation restrictions to reduce the likelihood of teens encountering potentially sensitive or low quality content. Meta's integrity systems classify content as being age-appropriate or not based on relevant signals. |
| Systems and Product Integrity | Age-Related Safety Measures | Meta has built age-related safety measures through features that minors and parents/guardians can access to further protect minors. Minors receive in-product reminders on how to adjust their settings and curate the content they see, and parents have access to supervisional tools and controls to adjust their child's settings. |
| Systems and Product Integrity | AI User Self-Disclosure Labelling | Meta maintains a feature for users to self-disclose when they share photorealistic video or realistic-sounding audio on Meta's platforms. Self-disclosure for photorealistic video or realistic-sounding audio is required and results in a label on posted content. Meta may apply penalties to users who fail to properly self-disclose AI-generated content. |
| Systems and Product Integrity | Behaviour | Meta maintains a Behaviour Pillar to address and fight abuse across a spectrum of scale and sophistication from scripted and coordinated activity to deceptive links to various types of scams. Meta builds on multiple enforcement strategies (a mix of actions on actors, content, pages, links, etc.) and complex sets of signals to respond to sophisticated adversaries. |
| Systems and Product Integrity | Blocking Function | Meta operates a technical blocking mechanism that allows a user to prevent another user from seeing their activity on Meta's products, including the user's profile, posts, or stories. Users can either go into their settings or profiles of another user to block. |
| Systems and Product Integrity | Bug Bounty Programme | Meta maintains and promotes a Bug Bounty Programme to incentivise external individuals to responsibly disclose security bugs that could compromise the integrity of Meta user data, circumvent the privacy protections of Meta user data, and/or enable unauthorised access to a system within the Meta infrastructure. Individuals submit their findings, which are then triaged, validated, and remediated, as needed. The Bug Bounty Programme then compensates individuals for discovering impactful vulnerabilities to the organisation accordingly, with cash or other incentives. |

| | | |
|---|---|---|
| Systems and Product Integrity | Celeb-Bait Playbook | Meta maintains the Celeb-Bait Playbook which provides holistic guidance for the detection and enforcement of Celeb-Bait Ads. The Playbook guides Market Specialists on detecting both content, as well as actor behaviour signals, to retrieve relevant data points for effective enforcement. |
| Systems and Product Integrity | Civic Actors (CVA) List | Meta maintains and updates the Civic Actors List to quickly provide protections for people in high risk election countries, as deemed by the Global Response team, due to their contributions to civic affairs, either online or offline. Individuals are identified to be CVAs through an internal submission form where approved employees may provide evidence of a high-risk individual's authentic account for consideration. Individuals are also identified based on vetted sources and through various platform signals. The CVA List is used to apply specialised protections, including enforcements related to impersonation, account compromise, and harassment. |
| Systems and Product Integrity | Content Interstitials | Meta maintains content interstitials to warn or provide contextual information to users with regards to content that is not policy-violating but that may be sensitive for certain users. The interstitial is introduced to enable users to decide how to engage with potentially sensitive content. |
| Systems and Product Integrity | Coordinated Attack Prevention | Meta's Coordinated Attack Discovery System is an ongoing prevention measure implemented to defend against coordinated attacks on Meta's system. This system includes near-real-time visualisations of possible attacks, along with investigation tools to find and explore connections among attack clusters, particularly looking at recent traffic patterns. |
| Systems and Product Integrity | Dangerous Organisation and Individuals (DOI) Designation | Meta oversees the process by which entities (organisations and individuals) that qualify as a 'Dangerous Organisation and Individual" are added to the DOI Designations List. Designations are proposed through a designations nomination form and assessed by a cross-functional team of experts. This cross-functional team conducts a review of both on-platform and off-platform information to see if the nominated organisations or individuals meet the threshold for designation. If designated, measures are taken to remove all of the DOIs assets from the platform, as well as bank terms and images that represent the DOI in order to bolster continued scaled enforcement. All glorification, support, and representation of the designated DOI is also prohibited. |
| Systems and Product Integrity | Design Taxonomy | Meta maintains a taxonomy of design patterns that have been identified as likely to be deceptive in order to avoid implementing such designs on Meta's platforms. The taxonomy utilises patterns flagged as noteworthy by key external sources (e.g., DSA, European Data Protection Board (EDPB), and Federal Trade Commision (FTC)) to provide examples that represent deceptive designs that may lead a reasonable person to feel tricked, misled, coerced, or unduly pushed into doing something they wouldn't have otherwise chosen to do. The taxonomy is used by product designers as guidance for their designs to avoid deceptive designs. |
| Systems and Product Integrity | Differential Review | Meta maintains a machine learning model and an algorithm code update process that requires any changes to be reviewed by at least one other engineer. This ensures that no one person can make changes on their own and the process always tracks who made the changes so that any bugs can be identified. |
| Systems and Product Integrity | Dogfooding | Meta maintains dogfooding programmes over Family of Apps (FoA) products to ensure any bugs and issues associated with products, services, and features are identified and rectified prior to going live to external users. This means internal users (employees) test new or updated products, services, and features as external users would and identify and file bugs, locate internationalisation issues, and provide feedback to ensure the final build is as high quality and polished as possible. |
| Systems and Product Integrity | Epsilon | Meta maintains the Epsilon checkpoint which protects accounts that have been compromised and blocks bad actors from continuing to access the account. When there is suspicion that an account has been accessed by a bad actor, the account goes through four steps: block (lock down account so no further harm can be caused), authenticate (make person attempting to access account prove they are account owner), secure (help account owner secure account from future attacks),and restore (return account to former state). |

| | | |
|---|---|---|
| Systems and Product Integrity | Facebook Well-Being | Meta maintains a programme to support users on Facebook and prevent problematic use on the platform. Users are encouraged to access time management tools and resources to have better control and feel more agency over their experience on Facebook. |
| Systems and Product Integrity | Hack and Leak Playbook | Meta removes content claimed or confirmed to be from hacked sources, except in limited cases of newsworthiness. When a suspected hack and leak of material becomes known, the Strategic Response Policy team will drive a review of the content and actors and escalate to leadership for a decision. |
| Systems and Product Integrity | Hidden Words Function | Meta maintains the Hidden Words Function to empower users to filter out potentially offensive messages and comments on Meta's platforms. Users can better control their experience on Meta's platforms by hiding what they may find to be offensive, abusive, or otherwise unwanted interactions. |
| Systems and Product Integrity | Identification of Suspicious Adults (Detection) | Meta maintains processes to proactively identify adults who display suspicious behaviour towards minors before they potentially perform policy violating actions. Meta uses detection systems to identify risky interactions which are likely to be policy violating behaviour. |
| Systems and Product Integrity | Identity Verification and Authenticity | Meta has processes in place to verify a user's identity on Meta's platforms. Meta requests users provide a copy of an item with their name and date of birth or name and photo on it as proof of identification, such as a photo identification (ID) issued by a government, an ID from a non-government organisation, an official certificate, a licence that includes the user's name, with potential special ID requirements for some cases, such as advertisers running ads about social issues, elections or politics. Proof of identification is checked for usability and forgery via machine learning models or human reviewers, then matched against the account profile. |
| Systems and Product Integrity | Individual Verification for Account Ownership | Meta maintains processes for individuals to regain access to their accounts on Meta's platforms. When individuals are unable to access their accounts, they can recover them by following steps detailed on the specific platform's Help Centre based on the login issue they are experiencing, such as forgot password, forgot username, and disabled account, following necessary individual verification processes. |
| Systems and Product Integrity | Integrity Review | Meta manages the Integrity Cross-Functional (IXFN) Review Process which reviews applicable new product launches that impact user generated content and/or what users can view on Meta's platforms for integrity risks in an effort to comply with various regulatory obligations. New products are reviewed against the MIS to help identify and mitigate potential risks to user safety, rights, and protection. If potential integrity risks are identified, teams are provided with mitigations that aim to address those risks. The review consists of five stages: Intake, Risk Assessment, Implementation, Verification, and Launch. |
| Systems and Product Integrity | Log Out | Meta has mechanisms in place to protect user accounts through a Log Out feature on Facebook and Instagram. "Where You're Logged In" section of "Security and Login" settings shows users a list of devices that have been recently used to log in to their account and users can log out of those accounts following instructions provided. |
| Systems and Product Integrity | Login Challenges | Meta maintains a Login Challenges flow that is triggered after a user successfully logs into an account on a device not attributed to the user in the past. After a user successfully logs into an account on a new device, the user is required to approve the login via an approval push or jewel notification from another device which has already been attributed to the user or pass another authentication challenge. |
| Systems and Product Integrity | Malicious Scam Actors (MaSA) | Meta maintains the Malicious Scam Actors (MaSA) review protocol framework for scam actors on Meta's platforms. MaSA is the protocol used to label FB and IG accounts as scam actors and to produce metrics and train ML models. |
| Systems and Product Integrity | Pages and Groups Admin Controls | Meta has in-product, ongoing controls for admins running pages and / or groups in apps to moderate certain types of content, keywords, behaviour (e.g., harassment), etc. even if it doesn't violate policy. This allows those admins to create a better space for the conversations they want to promote with their page / group and community of users as they see fit. Meta reminds admins, especially political actors, that they may have their own legal requirements to not block certain content or use certain features in some ways, including a limit to the number of content filtering rules that can be applied. |

| | | |
|---|---|---|
| Systems and Product Integrity | Privacy Checkup | Meta manages the Privacy Checkup feature designed to help inform users on various privacy-related topics and adjust their privacy settings on Facebook. The feature is available on the settings page where users can review and change their privacy settings related to who can see what they share. Users can also explore a module of topics with information on how to keep their accounts secure, how people can find them on Facebook, their data settings on Facebook, and their ad preferences. |
| Systems and Product Integrity | Privacy Review | Meta maintains a Privacy Review Process to address privacy risks before Meta releases a new product or makes material changes to products or features. Through the Privacy Review Process, proposed new products, services, or practices or proposed modifications to products, services, or practices that collect, use, or share user data, as well as any external commitments that Meta makes regarding the privacy or security of user data are assessed, reviewed, and approved. During Privacy Review, Meta flags potential deceptive design risks for new experiences that are being developed. If deceptive design risks exist, the experience is required to be changed prior to Privacy Review approval. In addition to review flags, Meta provides internal guidance through an updated and expanding library of standards on deceptive design. A cross-functional group remains up-to-date on deceptive design developments and creates educational material available across the business. |
| Systems and Product Integrity | Private Accounts | Meta maintains a private account feature that operates on an ongoing basis to protect user accounts, including teens, from being open to users outside of their network. Users that are under 18 when signing up for an account, have the option to choose between a public or private account. However, Private is selected by default for users under 18 in the EU/UK and for users under 16 in the rest of the world. For users over 18 years old, the account is public by default and users can choose to make their account private at any time. |
| Systems and Product Integrity | Restrictions of Suspicious Adults | Meta maintains and operates Discovering and Connecting Prevention Tools. Once potentially suspicious adults are identified, Meta works to prevent them from discovering and connecting with teen accounts. These interventions can take a number of different forms, including messaging reachability limits for unconnected adults and teens, account recommendation filtering between risky adults and minors, and discovery and connection limits. |
| Systems and Product Integrity | Scam Account Score (SAS) | Meta maintains and develops an actor level signal, the Scam Account Score, that predicts the likelihood of an active account being a scammer on Instagram or Facebook. Teams across Meta can leverage the score to take action against accounts, such as restrictions or disable, depending on relevant thresholds. |
| Systems and Product Integrity | Security Checkup | Meta maintains security resources for users to access on Facebook and Instagram, respectively. Security Checkup enables the review and addition of security measures to Meta users' accounts. Facebook users can enable Security Checkup at their own discretion and review password, two-factor authentication (2FA), and login alert features. |
| Systems and Product Integrity | Snooze Function - Facebook | Meta maintains a mechanism for users to control their online experiences by selecting what type of content and activities they see from other users on Facebook. On Facebook, users can enable this feature and choose who to mute which stops the user from seeing the muted user's stories. Facebook users are further able to control their online experiences with the snooze feature which is a mechanism for users to stop seeing posts in their feed from a snoozed user's profile, page or group for 30 days. |
| Systems and Product Integrity | Strategic Network Disruption (SND) | Meta maintains the SND process which guides the strategic removal of actors and organisations involved in coordinated behaviours that facilitate real world harm via the Facebook family of apps and services. SND is done via i3's intelligence lifecycle which involves developing intelligence to understand the threat landscape, delivering targeted deference by proactively discovering and disrupting complex cases, enabling scaled problem reduction through institutional knowledge, and engaging the security and safety stakeholder community to build legitimacy and build programming. |
| Systems and Product Integrity | Tag and Mention Controls | Meta maintains Tag and Mention controls to allow users to choose whether they want everyone, only people they follow, or no one to be able to tag or mention them in a comment, caption or Story. This protects them from seeing unwanted behaviour. |

| | | |
|---|---|---|
| Systems and Product Integrity | Two-Factor Authentication | Meta maintains a 2FA process to increase security and decrease an account's vulnerability to hacks or unauthorised access on Meta platforms. There are several 2FA methods which send a notification to users at the time of login and allow users to utilise a second factor during the login process. A second factor can be a Short Messaging Service (SMS) sent with a code once or a security key that generates a code (a first factor is generally a password) which helps Meta authenticate a user's identity. |
| Systems and Product Integrity | User Consent Flow | Meta maintains a process for User Consent Flows which are product experiences that request permission to collect, use, or share user data. Each consent flow contains privacy elements including what data is being requested, who is requesting the data, the purpose for which the data is being used, what happens if the user does/does not consent, how the consent can be changed in the future, and a link to further information about the data processing. |
| Systems and Product Integrity | User Facing Privacy Controls | Meta maintains user facing privacy controls to permit users to manage their privacy settings on Meta's platforms. Users can navigate to the Privacy Settings page to adjust settings related to their privacy preferences. |
| Systems and Product Integrity | User Mental Wellbeing Support | Meta maintains a repository of in-app mental health resources, including 'The World Health Organisation Digital Stress Management Guide' and the WHO Health Alert chatbot, which provide easy-to-follow techniques designed to reduce stress and promote mental well-being. In addition, Meta partners with the Crisis Text Line to support suicide and self-injury crisis support. |
| Systems and Product Integrity | Why Am I Seeing This (WAIST) | Meta maintains the WAIST tool to provide users with information about why a particular ad appears in their feed or other surfaces across Meta's platforms. WAIST generates an explanation, based on machine learning models and algorithms to personalise a user's experience, to highlight the most relevant factors that contribute to the ad shown to a user. |

#### 6.2.1.3 Detection

Facebook has automated detection mechanisms in place to respond and take action against content that goes against our Facebook Community Standards, Meta's Advertising Standards, Commerce Policies and other applicable policies and guidelines. Of the violating content we take action on, our technology proactively detects and removes the vast majority across formats (e.g., image, text, and video) before anyone reports it, as demonstrated in our Quarterly Community Standards Enforcement Reports. Engineers, data scientists, and review teams work together to continuously update and improve our detection technology and better understand the effectiveness of our detection mechanisms. Meanwhile, our labelling strategy enables high accuracy and quality, which is measured through precision metrics.

The inherent trade-off of precision and recall that arises in any system means it is not possible for detection systems to be 100% accurate.[72] As we develop and improve our systems during model development, we take steps to respect the fundamental rights of users by ensuring that models are trained on the latest policies, which are developed based on feedback from civil society organisations, human rights defenders, marginalised groups, international organisations, Trusted Partners, investors, advertisers, and users. Stakeholder engagement also enables us to improve our detection of policy-violating content and better understand the impact of our services and the context of the diverse communities in which we operate around the globe.[73] Additionally, we continue to invest in technological capabilities, including training our models on safety and responsibility guidelines, content labelling, and people supporting detection efforts. This includes working with the EFCSN, in preparation for the 2024 EU Parliamentary Elections, to help train

[72] Precision tells us the percentage of cases that were genuinely true, out of the cases Meta labelled as true. Recall tells us the percentage of true cases Meta found, out of the possible true cases in the population.
[73] Bringing local context to our global standards | Transparency Center (fb.com)

fact-checkers across Europe on the best way to evaluate AI generated and digitally altered media, and on a media literacy campaign to raise public awareness of how to spot that type of content.[74]

The following table details the foundational controls that were assessed for this year's Systemic Risk Assessment as it relates to this control domain.

| Domain | Foundational Control Name | Foundational Control Description |
|---|---|---|
| Detection | Automated Proactive Detection Technology | Meta develops and maintains automated proactive detection tools for different content types across different surfaces. The automated proactive detection tools proactively detect and route potentially policy-violating content for automated decisions or manual review. |
| Detection | External Sharing of Blocklist | Meta develops and maintains processes and infrastructure to ingest and externally share signals related to harmful content on Meta's platforms. Depending on the programme, various APIs and processes are used to connect Meta and industry partners, NGOs, and trade groups, and governments to fight harm both on and off Meta. Sharing involves heavy privacy, legal, policy, and other stakeholder reviews, as well as a variety of contracts. |
| Detection | Fact Checkers | Meta partners with independent third-party fact-checkers in certain jurisdictions, who are certified through the non-partisan International Fact-Checking Network (IFCN), to address viral misinformation. Fact-checkers review a piece of content and rate its accuracy. This process occurs independently from Meta and may include, but is not limited to, calling sources, consulting public data, and authenticating images and videos. |
| Detection | Limits for Repeated Non-Violating Reports | Meta maintains mechanisms to identify users, entities, or individuals that frequently submit notices or complaints that are manifestly unfounded. Meta suspends the processing of these notices or the user's ability to submit notices for a reasonable period of time. A query is run on a quarterly basis to identify potential repeated non-violating reporters and findings are reviewed with Legal. If confirmed, limits are applied by tagging reporters via highlighting, which indicates to reviewers to ignore incoming reports from this reporter for a reasonable period of time. |
| Detection | Proactive Detection Quality and Governance | Meta develops and maintains quality and governance measures to ensure proactive detection systems are operating effectively and efficiently, while also minimising the risk of false positives, biases, and other potential negative consequences. Meta establishes policies and protocols, develops and maintains high-quality training, monitors and evaluates model performance, and conducts regular audits and reviews. |
| Detection | Proactive Detection Systems Implementation / Adaptation | Meta implements proactive detection systems through a combination of technology, data analysis and human expertise for relevant Problem Areas on Meta's platforms. Proactive detection systems are adapted by each Problem Area team to address concerns and nuances aligned with their respective policies by developing and maintaining high-quality training data, building and optimising detection models, and continuously improving and updating models. |
| Detection | Repeated Offender Identification - Manifestly Illegal Content (DSA-specific) | Meta maintains an internal tracking mechanism to identify individuals or entities that frequently post manifestly illegal content. This mechanism operates by tracking the number of times a user posts illegal content, and then on a case-by-case basis applying temporary suspension at the account-level based on set thresholds. |
| Detection | Repeated Offender Identification - Policy Violating | Meta maintains processes to identify repeat policy violating offenders on in-scope platforms. Meta identifies repeat offenders through a variety of detection methods and continuous review over organic, commerce, and paid content areas. |
| Detection | Rights Holders Reporting Tools | Meta maintains a variety of tools to support right holders to protect their content on Meta's platforms. The tools provide proactive and reactive levels of support, as well as the ability to request dedicated digital rights support, review channels, and enforcement methods. |

[74] https://about.fb.com/news/2024/02/how-meta-is-preparing-for-the-eus-2024-parliament-elections/

| | | |
|---|---|---|
| Detection | Takedown Requests - Specialised Platforms Intake System (SPInS) | Meta maintains the SPInS to provide external parties such as governments, regulators, and law enforcement, the ability to report and request the takedown of locally violating content. Meta administrators configure access or create new forms for external parties to submit relevant takedown requests. These submissions are then routed to the appropriate review teams. |
| Detection | Trending Events Tool | Meta launches a "Trending Event" during critical events that enables third-party fact-checkers to more easily review content related to the trending event. Using keyword detection, tags are applied to content that is likely related to the trending event, and made available within the Fact-Checking Product used by fact checkers so that they may prioritise reviews of that content. |
| Detection | Trusted Flaggers | Meta maintains Trusted Flaggers who are designated by the Digital Services Coordinator (DSC) and are prioritised through their onboarding to a dedicated reporting channel via SPInS. Trusted Flaggers submit reports of alleged illegal content within their designated area of expertise on an ad hoc basis through a single point of contact form which are prioritised for review. Meta has dashboards in place to track the quality of trusted flaggers with regards to submitting a significant number of insufficiently precise, inaccurate or inadequately substantiated notices, with the ability to report trusted flaggers' poor performance to the DSC. |
| Detection | User Data Disclosure Requests | Meta maintains the Law Enforcement Online Request System (LEORS), which is an external portal used by authorised law enforcement officials to submit and track requests for preservation and disclosure of user data. The portal enables Meta to review and respond to law enforcement requests for preservation and disclosure of user data in accordance with Meta's policies and applicable law. |
| Detection | User/Entity Initiated Reports - Escalated | Meta operates and maintains forms that allow users and non-users to escalate potentially policy-violating or illegal content that are high-risk. The forms allow users and non-users to submit reports pertaining to potential policy violating or illegal content that may require faster review turnaround times and higher priority. The report submissions contain relevant information, including substantiation and explanation of the content, the electronic information of the content, and the name and email address of the individual or entity submitting the report, as applicable. |
| Detection | User/Entity Initiated Reports - Policy Violating | Meta operates and maintains forms that allow users and non-users to report potentially policy violating content. The forms allow users and non-users to submit reports, either through in-app reporting options or through a form linked on the Facebook Home Page, containing relevant information, including substantiation and explanation of the content's illegality, the electronic location of the content, and the name and email address of the individual or entity submitting the report, as applicable. |

#### 6.2.1.4 Enforcement

Technology and review teams help Meta detect and review potentially violating content and accounts on Facebook. When potentially violating content is identified, either through our automated detection mechanisms, or user reporting, this triggers our enforcement systems and processes to take action. Our Facebook Community Standards, other applicable policies, and data sets of human decisions are used to train our machine learning models and human review teams to enforce against violating content on Facebook. We have a three-part approach to content enforcement: Remove, Reduce, and Inform. Content that violates our Community Standards and other applicable policies is removed in accordance with these policies; the distribution of content that may be problematic but does not call for removal under our content policies is reduced/demoted; and/or users are informed and provided with additional context to make informed decisions about the content they consume.[75] Meta also has mechanisms, such as the strike system and associated account restrictions, in place to hold users accountable for repeated violations of Facebook's Community Standards. Additionally, on occasion, content may be restricted in a particular country due to requests from regulatory authorities and pursuant to court orders or other reports of locally unlawful content once it has been reviewed and vetted by Meta.

[75] https://transparency.meta.com/enforcement

Over the past year, Meta has increased its investment in managing recidivism on the platform and managing threat actors that consistently seek new ways to circumvent detection and enforcement methods, including evolving use of emojis, slurs with intentional misspellings, and new terms. This includes continuing to develop our cross-platform enforcement capabilities, ongoing innovation and adaptation of our moderation systems, signals, and processes, building a process to enable enforcement against accounts that have perpetrated off-platform harm, and building classifiers that can detect if a bad actor that has already been removed from the platform is behind the creation of new assets.

The following table details the foundational controls that were assessed for this year's Systemic Risk Assessment as it relates to this control domain.

| Domain | Foundational Control Name | Foundational Control Description |
|---|---|---|
| Enforcement | Ad Review Process | Meta proactively reviews all advertisements before they are able to be published on Meta's platforms. Some ads are also selected for reactive review after they are published for various reasons, including if a user reports the ad as potentially violating. An ad is broken down into its various components, such as the title, images, and other text, to review if any portion of the ad violates Meta's Advertising Standards. The decisions made during these reviews determine whether ads are approved and go live, or if they are disapproved and returned to the advertiser. |
| Enforcement | Automated Enforcement Decisioning | Meta develops and maintains enforcement technologies that review content and behaviours identified as potentially violating Meta's policies and determine whether to take enforcement action. The enforcement technologies leverage training data, including previous decisions made and policy language, to determine if the content or behaviour is in violation of Meta's policies. If the content or behaviour is deemed policy violating, the enforcement technologies assign the appropriate enforcement action based on the severity of the violation and language in Meta's policies. If there is insufficient information to make an automated decision, the content or behaviour is sent for manual review by the human review teams. |
| Enforcement | Automated Enforcement Decisioning - Accuracy and Consistency | Meta maintains mechanisms to ensure its automated systems for detection are operating as intended. Meta has both quality assurance processes (e.g. training of models) and monitoring processes (e.g. metrics) to maintain and improve its quality of automated review processes. |
| Enforcement | Content and Entity Removal | Meta utilises enforcement technologies to remove individual accounts, complex objects, content, commercial listings, and ads found to be violating Meta's policies. Enforcement technologies leverage signals from automated and/or manual decisioning processes and remove violating entities from Meta's platforms. |
| Enforcement | Content and Entity Restriction | Meta utilises enforcement technologies to restrict content and/or entities on both paid and organic surfaces that violate Meta's policies across violation types. Each problem team may take different actions to restrict content and/or entity based on policies and protocols, including using enforcement technologies and leveraging signals from automated and/or manual decisioning processes. |
| Enforcement | Content Review Prioritisation | Meta maintains a content review prioritisation process to rank and prioritise content in order of importance for reviewers to take action on potentially policy violating content on Meta's platforms. The content review prioritisation process for enforcement considers the severity, virality, and likelihood of violation when determining what the human review team should prioritise. |
| Enforcement | Enforcement Actions - Accuracy and Consistency | Meta maintains mechanisms to ensure its automated systems for detection are operating as intended. Meta has both quality assurance processes (e.g. training of models) and monitoring processes (e.g., metrics) to maintain and improve its detection capabilities, the quality of automated review processes, and its enforcement mechanisms. |

| | | |
|---|---|---|
| Enforcement | Integrity Actions Platform (IAP) | Meta maintains a centralised actioning tool to execute various enforcement actions. The centralised actioning tool is integrated with decision processes and tools to intake appropriate enforcement actions to apply. |
| Enforcement | Language and Cultural Coverage | Meta incorporates language and cultural coverage into its products and services, such as translation tools, to support agnostic review for automated and manual review processes. Meta also develops and maintains knowledge management tooling and internal repositories to be utilised in its automated and manual review processes. |
| Enforcement | Manual Enforcement Decisioning | Meta has dedicated review teams for manually reviewing potentially violating content and entities identified through proactive detection systems and third party reports and determining whether the content and/or entities are in violation of Meta's policies. Human reviewers utilise review tools and guidelines to make enforcement decisions in accordance with Meta's policies. |
| Enforcement | Manual Enforcement Decisioning - Accuracy and Consistency | Meta maintains mechanisms to ensure the manual enforcement processes are operating as intended. Meta uses operational guidelines to establish standards, measurement systems to evaluate performance, and performance monitoring and continuous improvement processes to maintain performance. |
| Enforcement | Repeated Offender Enforcement - Manifestly Illegal Content (DSA-specific) | Meta applies enforcement actions on repeated offenders who frequently post manifestly illegal content by leveraging records of previous violations and determining appropriate action based on the number of violations and their severity. |
| Enforcement | Repeated Offender Enforcement - Policy Violating | Meta applies enforcement actions on repeated offenders who frequently violate Meta's policies by leveraging records of previous violations and determining appropriate action based on the number of violations and their severity. |

#### 6.2.1.5 Response and Notification

We have specific response actions for when crises arise that could impact how risks materialise on Meta's platforms, when there is an imminent threat to life, and when there is illegal activity suspected to enable matters to be escalated and addressed accordingly. Those processes include reporting to law enforcement and other authorities and providing relevant information via established processes in accordance with our Terms of Service and applicable laws. Working with law enforcement agencies and civil society organisations via dedicated reporting channels, is a key part of our approach to crisis response. A recent example of this has been in response to the assassination attempt on Slovak Prime Minister, Robert Fico. In response to this attack, Meta removed the alleged shooter's account under our Dangerous Organisations and Individuals Community Standards and notified the appropriate law enforcement bodies accordingly. Additionally, we provide notifications to users when crises arise and, in line with our usual processes, inform them on decisions taken related to their content and account.

Over the past year, Meta has been investing in improving consistency in how we inform users of actions taken against their account or content through the implementation of our statements of reasons, as required by the DSA.

The following table details the foundational controls that were assessed for this year's Systemic Risk Assessment as it relates to this control domain.

| Domain | Foundational Control Name | Foundational Control Description |
|---|---|---|

| | | |
|---|---|---|
| Notification | Enforcement Notification Processes | Meta maintains a notification process to notify users when Meta takes an enforcement action, including in response to a report of potentially illegal content or community standard violations, by sending notifications and statements of reason. The notification and statement of reason informs the recipient of the use of automation in decision-making and whether a review can be requested to appeal the decision. Where applicable, other potential options of redress, including out of court dispute settlement rights, are also provided to the user. Notifications are communicated to recipients in a timely manner, for both initial enforcement and relevant appeal decisions. |
| Notification | Mandatory User Data Preservation and Disclosure Obligations | Meta maintains processes to identify mandatory user data preservation and disclosure obligations. Meta reviews these processes to ensure compliance with user data preservation and disclosure obligations. |
| Notification | Policy Update User Notification | Meta notifies users in the event of a material change to the Terms of Service and has mechanisms that enable this notification (such as jewel notifications, and megaphone notifications, or email, as appropriate). Users can also see these changes in their notification settings within apps or Meta's online web interface. |
| Notification | Reporting Acknowledgement Processes | Meta maintains a report acknowledgement process to inform users that their reports have been received. For in-app reporting, users receive automatic in-app confirmations on their reports, and there is a display for the status of their requests. For out-of-app reporting using the contact form, users receive the acknowledgement via email. |
| Notification | Reporting Credible Threats to Authorised Government Authorities | Meta maintains processes for reporting credible threats to life or safety, when identified on Meta's platforms, to authorised government authorities via established processes on a case-by-case basis. Meta reports credible threats to life or safety in accordance with Meta's Terms of Service and applicable laws. |
| Notification | Statement of Reasons | Meta provides a statement of reasons in notifications sent to users after Meta takes an enforcement action that explains why the respective enforcement action was taken and informs users which of the Facebook Community Standards or other relevant terms or policies have been violated. Meta notifies users of their decisions, and provides a formal “statement of reasons” explaining any decision to remove or restrict visibility to a specific item of content and/or a user’s account and to inform users of what redress mechanisms are available to appeal the decision. |
| Response | Crisis Response Protocols | Meta implements crisis response protocols when an external event arises that could impact how risks materialise on Meta's platforms. Meta's response involves cross functional groups with varied areas of expertise that evaluate risks associated with the crisis and their potential impact, assess the effectiveness of controls to mitigate these risks, determine whether additional measures or actions are needed to address impacts, and implement measures to mitigate risks. Meta shares external and crisis event signals and information with governments through the appropriate channels as needed. |
| Response | Elections Readiness | Meta maintains elections readiness processes to ensure rapid support for the period immediately around critical events. XFN stakeholders collaborate to establish operational handbooks, policy clarifications, and response processes, and training for relevant external parties on how to leverage existing content reporting channels, and may also create new content reporting channels within the requirements set out by local law, as appropriate. |
| Response | Imminent Threat Escalation | Meta has specialised human reviewers that review information available to them to determine whether this gives rise to a suspicion of a threat to the life or safety of a person or persons. |
| Response | Targeted Search | Meta maintains Targeted Search, an ongoing programme for investigators with relevant clearance to search Facebook on an ad-hoc basis for measurement, intelligence gathering, and/or enforcement purposes, and that meet a defined search criteria. Targeted Searches may arise in response to external developments, enquiries or triggers, or identified on-platform trends, concerns or incidents. Targeted Search identifies content violating the Community Standards, or content which is equivalent to such specific content, and where there is a sense that searching for such content could be broadly justified (e.g., real world threat of harm). The purpose of this activity is to determine if enforcement is necessary and should material be found indicating improper content, appropriate enforcement action will take place. |

| | | |
|---|---|---|
| Response | Temporary Risk Mitigation Strategies | Meta deploys temporary processes or tools to reduce the likelihood that people see harmful content on its platforms during critical moments or in situations with elevated risk of violence or other severe human rights risks. Meta's teams of specialised cross-functional experts closely monitor trends on and off platform and investigate situations to determine whether and how best to respond. As appropriate, Meta may apply limited, proportionate, and time-bound measures that can be quickly implemented to address a specific, emerging risk. |
| Response | Voter Interference | Meta prohibits content conveying incorrect information regarding methods and logistics of voting (how to vote/when to vote), or instructions or intent to commit voter fraud. Meta aims to remove all instances of voter interference by identifying new violating claims, finding similar violating claims, and removing any violating claims. |

#### 6.2.1.6 User Rights and Recourse

In an effort to protect users' fundamental rights, including the right to free expression, and given the nuances of moderating content, Facebook has well-established processes to empower people to challenge the content decisions we make. Users may be given the option to appeal our decision to remove content after receiving a notification that their content has been removed. When someone appeals a decision, Meta reviews the content again and determines whether or not it follows our Community Standards, using a combination of human review and technology. After we review the content, we notify users whether their content has been reinstated or if we confirmed it did not follow our Community Standards.

If our original decision is not overturned or reversed, users may have the opportunity to appeal to the Oversight Board, an independent group of experts who protect free expression by making principled, independent decisions regarding the most difficult content moderation challenges on Facebook and by issuing recommendations on the relevant Meta content policies and operations. These recommendations are informed by international human rights standards and external perspectives from globally-representative free speech and human rights experts. Since its inception, the Oversight Board has reviewed 121 cases, and Meta has responded to 268 of the Oversight Board's recommendations.[76] These include decisions about the takedown of content regarding the Greek Elections that violated our policy on Dangerous Individuals and Organisations due to the use of symbols and praise for the Spartans party;[77] and posts in Poland targeting transgender people with violent speech advocating for members of this group to commit suicide.[78] Meta publishes regular updates on the Oversight Board to provide transparency into our responses to the Oversight Board's independent decisions about some of the most significant and difficult content moderation cases. These updates provide insights on the progress of our ongoing efforts and how Meta approaches decisions and recommendations from the Oversight Board.

Additionally, in line with Article 21 of the DSA, Meta has established the ability for users across EU member states to refer a decision to a relevant Out-of-Court Dispute Settlement Body. These processes allow people to let us know if they think we have made a mistake, which is essential to help us operate a fair system that respects users' voices.

The following table details the foundational controls that were assessed for this year's Systemic Risk Assessment as it relates to this control domain.

[76] https://transparency.meta.com/en-gb/oversight/overview
[77] https://transparency.meta.com/en-gb/oversight/oversight-board-cases/greek-2023-elections-campaign-cases
[78] https://www.oversightboard.com/decision/fb-uk2rus24/

| Domain | Foundational Control Name | Foundational Control Description |
|---|---|---|
| User Rights and Recourse | Ads Profiling | Meta maintains technical and organisational measures to ensure information with special protections (including Sensitive Categories of Data (SCD) are not used to show users personalised ads. Meta uses automated systems to prevent this information from being used for ads profiling and verifies the effectiveness of these measures. |
| User Rights and Recourse | Appeals Handling SOPs | Meta maintains SOPs that detail processes and activities to ensure complaints such as appeal processes are handled in a timely, non discriminatory, non arbitrary manner. SOPs include processes and criteria for consistent decision-making, training and human reviewers, user notifications, record-keeping, and turnaround times for handling of appeals. |
| User Rights and Recourse | Appeals Process - Actors | Meta maintains a process for actors to appeal Meta's enforcement decisions over content the actor created / posted and that was taken down. If the actor chooses to appeal Meta's original enforcement decision, the content is re-reviewed to determine if it violates Meta's content policies. Actors may appeal the original enforcement decision within six months and are able to provide further context for Meta to review. |
| User Rights and Recourse | Appeals Process - Reporters | Meta maintains a process for reporters to appeal Meta's enforcement decisions. If reporters choose to appeal Meta's original decision, the content is re-reviewed to determine if it violates Meta's policies or where applicable, local law. Reporters may appeal the original enforcement decision within six months and are able to provide further context / rationale for Meta to review. |
| User Rights and Recourse | Automated Review of Appeals | Meta utilises automated review systems to determine whether Meta's policies are in violation for appealed content. These systems perform tasks including recognising content in photos, videos, or text in accordance with Meta's policies to inform appropriate enforcement actions for appealed content. This excludes unlawful content appeals and other appeals that are regulatorily required to be reviewed manually. The automated review systems undergo governance measures including sampling and feedback loops to update the systems as needed to ensure they are operating effectively and efficiently. |
| User Rights and Recourse | Content and Account Restoration and Reinstatement | Meta maintains processes to restore content, reinstate suspended accounts, or reverse an enforcement action after a successful appeal. After an appeal is accepted, the applicable enforcement action is reversed by restoring removed content, removing warning labels, removing demotion actions, or reinstating accounts and account features. |
| User Rights and Recourse | Content Experience Personalisation | Meta offers users the ability to control how much of certain types of content (including sensitive or low quality content) they see in their feeds. Users can control if they wish to see more or less (depending on the content type) of this content in their feed. |
| User Rights and Recourse | Edit Rejected Ad | Meta provides capabilities in Ads Manager for advertisers to edit ads that have been rejected due to non-compliance with Meta's policies. Advertisers can edit an ad by reviewing the rejection details in Ads Manager and can edit the ad, request another review of the ad, and monitor the review status. Editing the ad also triggers an integrity re-review for violation of Meta's ads policies. |
| User Rights and Recourse | Edit Rejected Content | Meta provides in-app capabilities for users to edit content that has been rejected due to non-compliance with Meta's policies. Meta also provides an out-of-app process via the Contact Form. Users can edit their content by first reviewing the rejection reason details and then making relevant changes to the content. Users who submit via the Contact Form will receive back via email the final decision. The email decision option does not have a status update the way there is for in-app. |
| User Rights and Recourse | Newsworthy Inform Treatment | Meta maintains and operates a tool to label content as newsworthy. Upon escalation, Meta determines and may allow newsworthy content on their platforms for public awareness, even if it violates the Meta Community Standards. |

| | | |
|---|---|---|
| User Rights and Recourse | Non-Personalised Experience | Meta offers experiences not based on profiling across different product surfaces, which allows users in specific jurisdictions to experience those products without the use of personalised content ranking or recommendations. Users are able to access the non-profiling options through an easily-accessible entry point and are provided with a choice if they do not want to use recommender systems based on inferences that Meta's systems make based on the user's data. |
| User Rights and Recourse | Out-of-Court Dispute Settlement Process | Meta maintains a complaints handling process policy available on the Transparency Centre, which includes information on the right to submit a request to a certified out-of-court settlement body to resolve an issue. Users are also notified of out-of-court dispute settlement rights through in-app notifications. Out-of-court dispute settlement bodies have yet to be established and certified by regulatory authorities. |
| User Rights and Recourse | Quick Promotions | Meta updates, deploys, and monitors the Quick Promotion (QPs) feature on off-app channels (i.e., notifications and emails) and in-app Meta-to-User communication platform across the Family of Apps and Reality Labs. This feature can be used to drive growth/engagement for products, deliver critical legal notices, educate users, deliver surveys, or provide any other communication with users. The QP tool allows for the creation and management of QPs, including choosing the surface it will appear on, defining the eligibility rules and content, testing the QP, and deploying it in production and monitoring it. |

#### 6.2.1.7 External Awareness and Support Resources

At Meta, we continuously develop resources on how to keep our users safe on our platforms. We believe that user education can help users understand why particular content and behaviour are violating, so that they may reform their practices on our platforms. It can also help create a culture of respect and empathy in our services. Our library of tools and resources for improved online safety supports and reflects our Community Standards. Our goals for developing safety tools are for our users to learn how to stay safe on our platforms, keep their accounts secure, and protect their personal information.

Our Safety Centre has support resources around a variety of topics, including mental health and well-being, bullying and harassment, online child protection, intimate image abuse and sextortion. It also has tailored resources for communities that can be most impacted by Problem Areas, including women, youth, journalists, activists, public figures, vulnerable users, and the LGBTQ+ community. We also have resources for parents, educators, and law enforcement to support them in their respective roles. Additionally, we partner with external organisations and experts to provide educational materials, such as our Orygen #chatsafe for Educators that equips users to talk with minors about safety on social media (e.g., suicide and self harm), as well as our extensive materials listed in our Safety Centre for Parents to combat child safety concerns built in partnership with Thorn and their NoFiltr brand. Furthermore, we have a dedicated Facebook Help Centre that provides resources and information for users on how to stay safe on the service.

Over the last year, Meta has invested in updating its external resources, including updates to our Privacy Centre, updating Domestic Violence hotlines for EU member states, and generative AI labelling.

The following table details the foundational controls that were assessed for this year's Systemic Risk Assessment as it relates to this control domain.

| Domain | Foundational Control Name | Foundational Control Description |
|---|---|---|
| External Awareness and Support Resources | Ad Blueprint Training | Meta provides external e-Learning training through its Meta Blueprint programme for agencies, marketing partners, clients and Small Business Groups. Content is readily available on Meta Blueprint and enables people to gain foundational knowledge on Meta's advertising practices, including explaining how ads work and restrictions that apply. |

| | | |
|---|---|---|
| External Awareness and Support Resources | Branded Content Tool | Meta maintains the Branded Content Tool, which provides users with the opportunity to declare in real-time that their content contains commercial communications. If a user declares that their content contains commercial communication, this information is displayed prominently to other users using the “Paid Partnership” label on the commercial content. Any piece of content can be marked as Branded Content, and the user must indicate a relationship with the affiliated Brand/Company before being able to tag them. |
| External Awareness and Support Resources | Family Centre | Meta maintains the Family Centre, which provides users with resources, insights and expert guidance to help users support their family’s online experiences on Meta's apps and across the internet. The Family Centre has information about the tools across products, as well as the Education Hub which provides informational resources and tools across Meta's products. |
| External Awareness and Support Resources | Help Centre | Meta maintains the Help Centre, which provides users with a centralised hub for support across Meta products and services. The Help Centre provides links to product support, Meta shops, Meta help, specific help centres for Meta apps, and support for different types of users. |
| External Awareness and Support Resources | Parent's Guide | Meta maintains the Parent’s Guide on an ongoing basis to teach parents and guardians how to help their teens navigate its platforms. Parents can access these resources at their own discretion, which include conversation starters, information about safety and well-being tools and features, glossary of terms and more. Parent's Guide is issued in 40 languages. The guide provides details about how to manage privacy, manage interactions, comments, time, security, or supporting other people that need it due to eating disorders, among others.<br><br>Meta supports parents, caregivers and educators on an ongoing basis through Safety Centre with policies, resources and tools that help protect the safety and well-being of young people online. |
| External Awareness and Support Resources | Privacy Centre | Meta operates and maintains the Privacy Centre and its resources for users, including minors, to access and learn about common privacy topics. The Privacy Centre provides helpful information to users about common privacy topics, how Meta protects users' data, and what users can do to protect themselves. |
| External Awareness and Support Resources | Safety Centre | Meta maintains a Safety Centre to house documentation about Meta's approach to safety across Facebook, Instagram, and other products. The Safety Centre, available in over 60 languages, can be accessed by users and non-users and includes information, resources, and news to document Meta's approach toward safety for all users on their platforms. These resources can be accessed by users and non-users for specific safety topics and communities to empower individuals to obtain the support they need, specific to the area they are looking for. |
| External Awareness and Support Resources | Sponsored and Paid Partnerships Labels | Meta communicates to users that content being viewed is being promoted in an advertisement for paid ads. This is done through applying a prominent "sponsored" label on this content. |
| External Awareness and Support Resources | Suicide and Self-Injury Content Response | Meta operates and maintains a feature that sends resources to users who have posted content that is identified as being suicidal or self-harm related. Upon this content being identified, the user is directed to resources which include relevant helplines and a prompt to contact emergency services if the content indicates the user is in imminent harm. Resources are also available to users reporting such content; during the reporting flow, users are given an option to click "See Resources" and are taken to a Help Centre page with support resources, such as a suicide hotline. Additionally, there is an escalation pathway for this content to enter Credible Intent of Suicide triage review to determine if immediate assistance is required, in which case content may be communicated to first responders for intervention where allowed by local law. |
| External Awareness and Support Resources | Voter Empowerment Features | Meta maintains and launches election-specific product features to encourage voter participation in free and partly-free elections globally. Resources with information relevant to voting-age users are provided to connect them to authoritative information regarding elections. |

#### 6.2.1.8 Internal Training and Resources

To support the enforcement of our Community Standards and other policies, we provide extensive training to our employees and contingent workforce, including human reviewers. The training is aimed at enhancing knowledge and understanding of Meta's Community and Integrity Standards, moderation processes, Problem Area trends and insights, risk and compliance best practices and processes, and DSA and other content regulatory requirements. We also have routine processes in place to update reviewers on changes to policies or retrain reviewers when we detect enforcement issues. Additionally, Meta supports reviewers by proactively providing resilience, health, and wellness resources because they often work with content that may be objectionable and/or graphic.

The following table details the foundational controls that were assessed for this year's Systemic Risk Assessment as it relates to this control domain.

| Domain | Foundational Control Name | Foundational Control Description |
|---|---|---|
| Internal Training and Resources | Role Based Training | Meta develops and provides training to employees based on the role they undertake in order to equip them with the necessary knowledge and skills to perform their specific roles effectively. The exact nature of the training varies depending on their role, but includes onboarding, technical, and policy and compliance training, as needed. |

#### 6.2.1.9 Risk Assessment

Due to the global reach and scale of our services, the increase in online problems, and the exponential growth of user-generated content, Meta manages an increasingly complex and evolving integrity risk landscape. As such, Meta has designed and implemented risk assessment processes to identify, analyse, and mitigate integrity risks that could surface on our platforms, which includes our annual DSA SRA and our CIRAs.

As Meta continues to evolve and mature its compliance programme, we are committed to strengthening our risk management practices and executing risk assessments in a coordinated manner for ad-hoc risk assessments and annual requirements.

The following table details the foundational controls that were assessed for this year's Systemic Risk Assessment as it relates to this control domain.

| Domain | Foundational Control Name | Foundational Control Description |
|---|---|---|
| Risk Assessment | Crisis Response Risk Assessments | Meta reports on crisis response risk assessments to identify and assess external events impacting how risks materialise on Meta's platforms. Meta's crisis response reporting includes specifically internal crisis response protocols (IPOC, Global Response Operations Crisis Protocol, etc.) and using internal protocols to evaluate the impact of the risks associated with crisis response. Assessment of the risks associated with sensitive external events, such as wars, protests, and elections, may require controls to be updated to ensure the external events are managed appropriately. |
| Risk Assessment | Election Risk Assessments | Meta performs risk assessments across its platforms on a periodic basis during election periods / cycles. |
| Risk Assessment | Investigations | Meta conducts internal, governance and special investigations. Meta investigates holistic networks of abuse to detect current and continued on-platform activity and uses their capabilities to disrupt, deter, deny, and degrade adversarial harm. The investigations function consists of two pillars, Threat Mitigation and Threat Disruption. Additionally, investigative teams |

| | | |
|---|---|---|
| | | conduct deep dives to understand and mitigate trends of abusive accounts, conduct root cause analysis to improve manual and proactive detection, and manually enforce against violating content and accounts. |
| Risk Assessment | Salient Human Rights Risk Assessment | Meta manages a company-wide human rights risk assessment performed by a third party vendor to identify and document Meta's salient risks in order to provide recommendations to mitigate identified risks. The assessment is performed to develop recommendations to help mitigate risks and inform future strategy. In the event the salient risk assessment cannot be successfully completed, the Meta Human Rights Director will document the reasoning with appropriate sign-off. |

#### 6.2.1.10 Governance

At Meta, we have robust governance frameworks and processes in place across the three lines of defence for integrity matters, which includes structures for decision making, accountability, compliance, and oversight. Our layered approach to governance includes, but is not limited to, the following:

- **Enforcement Decisions (First Line of Defence - 1LOD):** Monitoring and measuring policy violating content and behaviour enforced against on Facebook through our accuracy monitoring processes;
- **Integrity System Changes (1LOD):** Managing changes to our Facebook integrity systems, including detection systems whereby multiple people are required to review and sign off on changes;
- **Integrity Reviews (1LOD):** Reviewing planned services and feature launches against defined integrity standards before they are launched to help ensure the appropriate safeguards are implemented;
- **Risk and Compliance Oversight (Second Line of Defence - 2LOD):** Executing routine risk and compliance processes to identify risks and assess the effectiveness of our controls carried out by the Governance, Risk and Compliance team with oversight from our DSA Compliance Office;
- **Internal Audit (Third Line of Defence - 3LOD):** Providing independent assurance that the risk, compliance, governance, and control processes and activities in place to manage integrity risks are designed and operating effectively;
- **External Audit:** Providing an objective independent examination of the risk, compliance, governance, and control processes and activities in place to manage DSA requirements and verify Meta is compliant; and
- **Management Body Oversight**: Our DSA Head of Compliance reports directly to the board of directors of MPIL, which provides Facebook to users within the European Union. The MPIL Board has an oversight role and is actively involved in decisions related to risk management.

As Meta evolves its compliance programme, it continues to strengthen and adapt its governance mechanisms and capabilities. Over the last year, Meta has focused on standing up risk decisioning forums, training the 1LOD and 2LOD on control ownership and audit requirements, and developing and scaling its operational practices within the 2LOD.

The following table details the foundational controls that were assessed for this year's Systemic Risk Assessment as it relates to this control domain.

| Domain | Foundational Control Name | Foundational Control Description |
|---|---|---|
| Governance | Human Rights Due Diligence | Meta conducts human rights due diligence to identify salient human rights considerations in the context of products, policies, and operations, using the United Nations Guiding Principles Framework of likelihood and severity and helps to create strategies to avoid, prevent and mitigate related potential risks on Meta's platforms. The Human Rights Team works with key stakeholders in policy and product to identify and support opportunities to embed the protection |

| | | |
|---|---|---|
| | | of human rights across processes, systems and activities, using a variety of due diligence methodologies, including, but not limited to, human rights impact assessments. |
| Governance | Local Law Content Restrictions | Meta has processes in place to identify, process, address, and restrict, where applicable, content that is reported to Meta as violating local law. Meta has reporting mechanisms to enable governments, regulators, courts, non-government entities, and members of the public to report illegal content. If the content does not violate Meta's policies, a review is conducted to validate the reported illegality and a separate review may be conducted to validate requests are in line with Meta's Corporate Human Rights Policy and commitments as a member of the GNI. If it is determined the content is illegal, Meta actions the content (e.g., by blocking the content in the relevant jurisdiction(s)) and notifies the reporter accordingly. |
| Governance | Local Law Review | Meta updates procedures for how they review locally illegal content in response to local law and scale these reviews. |
| Governance | Meta Privacy Programme and Governance | Meta maintains a programme for Meta's compliance to global regulatory privacy obligations. Meta develops frameworks to help place user privacy at the centre of Meta's products and services in accordance with our regulatory obligations. Meta partners with cross-functional teams to document current practices, scale safeguards, and identify and remediate gaps as needed. |
| Governance | Oversight Board | Meta maintains an independent content appeals process that is overseen by the Oversight Board. The Oversight Board is a global body of experts that provides independent review and decisions of Meta's product content decisions. The Oversight Board hears cases in instances where users disagree with the outcome of Meta's content decisions and have exhausted appeals or Meta directly submits the case for review. Decisions by the board are independent, binding (unless implementing the recommendations could violate the law), accessible to users, and transparent. Recommendations are not binding and are implemented at Meta's discretion. The only binding requirement for recommendations is for Meta to publicly disclose the action it takes in response. |
| Governance | Partnerships and Collaborations | Meta establishes partnerships with external groups to inform content policy development, enhance transparency around the policies and their implementation, and provide additional resources for Meta platform users. Meta's partnerships and collaborations help inform Meta's content policies, and they help Meta develop integrity protections and informational resources that are made available to users, in an effort to minimise harm and protect voice and well-being. Resources are shared with users via designated locations on Meta's website (e.g., the Newsroom, the Safety Centre, and the Community Standards page). |
| Governance | Response to Law Enforcement Requests for Preservation or Disclosure of User Data | Meta responds to requests from law enforcement for preservation or disclosure of user data in accordance with applicable laws and its Terms of Service. |
| Governance | Traffic Light System (TLS) / Geo-Blocking Policies | Meta maintains Traffic Light System Policies based off of the RCP Due Diligence Assessment to support Operations teams in making efficient review decisions based on identified content trends. The policies are created in collaboration with XFN to guide Operations teams in how to respond to takedown requests from regulators/courts/users from a particular country. The TLS policies create a scalable, responsive, and future proof review model to empower Operations teams to make content decisions which reduces escalations to Legal and RCP. |
| Governance | User Understanding | Meta maintains processes and systems to understand users' feedback on Meta's approaches for protecting users. These processes and systems include, but are not limited to, the measurement of user behaviour at scale, channels for users to provide feedback, and tools for data storage, analysis, and visualisation. |
| Governance | User/Entity Initiated Reports - Locally Illegal Content | Meta operates and maintains a service that allows users and / or entities to report potentially illegal content in their local jurisdiction. Users and entities can submit reports over potentially illegal content with relevant substantiation, such as the post, comment, profile, or other information over the content in question. The content is then reviewed to determine if it is in |

| | | |
|---|---|---|
| | | violation of Meta's policies and local laws. Users and / or entities are updated of report decisions via in-app, or email notifications. |

### 6.2.2 Detailed Risk Observations and Mitigating Measures

The following subsections detail some of the key trends identified in this assessment, the critical controls that are in place to manage these Problem Areas, including some of the ecosystem of controls that are detailed in Section 6.2.1, and limitations identified through the assessment period.

#### 6.2.2.1 Account Integrity and Authentic Identity

Authenticity is the cornerstone of our community. Meta believes that authenticity helps create a community where people are accountable to each other and on Meta's platforms in meaningful ways. We want to allow for the range of diverse ways that identity is expressed across our global community, while also addressing impersonation and misrepresentation.

Account Integrity and Authentic Identity is associated with the Deceptive and Misleading, Civic Discourse and Elections, and Protection of Minors Systemic Risk Areas in the DSA. This Problem Area relates to the risk of Facebook being used to deceive or mislead users through the use of compromised accounts, accounts and advertisers that are not authenticated or verified correctly, and threat actors returning to the service after being banned (e.g., recidivism). In some instances, threat actors build tools to create many accounts at once, known as “scripted abuse”. Additionally, with industry generative AI and evolving technology, it becomes more difficult to verify identity using government IDs as it has become easier to forge official identification.

**What are we doing to try to prevent and mitigate these risks?**

**2024 Trends:** During the assessment, it was identified that the high number of elections in the EU, including the EU Parliamentary Elections and generative AI could increase the volume of activity as it relates to this Problem Area. As highlighted in our Quarterly Adversarial Threat Report for the first quarter of 2024, this could lead to inauthentic amplification of authentic accounts or Pages of domestic politicians through likes, shares and comments to make them appear more popular than they were.[79] Additionally, it was identified that impersonation on the platform is reactive to geopolitical conflicts, possibilities of war, and other crisis events, which Meta cannot predict, and often sees threat actors piggybacking or taking advantage of such sensitive events. As a result, Meta has put in place dedicated election teams and mechanisms to combat the likely increase in adversarial behaviour and implemented several initiatives to manage the rapid expansion of generative AI, from labelling of generative AI content when we become aware that it’s generated and using synthetic data to improve classifier performance. Furthermore, we singled out the discrete risk of “Under Thirteen Year Olds on the Platform'' to account for users below the age of 13 creating accounts in violation of our Terms.

**Problem Area Mitigation Overview:** As detailed in Section 6.2.1, we have an extensive ecosystem of controls that work together to manage these Problem Area risks on Facebook. Specifically for Account Integrity and Authentic Identity, we have built a combination of automated and manual mechanisms to block and remove accounts that are used to persistently violate our Community Standards. We also implement mechanisms for searching and disabling accounts that have been dormant or inactive for a specific period of time. Meta has built classifiers that can help detect if a bad actor that has already been removed from the platform is behind the creation of new assets. Over the past year, Meta has improved its ability to detect scripted abuse accounts, and put in place net new actions to mitigate account takeover by introducing additional verifications, such as log-in challenges, age checks, and verification of contact point checks. To prevent any further activity from compromised accounts, Meta constantly works to improve its ability to identify compromised accounts closer to the point of compromise. For example, we work to refine classifiers to increase recall and minimise false positives. We also use a Fake Account Index and a Compromised Account Index as indicators of account integrity and authentic identity issues.

In the first quarter of 2024 alone, **Meta removed 631 million potentially fake accounts** on Facebook globally.[80] As account level removal is a severe action, whenever possible, we aim to give our community a chance to learn our rules and follow our Community Standards. We have also built many impersonation related user education interventions on the platform, including in-product user messages / warnings and links

[79] https://transparency.meta.com/metasecurity/threat-reporting
[80] https://transparency.meta.com/reports/community-standards-enforcement/fake-accounts/facebook/

to Help Centre educational resources in the event a user receives a friend request or message from an account that is new or with no mutual friends. Meta also has processes in place to verify a user's name on Facebook for accounts when a user has lost access, including sharing a valid ID with their name on it. In some cases, such as authorisation for advertisers running ads about social issues, elections or politics, there may be special identification requirements.

Furthermore, Meta has a **dedicated team** that **works on Account Integrity and Authentic Identity 24 hours a day, seven days a week**. Meta protects the integrity of user accounts through the use of security measures, such as two-factor authentication and sophisticated machine learning models applied at login and account notifications, such as login alerts after suspicious login attempts. Additionally, we provide users with an in-application mechanism that enables them to view devices that have recently logged into their account and remotely log out, as needed. Additionally, we require users to be at least 13 years old to sign up for Facebook and provide a reporting mechanism for users to report an account belonging to someone under 13. If an account is reported for someone who is reasonably verifiable as being under 13, the account will promptly be deleted.[81]

**Limitations**: Throughout our assessment, we identified areas for continued improvement as it relates to this Problem Area. Unless self-reported by the user, identifying compromised accounts is a challenge and more so if the user's device itself has been compromised. However, where we cannot easily verify the owner of the compromised account, we are able to use personal documents, such as a passport and identification card, to match the account profile and recover the account. Significant investment is being made in this area to identify and ingest more signals to identify compromised accounts. In addition, actors purposefully share Meta-banked CSAM content to get an account disabled and gain control of linked business accounts. With the rapid expansion of generative AI, certain risks may increase including impersonation of high profile individuals and content piracy. However, dedicated teams are working on identifying and enforcing against bots, document forgery, and related content, which includes staying aware of evolving AI and automation tools and partnering with external organisations, in order to keep pace with adversarial behaviour trends.

#### 6.2.2.2 Adult Sexual Exploitation and Adult Nudity

Meta recognises that its platforms may be used as a place to educate users on and draw attention to sexual violence and exploitation, and where such intent is clear, we make allowances for the content. However, to protect human dignity, right to privacy and family life, and the rights of the child, we default to removing sexual imagery unless it is posted as a form of protest, for educational or medical reasons, or to raise awareness about a cause, as published in our Adult Nudity and Sexual Activity Community Standards.

Adult Sexual Exploitation and Adult Nudity is associated with the Gender-based Violence, Protection of Minors, and Fundamental Rights Systemic Risk Areas in the DSA. This Problem Area relates to the risk of Facebook being used by threat actors to sexually exploit adult users, including non-consensual sexual touching, necrophilia, or forced stripping, sextortion, non-consensual intimate imagery, creepshots, or rape threats. This also potentially includes Facebook being used by threat actors to depict or promote imagery or ads of real nude adults, adult sexual activity, or extended audio of adult sexual activity. In some instances, threat actors use obfuscation techniques, embed brief violating videos within longer ones, and use benign/permitted nudity tags (e.g., breastfeeding) to evade detection systems. In some instances, threat actors use depictions of adult nudity and sexual activity as a means to bait users for other purposes or goals. Additionally, this risk potentially includes the adverse impact on users' fundamental rights, specifically the right to human dignity and the right to private and family life as enshrined in the EU Charter.

**What are we doing to try to prevent and mitigate these risks?**

**2024 Trends:** During the assessment, there were no new trends identified that could potentially change the inherent risk exposure associated with this Problem Area.

**Problem Area Mitigation Overview:** As detailed in Section 6.2.1, we have an extensive ecosystem of controls that work together to manage

[81] https://www.facebook.com/help/157793540954833?helpref=related_articles

these Problem Area risks on Facebook. Specifically for Adult Sexual Exploitation and Adult Nudity, we engage with external parties throughout our policy development process which influences the development of our policies. Additionally, we strive to balance voice and safety within our policies as it relates to detection and enforcement.

As it relates to our detection controls, we are aware of patterns associated with sharing of non-consensual intimate imagery (NCII) and non-consensual sexual touch (NCST) as well as sextortion and have automated systems that detect and remove these accounts at scale. Additionally, over the last year we increased our investments in enforcement on recidivism, increased the number of human reviewers to address risks across this Problem Area, and improved the effectiveness of manual reviews. When we take enforcement action against a user for sextortion, the network of accounts and devices owned by the threat actor are taken down and we make it difficult for them to create new accounts. We also encourage people to report NCII content. **Our teams review NCII reports 24 hours, 7 days a week in more than 70 languages globally.**[82] In the first quarter of 2024, globally, we have seen an increase in actioned content for adult nudity and sexual activity on Facebook after rectifying a loophole by which threat actors were previously able to share links for violating content, and **we took action against 39.4 million pieces globally** of potential adult nudity and sexual activity content on Facebook, with 94.70% being identified by us before users reported it.[83] Additionally, we have developed more than **50 tools and features to help support the safety of teens and families** across our apps, including supervision tools for parents and guardians and specific education and resources about sextortion in our Safety Centre.[84]

**Limitations:** Throughout our assessment, we identified areas for continued improvement as it relates to detection and enforcement. Adult nudity and sexual activity remains a highly adversarial and profit motivated space, particularly as threat actors develop new ways to circumvent detection and enforcement such as utilising multiple accounts, creating fake profiles, signposting where threat actors lead users to harmful external sites, or embedding violating video clips lasting only a fraction of a second long within an otherwise benign video. Specifically regarding NCII, it may be challenging for our detection tools to proactively identify intent or consent. Therefore, we may rely more on user reporting and human review. Similarly, we rely more heavily on user reporting and human review to identify and enforce against policy-violating content that occurs more sporadically, such as necrophilia. Additionally, due to our severity prioritisation approach, some lower severity risks could be auto-closed without human review, such as adult nudity. Meta is continuously working to improve and enhance our detection and enforcement capabilities to implement further mitigations on Facebook.

#### 6.2.2.3 Bullying and Harassment

Meta prohibits bullying and harassment as it can create unsafe and disrespectful environments on our platforms. We remove content that is meant to degrade or shame private individuals and remove targeted mass harassment when there is a heightened risk of real-world harm, as published in our Bullying and Harassment Community Standards.

Bullying and Harassment is associated with the Civic Discourse and Elections, Gender-based Violence, and Protection of Minors Systemic Risk Areas in the DSA. This Problem Area refers to the risk of Meta's systems being used to promote content that degrades or shames users or to make repeated contact with a user that is unwanted, such as cyberbullying, threats of harm, mass harassment, and sexual harassment. We recognise that bullying and harassment can have disproportionate effects on minors' well-being and mental health, which is why we provide heightened protections for users between the ages of 13 and 18. We also recognise that the LGBTQIA+ community as well as public figures like female politicians, especially female politicians of colour, are targets of bullying and harassment at a disproportionate rate. This can cause silencing of the LGBTQIA+ community and women's voices and intimidation and/or fear for their safety. This risk can manifest itself on Facebook when adversarial networks work together to engage in repetitive behaviour, which is challenging to manage as brigading and coordination of mass harassment happens on and off the service and can take various forms, making it difficult to identify.

We also recognise that becoming a public figure isn't always a choice, and that this fame can increase the risk of bullying and harassment — particularly if the person comes from an underrepresented community.

[82] https://about.meta.com/actions/safety/topics/bullying-harassment/ncii/
[83] https://transparency.meta.com/reports/community-standards-enforcement/adult-nudity-and-sexual-activity/facebook/
[84] https://www.meta.com/help/policies/safety/tools-support-teens-parents/

Consistent with the commitments made in our Corporate Human Rights Policy, we now offer more protections for public figures like journalists and human rights defenders who have become famous involuntarily or because of their work. These groups now have protections from harmful content, for example content that ranks their physical looks, as other involuntary public figures do. The full list of protections for public figures, including involuntary public figures, can be found in our Community Standards.

**What are we doing to try to prevent and mitigate these risks?**

**2024 Trends:** During the assessment, we identified that bullying and harassment risks may disproportionately impact minors and thus could potentially increase the inherent risk exposure associated with this Problem Area. As a result, Meta has developed processes and tools specifically targeting minor protection, improved automated detection of content related to bullying and harassment, and bullying prevention resources. Additionally, it was identified that the high number of elections in the EU, including the EU Parliamentary Elections, could increase the risk of bullying and harassment against political public figures on the platform. As a result, Meta put in place dedicated election teams to combat the likely increase in adversarial behaviour.

**Problem Area Mitigation Overview:** As detailed in Section 6.2.1, we have an extensive ecosystem of controls that work together to manage these Problem Area risks on Facebook. Specifically for Bullying and Harassment, given the potential disproportionate impact on minors' and vulnerable users' well-being and mental health, we provide heightened protection for users under the age of 18, including protection from allegations about criminal or illegal behaviour and videos of physical bullying against minors, in addition to all other protections provided.

We have made strong investments in classifier performance through continuous improvement of our models, extensive experimental periods to test accuracy and stability, and settings to take automated enforcement actions. Additionally, we are continuing to build new features to improve detection of new content types, such as generative AI. This is exemplified by the fact that in the first quarter of 2024, globally, **we actioned 7.9 million pieces of bullying and harassment content on Facebook globally, with 85.6% detected proactively** before being reported by users.[85] Furthermore, we also include a link on nearly every piece of content for reporting abuse, bullying and harassment, and other issues and encourage self reporting as it helps us understand when a person feels bullied or harassed.[86]

We continue to provide many options for users to control their experiences on Facebook and limit unwanted interactions with other users to prevent bullying and harassment. These include blocking other users, restricting other users' ability to comment on their posts, and restricting visibility of posts and profile information for specific users.

There are also many teams at Meta, including the policy and safety teams, that routinely work with external parties to understand new trends and behaviours to help improve Meta's policies and resources. For example, **after working with over 400 women's safety organisations and experts, we established Meta's Global Women's Safety Expert Advisors to advance the safety of women online**. Additionally, we work with bullying prevention experts, such as the Diana Award Anti-Bullying Ambassador Programme, International Bullying Prevention Association, and Cyberbullying Research Centre to stay informed on bullying trends, and maintain our bullying prevention resources, such as Bullying Prevention Tips for Youth, Online Bullying Prevention Tips for Parents, and Managing Bullying and Harassment in Facebook Communities. We also engage with multiple governments during rollouts of EU hate speech tests, to collect feedback for improvements regarding perception of hostile speech mitigation measures on our platforms and services.

**Limitations:** Throughout our assessment, we identified areas for continued improvement as it relates to detection and enforcement. Bullying and harassment continues to evolve with the landscape of social interaction and digital connection. Threat actors continue to explore ways to circumvent detection and enforcement, such as using emojis, intentional misspellings or symbols. Additionally, with bullying and harassment being highly individualised and context-dependent, it often requires moderators to understand the relationship between users, the meaning behind content and behaviour, and the nuances of language and regional context to avoid over-enforcement of content moderation in benign scenarios. As cultural context changes and new generations emerge, new trends, terms and phrases that are not yet able to be flagged can emerge as well. Meta is continually keeping on top of culture shifts and adapting mechanisms to account for changing landscapes. Additionally, there are no automated detection or classifiers to detect bullying and harassment violations in ads. Therefore, we may rely more on user reporting and human review. However, new features are being built to improve detection of new types of content.

#### 6.2.2.4 Child Sexual Exploitation, Abuse and Nudity

Meta does not allow content or activity that sexually exploits or endangers children, as published in our Child Sexual Exploitation, Abuse and Nudity Policy Community Standards.

[85] https://transparency.meta.com/policies/community-standards/bullying-harassment/
[86] https://about.meta.com/actions/safety/topics/safety-basics/tools/stay-safe

Child Sexual Exploitation, Abuse and Nudity is associated with the Protection of Minors and Fundamental Rights Systemic Risk Areas. This Problem Area relates to the risk of Facebook being used to promote or disseminate content or activity of non-sexual child abuse, child nudity, child endangerment, child sexual exploitation, child sexual abuse materials (CSAM), including self-generated CSAM and solicitation of CSAM, sexualisation of minors, and exploitative intimate imagery and sextortion of minors. This risk also includes inappropriate interactions with minors and the adverse impact on minors' fundamental rights, specifically the respect for the rights of the child and the right to human dignity as enshrined in the EU Charter.

Additionally, the behaviour of threat actors is currently evolving, including through intentional manipulation by threat actors to persistently adapt to evade detection, including using implicit signals like keywords and hashtags (e.g., "chicken soup"), moving conversations off the service to take advantage of minors, and returning to the service through new accounts despite being blocked. Additionally, some minors may attempt to circumvent Facebook's age gating processes by misrepresenting their age, which makes it challenging to protect these users from certain types of content.

**What are we doing to try to prevent and mitigate these risks?**

**2024 Trends:** For this year's risk assessment, Meta continued to actively monitor the potential for generative AI to impact risks in this space as the use of generative AI advances. Several investments on this front include and committing to Safety by Design principles from Thorn and All Tech is Human to proactively address child safety risks.

**Problem Area Mitigation Overview:** As detailed in Section 6.2.1, we have an extensive ecosystem of controls that work together to manage these Problem Area risks on Facebook. Specifically for Child Sexual Exploitation, Abuse and Nudity, we are continuously evaluating our controls as it relates to this Problem Area and updating and creating new child protection tools and mechanisms, as appropriate. Significant work has been done to improve our controls, including key word interstitial updates. Furthermore, in the third quarter of 2023, we launched a new mechanism to proactively find, disable, and remove Facebook accounts if they exhibit a number of signals which we monitor for potential suspicious behaviour. We have expanded our work to detect and remove networks that violate our policies with our account enforcement propagation efforts. **Between 2020 and 2023, our teams disrupted 37 abusive networks and removed nearly 200,000 accounts associated with abusive networks globally.**[87] Additionally, if we take action against an account on Facebook for violations related to child sexual exploitation, abuse, and nudity, we look for linked accounts and devices and take them down.

Additionally, in further efforts to strengthen our protections for young people, Meta introduced controls to restrict Recommendations and Discovery features and expand search intervention methods. When it comes to Recommendations, we have systems that proactively find, remove, or refrain from suggesting content across most surfaces, Groups and Pages, and we have improved these systems by combining them and expanding their capabilities. We also expanded the existing list of child safety related terms, phrases and emojis for our systems to find. We have many sources for these terms, including non-profits and experts in online safety, our specialist child safety teams who investigate predatory networks to understand the language they use, and our own technology which finds misspellings or spelling variations of these terms. We have also introduced novel techniques to find new search terms. For example, we are using machine learning technology to find relationships between terms that we already know could be harmful or that break our rules and other terms used at the same time. These could be terms searched for in the same session as violating terms, or other hashtags used in a caption that contains a violating hashtag. We combined our systems so that as new terms are added to our central list, they will be actioned across Facebook and Instagram simultaneously.[88]

Our policies against Child Sexual Exploitation, Abuse and Nudity apply to both content and on-platform activities beyond content. We enforce these policies in two main ways; via content reviewers responding to user reports and proactively through automated systems. Meta deploys classifiers that can proactively identify violating or potentially violating content. We also leverage technologies such as Google's classifier, in order to prioritise content for reviewers. In addition to removing content that violates our policies, our automated systems consider a broad spectrum of on-platform activity signals to identify and disable accounts engaged in violating activity, and to help prevent potentially unwanted or unsafe interactions. We may restrict the visibility and discoverability between adult and teen accounts, as well as removing the ability to initiate new connections (e.g., via friending/following). We may also remove access to products and features (e.g., the ability to message certain other users) for adults based on their interactions with other accounts, searches for or interactions with violating

[87] https://about.fb.com/news/2023/12/combating-online-predators/
[88] https://about.fb.com/news/2023/12/combating-online-predators/

content, or membership in communities (e.g., Groups) we have removed for violating our policies. In the first quarter of 2024, globally, we have seen an increase in actioned content for child sexual exploitation due to improvements and actioned **14.4 million pieces of potential child sexual exploitation content on Facebook, with 94.3% being identified by us before users reported it.**[89]

Additionally, Meta engages with civil society organisations, academics, child safety experts, NGOs and other thought leaders to gather knowledge and experience as we develop our content policies. Our efforts include developing industry best practices, building and sharing technology to fight online child exploitation, and supporting victim services. The Take It Down platform, which allows the hashes of young people's intimate images to be shared with Meta, has enabled more effective automated detection and enforcement of this type of content.[90] Meta is also a founding member of the Technology Coalition where we collaborate with leading internet safety organisations from around the world to develop industry best practices, build and share technology to fight online child exploitation, and support victim services. We work with experts especially focused on child safety to build a collection of resources that foster conversations between parents, caregivers, and teens as they navigate and develop online safety habits in our Family Centre Education Hub. Furthermore, we collaborate with our external trusted third parties, including the Facebook Safety Advisory Board, law enforcement and governments, to discuss and improve our policies and enforcement around online safety issues, especially with regard to children, and to share information, resources, and trends regarding child sexual exploitation, abuse, and nudity.

**Limitations:** Throughout our assessment, we identified areas for continued improvement as it relates to detection. Due to the time-bound nature of live streamed and co-broadcasting content, our human reviewers may not be able to review all CSAM cases escalated for manual review in real-time. However, such content may be reviewed once the live stream has ended. We have also made improvements on ranking to help with the timely review and closure of CSAM cases. Lastly, some abuse types can be more difficult to detect than others, such as non-sexual child abuse, so managing this type of content may rely more on user reporting and human review. CSAM continues to be a highly adversarial space where threat actors identify new ways to evade detection and enforcement, including posting links to off-platform sites that could contain policy-violating content, making it difficult to detect. Meta is continuously working to improve and enhance our capabilities to implement further mitigations on Facebook.

#### 6.2.2.5 Coordinating Harm and Promoting Crime

In an effort to prevent offline harm and copycat behaviours, we prohibit the facilitation, organisation, promotion or admission to certain criminal or harmful activities targeted at people, businesses, property or animals, as published in our Coordinating Harm and Promoting Crime Community Standards. While discussions and advocacy regarding the legality of such activities are permitted, coordinating or advocating for harm is not.

Coordinating Harm and Promoting Crime is associated with the Public Security and Civic Discourse and Elections Systemic Risk Areas in the DSA. This Problem Area relates to the risk of Facebook being used to facilitate, organise, promote, or call for voter or census fraud, illegal participation in elections, coordinated interference in elections, violence against people, potentially including high-risk viral challenges and violence against property, including vandalism of state property.

**What are we doing to try to prevent and mitigate these risks?**

**2024 Trends:** During the assessment, it was identified that the high number of elections in the EU, including the EU Parliamentary Elections could increase the risk of voter and/or census fraud and coordinated interference in elections. As a result, Meta put in place dedicated election teams to combat the likely increase in adversarial behaviour.

**Problem Area Mitigation Overview:** As detailed in Section 6.2.1, we have an extensive ecosystem of controls that work together to manage these Problem Area risks on Facebook. Specifically for Coordinating Harm and Promoting Crime, we have a global workforce of content reviewers in the markets that Meta operates in, including in the European Union. These reviewers review content against the Coordinating Harm and Promoting Crime Community Standards and other applicable policies and guidelines with expert or native understanding of the

89 https://transparency.meta.com/reports/community-standards-enforcement/child-nudity-and-sexual-exploitation/facebook/

90 https://about.meta.com/actions/safety/topics/bullying-harassment/ncii

language the content was posted in. This helps ensure the policy is correctly enforced and accounts for cultural and linguistic nuances.

Additionally, our security teams work to dismantle manipulation campaigns and identify emerging threats, including investigating and taking down coordinated networks of inauthentic accounts, Pages and Groups. Our team leverages image banks to detect content in regions with high-risk elections to combat voter suppression, and we implement additional processes to perform a secondary, holistic review of politically viral content. We also have a Violence and Harm Team that actively operates an ongoing process for proposing changes to our Market-specific Implicit Threat Terms List, which includes market-specific idioms or proxy language that enables us to identify escalation cases. As described in Meta's Adversarial Threat Report for the first quarter of 2024, our teams have continued research on Doppelganger, sharing information and insights with industry peers and relevant governments and engage in daily efforts to find and block Doppelgangers' attempts to acquire new accounts, run ads, and share links before these are ever shared on our apps, as well as other coordinated interference clusters. We have responded to a major shift in tactics on our platform by Doppelgangers and focused on reducing Doppelgangers' ability to seed links directly or redirect URLs. Since April 2024, our research shows the operators have stopped attempting to share links altogether.[91]

**Limitations:** Throughout our assessment, we identified areas for continued improvement as it relates to detecting content that is coordinating harm and promoting crime. Nuanced language and evolving context pose a challenge to moderating content for this Problem Area as users are constantly coming up with new themes and trends and cultural context is continuously changing. Additionally, regional culture barriers make it difficult to gain an adequate understanding of all regional or cultural nuances, such as slang or dialects from a particular neighbourhood, city, or region, when determining what is classified as coordinating harm. However, our human reviewers receive in-depth training and often specialise in certain policy areas and regions in order to account for those linguistic and cultural nuances. Furthermore, while we detect for certain instances of coordinated harm under our hostile speech classifiers and banked terms, aside from voter interference, there is no automated detection for coordinating harm and promoting crime violations in place.

#### 6.2.2.6 Dangerous Organisations and Individuals

In an effort to prevent and disrupt real-world harm, we do not allow organisations or individuals that proclaim a violent mission or are engaged in violence to have a presence on Facebook and remove any content that glorifies, supports or represents individuals or groups engaging in terrorist activity or organised hate, as published in our Dangerous Organisations and Individuals Community Standards.

Dangerous Organisations and Individuals is associated with the Public Security and Civic Discourse and Elections Systemic Risk Areas in the DSA. This Problem Area relates to the risk of Facebook being used by terrorists, hate and/or criminal organisations, militarised social movements, violence-inducing conspiracy networks, groups promoting hatred, or violent non-state actors to advocate for and facilitate violence. This risk also includes the risk of Facebook being used by terrorists to recruit and/or radicalise users and the use of Live by such actors or entities to disseminate content in association with a terrorist act.

**What are we doing to try to prevent and mitigate these risks?**

**2024 Trends:** During the assessment, it was identified that the high number of elections in the EU, including the EU Parliamentary Elections could increase the inherent volume of activity as it relates to this Problem Area, specifically the growth in hate groups. As a result, Meta has put in place dedicated election teams to combat the likely increase in adversarial behaviour.

**Problem Area Mitigation Overview:** As detailed in Section 6.2.1, we have an extensive ecosystem of controls that work together to manage these Problem Area risks on Facebook. Specifically for Dangerous Organisations and Individuals, Meta has robust policies in place that are routinely reviewed and updated to adapt to the current climate. For example, at the end of 2023, we refined our Dangerous Organisations and Individuals Community Standards, including more comprehensive definitions of dangerous organisation types and tiers and prohibiting "glorification" of the violence and hate of dangerous organisations and individuals. We also updated our delisting process, which now provides more detailed and comprehensive criteria across all types of dangerous organisations and individuals that must be satisfied for a dangerous organisation or individual to be considered for delisting. We also updated our Dangerous Organisations and Individuals designations from 3 Tiers to 2 Tiers; with Tier 1 focusing on entities that engage in serious offline harms and Tier 2 that engage in violence

[91] https://transparency.meta.com/metasecurity/threat-reporting

against state or military actors in an armed conflict but do not intentionally target civilians. Tier 1 continues to result in the most extensive enforcement because of their most direct ties to offline harm. However, views of violating content that contains terrorism are very infrequent, and we remove much of this content before people see it. **In the first quarter of 2024, we actioned 8.4 million pieces of content related to terrorism on Facebook globally, with 99.3% being identified by us before users reported it.**[92]

We utilise advanced detection technology to identify dangerous or violent actors on Facebook and take an actor-centred enforcement approach. We perform investigations and use intelligence to identify actors and objects that are connected from a network with our account enforcement propagation efforts. Once a threat actor is identified with a sufficient degree of certainty, we use SND to take action against the identified network, which includes all accounts and devices owned by the threat actor, in an effort to also combat recidivism. Additionally, we have started engaging with Trusted Flaggers to triage allegedly illegal content posted by users in accordance with Article 22 of the DSA. Additionally, we have developed features to inform users of dangerous organisations and individuals on our platforms. For example, our Search Intercept and Search Redirect features are designed to redirect a problematic search to the relevant help resources and/or display warning screens that indicate the problematic nature of their search. To help prevent glorification, support or representation by individuals or groups engaging in terrorist activity or organised hate, we make hashtags associated with designated dangerous organisations or individuals unsearchable.

To help tackle dangerous organisations and individuals more broadly, we work closely with external organisations and authorities, including law enforcement. We also hold and collaborate in forums, such as the Global Internet Forum to Counter Terrorism (GIFCT). Through GIFCT, we collaborate with industry peers via signal sharing and a hash matching service, which includes hashtags and URLs that are sourced internally and externally, reviewed, and added to a shared bank of signals. This enables continuous process improvement of automated systems and updates to our Dangerous Organisations and Individuals Database. We are also in dialogue with the EU Internet Forum regarding programmes we can develop through partnership with the EU as well as the implementation of the EU Crisis Protocol. [93]

**Limitations:** Throughout our assessment, we identified areas for improvement as it relates to this Problem Area which include detecting and enforcing content related to dangerous organisations and individuals. This Problem Area is a highly adversarial space in which dangerous organisations and individuals constantly discover new ways to evade detection and enforcement on Meta's systems. There is also a challenge that something we categorise as a dangerous organisation or individual, such as far right movement, can be allowed or legal in another country. This makes moderating such content more challenging. And lastly, dangerous organisations and individuals that have been identified and banned can resurface and utilise our platforms before our teams and investigators detect and remove them. However, our dangerous organisations and individuals team are having continuous dialogues with other teams, such as Intelligence and Investigations, to help surface emerging risks and drive improvements in detection and enforcement and/or policy development.

#### 6.2.2.7 Discrimination / Discriminatory Actions

At Meta, we integrate anti-discrimination into our principles and operations, which are driven by our holistic approach to fundamental rights across all our Problem Areas. In addition to the anti-discrimination commitments in our Human Rights Policy, Meta also has a specific team focused on implementing civil rights principles globally, centring non-discrimination, justice, and fairness principles in this work. The team analyses civil rights risks, including discrimination, related to Meta products and policies. This team also includes fairness and non-discrimination requirements in its policy and product due diligence. We strive to protect our users and especially communities associated with protected characteristics, against hateful content. Additionally, we require advertisers to comply with applicable laws that prohibit discrimination, as published in our Community Standards.

Discrimination / Discriminatory Actions is associated with the Civic Discourse and Elections and Fundamental Rights Systemic Risk Areas in the DSA. This Problem Area relates to the risk of Facebook being used to adversely impact users' fundamental right to nondiscrimination as enshrined in the EU Charter. This can manifest on Facebook when threat actors engage in hate speech and violence and incitement. This is challenging to manage due to hateful terms constantly changing and threat actors deliberately circumventing measures.

[92] https://transparency.meta.com/reports/community-standards-enforcement/dangerous-organizations/facebook/
[93] The EU Internet Forum is an initiative of the European Commission that gathers member state representatives and selected members of the industry.

**What are we doing to try to prevent and mitigate these risks?**

**2024 Trends:** During the assessment, it was identified that the conflicts in adjacent regions and the high number of elections in the EU, including the EU Parliamentary Elections could increase the risk of discrimination on the platform. As a result, Meta has put in place dedicated election teams to combat the likely increase in adversarial behaviour.

**Problem Area Mitigation Overview:** As detailed in Section 6.2.1, we have an extensive ecosystem of controls that work together to manage these Problem Area risks on Facebook. Specifically for Discrimination / Discriminatory Actions, Meta is committed to upholding the fundamental right of non-discrimination by taking a holistic approach and integrating this principle across our operations and within various community standards. We employ dedicated Human Rights and Civil Rights Teams who collaborate with other teams across the company to build more equitable policies, products, and practices for Meta's communities. Their work is centred on non-discrimination, justice, and fairness and includes implementing our Corporate Human Rights Policy, addressing potential biases in artificial intelligence systems, and helping to inform the development of new technologies. The Civil Rights Team provides disparate treatment and impact analysis of policies, products, and practices at various stages in development, along with mitigations. We also have an Inclusive Product Council, which acts as a diverse consultative body that provides live experience feedback to product teams and advises on product development.

When reviewing content against Meta's policies, human reviewers work with region-specific teams to understand region-specific risks or trends and make sure regional, cultural, and linguistic nuances are considered when moderating content. We also maintain and leverage a Market-specific Slurs List, consisting of inherently offensive words that are used as an insult towards protected characteristics. This list is used to identify and flag slurs on our platforms and to train our classifiers to identify violating content. Additionally, we have updated our processes by shifting away from weighting based on certain signals like the number of comments and shares when ranking content for recommendations.

Additionally, the majority of full-time employees are required to take a Civil Rights and Meta Technologies training to help identify civil rights risks, including discrimination. This training helps employees, including certain policy and product staff, to understand what civil rights concepts and principles are, how to identify issues and concerns in their work, and where to go for help with issues or questions. The Civil Rights Team enhances this training module as needed and also engages in internal workshops and analysis to help teams build with civil rights in mind. Furthermore, Meta launched a support interface for managed partners and creators, with links to Meta's Help Centre and contact forms.

**Limitations:** Throughout our assessment, we identified areas for improvement as it relates to this Problem Area. One of the ways in which we address content targeting users based upon protected characteristics is through our hate speech policies in the Community Standards. As hate speech and attacks continue to evolve, detecting and enforcing against these attacks remains challenging. For example, threat actors continue to explore ways to circumvent detection and enforcement and new variations of slurs are continuously introduced. Additionally, as Meta does not collect data regarding certain protected characteristics to protect user privacy, there is no comprehensive data regarding differences in experiences on our technologies based upon a protected characteristic.

Our enforcement actions intended to reduce hate and attacks might have an impact on voice due to the less obvious boundaries for this Problem Area, when compared to others, such as Child Sexual Exploitation, Abuse, and Nudity. Additionally, there are limitations in automated detection, such as language specific slurs or Gen Z language. However, Meta is continuing to evolve its systems to detect and counter discrimination and hate on its platforms.

#### 6.2.2.8 Disinformation

Disinformation refers to promotion and distribution of false or misleading content spread with an intention to deceive or secure economic or political gain, including covert influence operations and coordinated inauthentic behaviour. Meta does not maintain a separate disinformation policy and disinformation risks are covered under aspects of both the Misinformation and Inauthentic Behaviour Problem Areas.

Disinformation is associated with the Deceptive and Misleading, Civic Discourse and Elections, Public Health, and Public Security Systemic Risk Areas in the DSA. This Problem Area relates to the risk of Facebook being used to promote and distribute false or misleading content that is spread with an intention to deceive or secure economic or political gain and which may cause public harm. This potentially includes the misuse of Facebook's systems to engage in covert information influence operations and coordinated inauthentic behaviour, such as manipulated media.

**What are we doing to try to prevent and mitigate these risks?**

**2024 Trends:** During the assessment, it was identified that the high number of elections in the EU, including the EU Parliamentary Elections, crises in adjacent regions, the assassination attempt on Slovakia's Prime Minister, and the increasing adoption of generative AI provides adversarial actors more intent and means to propagate disinformation, particularly related to ads. As a result, Meta has dedicated election teams, generative AI risk assessments, and classifier training initiatives to combat the expected increase in adversarial behaviour.

**Problem Area Mitigation Overview:** As detailed in Section 6.2.1, we have an extensive ecosystem of controls that work together to manage these Problem Area risks on Facebook. Specifically for Disinformation, while most of the manual and automated measures deployed to handle misinformation are also applicable to disinformation, there are certain additional measures focused on addressing the intentional sharing of misleading information; more information on our misinformation measures, including our partnerships with independent fact-checking organisations, can be found in Section 6.2.2.14. Disinformation is managed in three ways which include additional investigation by expert investigators to identify, remove and block users or groups engaging in coordinated inauthentic behaviour; automated systems trained to identify and prevent intentional and repeat offenders; and broader automated integrity systems like our fake account detection and removal systems. Our product design also includes interventions to reduce or add friction to actors' behaviour to propel disinformation. For example, targeted measures against state-controlled media, in-feed labelling, and interstitials for users are designed to reduce the distribution of disinformation. **Between July and December 2023, we attached fact-checking labels to over 68 million pieces of content viewed in the EU on Facebook and Instagram globally.** When a fact-checked label is attached to a post, the majority of our users do not click on the post to view it.[94] Meta also has a Disinformation Code Insights and Verification (CIV) programme that shares all endorsements that do not lead to immediate risks but are potentially misleading.

Additionally, we have processes and dedicated teams to track emerging trends and disinformation themes and plan for appropriate mitigation, such as during times of major elections or conflicts in adjacent regions. **To counter covert influence operations, we have built specialised global teams that have taken down over 200 adversarial networks since 2017** which we published in our Quarterly Threat Report.[95] Additionally, as it relates to emerging trends, Meta has also undertaken targeted measures to address the risk of disinformation propagating through the use of generative AI, including labelling of AI generated or edited content for user awareness, building a feature for people to disclose when they share AI-generated video or audio so we can add a label to it, building invisible markers to help detect deep fakes, requiring advertisers running ads on social issues, elections or politics to identify digitally altered and/or AI generated content, and collaborating with the industry on common standards and guidelines. **Between July and December 2023, we have removed 430,000 ads across the EU for failing to carry a disclaimer.**[96] Additionally, we are a member of the Partnership on AI and we recently signed on to the Tech Accord designed to combat the spread of deceptive AI content in the 2024 elections.[97]

Furthermore, advertisers who want to create or edit ads in the EU that reference political figures, political parties, elections in the EU or social issues within the EU are required to go through an authorisation process and have a "Paid for by" label. To help guard against foreign interference, advertisers (including political organisations and agencies) who want to run ads about social issues, elections or politics must have their ad pre-approved and run by a person who is authorised in the EU country that they are targeting. Meta is also directly engaging with the European Commission as a signatory of the European Union's Code of Practice on Disinformation. These reports provide further insight into our actions to fight disinformation.[98]

**Limitations:** Throughout our assessment, we identified areas for improvement as it relates to this Problem Area. Our fact-checking programme forms the basis to prevent the spread of large scale misinformation and disinformation and is capacity bound. The policy today is focused on fact-checking for widespread hoaxes, such as 'climate change is not real', and not for every piece of content, as that is not feasible. This may become increasingly challenging with the adoption of generative AI and propagation of AI-generated content on our platforms. We also want people to have the ability to share their thoughts freely, even if it is not the most accurate, as long as it is not harmful to anyone. However, implementing those boundaries to identify and moderate false content makes this area challenging. One of the ways we manage this is through behavioural enforcement, where we focus on the behaviour of the accounts rather than just the content itself. Meta is investing heavily to adapt our detection and enforcement processes and system to prepare ahead for these challenges.

[94] https://about.fb.com/news/2024/02/how-meta-is-preparing-for-the-eus-2024-parliament-elections/
[95] https://about.fb.com/news/2024/02/how-meta-is-preparing-for-the-eus-2024-parliament-elections/
[96] https://about.fb.com/news/2024/02/how-meta-is-preparing-for-the-eus-2024-parliament-elections/
[97] https://about.fb.com/news/2024/02/how-meta-is-preparing-for-the-eus-2024-parliament-elections/
[98] https://disinfocode.eu/reports-archive/?years=2024

#### 6.2.2.9 Fraud and Deception

At Meta, we remove content intended to defraud users or third parties and content that goes against our Fraud and Deception Community Standards. While we allow our users to raise awareness, educate, and condemn fraudulent and deceptive practices, we do not allow content that seeks to coordinate or promote these activities using our platform.

Fraud and Deception is associated with the Deceptive and Misleading, Fundamental Rights, Public Health, and Protection of Minors Systemic Risk Areas in the DSA. This Problem Area relates to the risk of Facebook being used to provide instructions on, engage in, promote, recruit, facilitate, or distribute fraudulent and deceptive ads that could potentially impact public health; public health related investment, financial, product, or inauthentic identity/fake engagement scams; stolen information, goods, and services related to public health; or deceive or misrepresent themselves to others for financial or personal benefit related to public health. Additionally, threat actors are using evolving technologies, such as generative AI, to come up with new ways to defraud and deceive people.

**What are we doing to try to prevent and mitigate these risks?**

**2024 Trends:** During the assessment, there were no new trends identified that could potentially change the inherent risk exposure associated with this Problem Area.

**Problem Area Mitigation Overview:** As detailed in Section 6.2.1, we have an extensive ecosystem of controls that work together to manage these Problem Area risks on Facebook. Specifically for Fraud and Deception, we use several machine learning models to identify users that are misrepresenting themselves or displaying common scam tactics. These models are constantly evaluating our entire user base and removing any accounts deemed to be violating as per our Community Standards. Over the last year, we have made significant investment in expanding the capability of these models. For example, we launched a new version of our Deceptive Identity and Scam Account Score models which significantly improved our detection capabilities and led to increased enforcement volumes.

To further supplement internal improvements, the Fraud and Deception Team engages with a number of external trusted third parties, including ScamHaters United, Zero Fox and Learning Lab Admins, to gather information about scam-related accounts and content. Meta actions any reported content that violates our Community Standards either through automation or manually. We use the insight from these cases to further iterate on our proactive detection and internal policies. We also intake other signals, such as URLs from financial institutions, to detect and address marketplace scams.

Additionally, we include a link on nearly every piece of content to report scam, fraud, false information, and other issues as part of our enhanced user initiated reporting experience, and **our teams work 24 hours a day, 7 days a week, to review content reported**. Furthermore, Meta has developed a library of tools and resources for improved online safety to help educate users, provide guidance about scams, and encourage our users to report content they believe violates our Community Standards and policies in our Facebook Help Centre, our Scam Safety Centre, and Anti-Scams Hub. We also deploy features to protect users such as product interventions with safety notices and prevention features like adding friction to the discovery of harmful content.

**Limitations:** Throughout our assessment, we identified areas for continued improvement as it relates to detection. Threat actors use evolving technologies, such as generative AI, to study how our detection and enforcement controls are designed and to evade them. There are also instances where ads may promote a service or product that may not exist in reality or do not match to their promises, such as hair growth shampoo, but are not necessarily illegal. Additionally, threat actors use signposting which leads users to harmful external sites. Our fraud and scam frictions are reliant upon effective external deterrents from law enforcement, which is addressed differently depending on the jurisdiction and makes effective management challenging. However, Meta is continuously working to improve and enhance our capabilities to implement further mitigations on Facebook, such as our policy harmonisation efforts and network disruption approach, where we take down each adversarial network of accounts and Pages as a whole, rather than removing them piecemeal.

#### 6.2.2.10 Hate Speech

At Meta, we define hate speech as a direct attack against people on the basis of protected characteristics such as race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease, as described in our Hate Speech Community Standards. We do not allow hate

speech on our platforms as this type of content can create an environment of intimidation and exclusion, and in some cases may promote offline violence.

Hate Speech is associated with the Civic Discourse and Elections and Gender-based Violence Systemic Risk Areas in the DSA. This Problem Area relates to the risk of Facebook being used by users to promote dehumanising speech or imagery against a user based on their protected characteristics; mock the concept, events, or victims of hate crimes, incite hatred towards a person and/or disparage a person based on their protected characteristics . This risk can be challenging to manage as the types of content and terms used for hate speech change frequently.

| What are we doing to try to prevent and mitigate these risks? |
| --- |
| **2024 Trends:** During the assessment, it was identified that the high number of elections in the EU, including the EU Parliamentary Elections could increase the inherent volume of activity as it relates to this Problem Area. As a result, Meta has put in place dedicated election teams to combat the likely increase in adversarial behaviour.<br><br>**Problem Area Mitigation Overview:** As detailed in Section 6.2.1, we have an extensive ecosystem of controls that work together to manage these Problem Area risks on Facebook. Specifically for Hate Speech,, we proactively detect and enforce against hate speech. In order to help manage the linguistic and cultural nuances of this Problem Area, Meta maintains and leverages a Market-specific Slurs List, consisting of inherently offensive words that are used as an insult for a protected characteristic in specific jurisdictions, which is used to identify slurs and surface them to our reviewers for review and labelling, where applicable. For images, we utilise classifiers and, in some cases, a central bank of content that has previously been flagged and enforced against, to try and proactively identify instances of these images being posted again. **In the first quarter of 2024, we actioned 7.4 million pieces of hate speech content on Facebook globally, with 94.70% being identified by us before users reported it.**[99]<br><br>We also have educational interstitials to deter users from posting violating hate speech content or from encountering potentially hateful content, Groups, and Pages by issuing warnings or, in more extreme cases, providing a link to relevant resources in our Safety Centre based on the terms used. We also empower our users with tools like blocking. When a user's content is reported and found policy-violating, the user is notified of the policy they have violated and may be given the option to edit their content for potential reinstatement. We also leverage a strike system where once the threshold for repeated policy violations is met, the user's account is removed. Furthermore, our Comment Warning mechanism is designed to make users aware of potentially offensive comments by displaying warnings on our platforms that remind users of Community Standards and inform them about enforcement actions, such as comment takedown or hide.<br><br>We also work with external stakeholders, such as governments, watchdog groups, and Trusted Partners, to help us identify instances of hate speech. For example, Meta is a signatory to the EU Code of Conduct on Countering Illegal Hate Speech, which involves us working closely with the European Commission and a network of civil society organisations located in different EU countries. In our Safety Centre, we have developed expert-backed safety resources and tools based on topics, such as Mental Health and Bullying and Harassment, and communities, such as women and LGBTQ+, to help our users as they face issues related to hate speech and other Problem Areas.<br><br>**Limitations:** Throughout our assessment, we identified areas for improvement as it relates to this Problem Area. As the types of content and terms used for hate speech are frequently changing, consistently detecting and enforcing against hate speech remains challenging. For example, threat actors continue to explore ways to circumvent detection and enforcement, such as implying instead of explicitly stating things, new trends can emerge as contentious depending on regional nuances. Additionally, often users can post content that is borderline hate speech, which makes over enforcement challenging to manage. Meta currently does not use actor and behaviour signals at scale, which makes it more difficult to handle recidivism. Currently, there are no policies targeting minors specifically in our Hate Speech Community Standards. However, Meta's Policy Team undertook a youth safety audit on all policies and identified the need to re-assess age-appropriateness of some policies and potentially restrict visibility of some types of content to minors. |

#### 6.2.2.11 Human Exploitation

At Meta, we do not allow content that facilitates or coordinates the exploitation of humans, including human trafficking and smuggling, as described in our Human Exploitation Community Standards. Meta defines

[99] https://transparency.meta.com/reports/community-standards-enforcement/hate-speech/facebook/

human trafficking as the business of depriving someone of liberty for profit and the United Nations defines human smuggling as the procurement or facilitation of illegal entry into a state across international borders.

Human Exploitation is associated with the Gender-based Violence, Protection of Minors, and Fundamental Rights Systemic Risk Areas in the DSA. This Problem Area relates to the risk of Facebook being used by threat actors to, ask for, or facilitate human smuggling; human trafficking; and/or facilitating content and activities that adversely impact the dignity and rights of users.

**What are we doing to try to prevent and mitigate these risks?**

**2024 Trends:** For this year's assessment, it was identified that global events, such as conflicts and the preparation for the Olympic Games in Paris, could increase the risk of human exploitation, including human smuggling. As a result, Meta increased its resource investment for events like the Olympics and has a crisis management tool to help decide on trends that should be deployed when conflicts arise.

**Problem Area Mitigation Overview:** As detailed in Section 6.2.1, we have an extensive ecosystem of controls that work together to manage these Problem Area risks on Facebook. Specifically for Human Exploitation, we are committed to raising human exploitation awareness globally, including by **making our Human Exploitation policies available in over 90 languages including the languages of the EU member states**. To help improve our Human Exploitation policies, we regularly collaborate with our trusted network of external stakeholders. Over the last year, this has resulted in adding more policy lines to our Human Exploitation Community Standards where we now differentiate between minor human trafficking and adult human trafficking. Additionally, improvements have been made to our human exploitation detection mechanisms by incorporating new trends and signals, updating our blocklist databases, ingesting signals from trusted partners, and routinely training classifiers to more effectively remove content that violates our Human Exploitation Community Standards. We leverage Media Match Service, Severity Framework, Integrity Brain, and High Risk Early Review Operations (HERO) to minimise human exploitation content on our platforms. We strive to include additional languages in our Search Interventions Programme in which we identify key words that may be associated with illicit activity and add friction to search results. In these interventions, we include links to resources for support. Additionally, we get feedback from our reviewers regularly to understand the trends they are seeing and update our policies accordingly to improve our detection capabilities. Furthermore, Meta maintains established local market teams to help identify dialects and trends related to human exploitation in order to improve our detection and enforcement capabilities.

**Meta has partnered with experts across academia, advocacy, victim services and support, and law enforcement to develop more than 50 tools and features to support the safety of its users and provide guidance to users.**[100] We encourage anyone who encounters content on Facebook that indicates someone is in immediate physical danger related to human exploitation to contact local law enforcement immediately and report this content to us. We provide links to local resources available in our Help Centre if anyone is a victim of human exploitation or would like resources to share with a potential victim. **We work with more than 400 safety organisations worldwide,**[101] and among them, we work closely with key anti-trafficking experts, including NCMEC, International Centre for Missing and Exploited Children (ICMEC), Polaris, Stop The Traffik, International Justice Mission, ECPAT International, and Tech Against Trafficking, where we work together with companies, non-profits, academics, and relevant stakeholders in a collaborative environment to support and accelerate the impact of technology solutions combating human exploitation especially human trafficking. This includes collaboration with organisations to provide ad credits and/or ad support to educate users on human trafficking, including across the EU (e.g., labour trafficking campaigns), to help prevent and bring awareness to human exploitation related risks.

**Limitations:** Throughout our assessment, we identified areas for continued improvement as it relates to detection. Human exploitation is a human-based problem and as humans move across borders, the trends we see related to human trafficking outside the EU can affect people within the EU. In addition, human exploitation is highly adversarial and it may not always be apparent online and can be difficult to detect without the proper context. For example, it may be challenging to determine whether someone is truly being exploited or if they offered their services as an adult without force, fraud or coercion. This Problem Area is also subject to adversarial spamming which requires resources to triage and may draw away necessary resources from managing policy violating events involving trafficking. However, we continue to make significant improvements to our detection mechanisms to incorporate new trends and signals and more effectively remove content that violates our Human Exploitation Community Standards. There were also challenges identified as it relates to consistency regarding human reviews and cross-platform enforcement, including device blocking, and regional and linguistic nuances as it relates to enforcement mitigation measures. As a result, Meta has invested heavily in strengthening its detection and enforcement capabilities to help manage these limitations accordingly.

[100] https://www.meta.com/help/policies/safety/tools-support-teens-parents/
[101] https://about.meta.com/actions/safety/audiences/women/#partners

#### 6.2.2.12 Inauthentic Behaviour

At Meta, we do not allow users to engage in or claim to engage in inauthentic behaviour, which we define as the use of Facebook services (accounts, Pages, Groups, or events) to mislead people or Facebook about the identity, purpose, or origin of the entity that they represent, the popularity of Facebook content or assets, the purpose of an audience or community, the source or origin of content, or to evade enforcement, as described in our Inauthentic Behaviour Community Standards. This includes the unauthentic use and exploitation of our services.

Inauthentic Behaviour is associated with the Deceptive and Misleading and Civic Discourse and Elections Systemic Risk Areas in the DSA. This Problem Area relates to the risk of Facebook being used to deceive or mislead users by misrepresenting themselves; organising coordinated attacks; and private users, organisations, and governments coordinating multiple assets and manipulating others. In some instances, threat actors attack new and immature features to target users with willfully deceptive content or misleading information. This also includes misuse of Facebook's systems, potentially including circumvention of Facebook's detection systems; hijacking and taking over accounts; imitating Facebook functionalities; and misusing Facebook reporting systems. In some instances, groups participating in covert influence operations create fictitious identities resembling media organisations and other credible sources to spread misleading and deceptive content.

**What are we doing to try to prevent and mitigate these risks?**

**2024 Trends:** During the assessment, it was identified that conflicts in adjacent regions and the high number of elections in the EU, including the EU Parliamentary Elections could increase inauthentic behaviour, such as targeting of political groups, increased methods to evade detection by coordinated groups (e.g. Doppelganger), and foreign influence operations. As a result, Meta has dedicated election teams, deployed targeted tactics to manage known groups, and has a crisis management tool to help decide on trends that should be deployed when conflicts arise.

**Problem Area Mitigation Overview:** As detailed in Section 6.2.1, we have an extensive ecosystem of controls that work together to manage these Problem Area risks on Facebook. Specifically for Inauthentic Behaviour, it is a complex problem area that is purely behaviour driven where we are focusing on the actor's behaviour itself instead of their actual content. When we investigate and remove threat actors on our platforms, we focus on behaviour, not just content, no matter who is behind them, what they post or whether they are foreign or domestic. To help manage this Problem Area, Meta leverages tools trained to detect and analyse behaviour related to coordinated threats. We actively maintain and track behaviour-based signals and indicators that we feed into our classifiers to detect coordinated activity which is then escalated to our investigation team for deep-dive analysis to identify gaps and support improvements. We also monitor efforts by networks to return to the platform that we previously removed. Some of these networks may attempt to create new off-platform entities, such as websites or social media accounts, as part of their recidivist activity. Using both automated and manual detection, our teams are engaged in daily efforts to find and block threat actors' attempts to acquire new accounts, run ads, and share links to their websites and redirect domains, before these are ever shared on our platforms. We strive to find and remove content related to inauthentic behaviour early, before threat actors are able to build audiences among our users.

To help manage this Problem Area, Meta has a dedicated threat disruption working group which consists of product, policy, and investigation team members that analyse novel cases and false negatives to identify inauthentic behaviour trends and provide feedback to product teams to help adapt and increase the maturity of our detection systems to enable us to find other threat actors engaged in similar violating behaviours. **Since 2017, globally, Meta has identified and removed more than 200 adversarial networks.**[102] In the lead up to the publication of Meta' latest Adversarial Threat Report (Q1 2024), we organised deep dive sessions with the relevant national authorities in France, Germany and Poland in May, as well as with the European Commission Team in charge of the Disinformation Code, in an effort to contribute to the security community's efforts to detect and counter malicious activity on the internet. However, while public discourse ahead of the EU Parliamentary Elections focused primarily on foreign threats like Doppelganger, we found that the majority of EU-focused inauthentic behaviour we disrupted was domestic in nature. This included both coordinated inauthentic behaviour in Croatia and more simple inauthentic clusters removed in France, Germany, Poland and Italy. These clusters and networks had small numbers of accounts, primarily targeted audiences in their own countries, were more focused on local elections rather than the EU Parliamentary Elections, and many were linked to

[102] https://transparency.meta.com/en-gb/metasecurity/threat-disruptions/

individuals associated with local campaigns or candidates. On the foreign threats side, the attempts we've seen so far (including Doppelganger and a handful of inauthentic behaviour clusters we took down) were primarily focused on undermining support for Ukraine among the EU member states, rather than directly targeting the EU Parliamentary Elections.[103]

Additionally, we provide users the ability to report the abuse of our reporting systems. However, to prevent misuse of this capability, we have built mitigations in 2023, including machine learning models, to cross validate user reports against other signals, email verification in contact forms and an Integrity Programme Reporting Centre that verifies a user's identity. Furthermore, Meta leverages feedback from government authorities, law enforcement, research organisations, security experts, civil society and other technology companies to identify and stop emerging threats, inform our protocols and policies, and build new detection and enforcement tactics. We also share information from our investigations, such as in our Quarterly Adversarial Threat Report, to provide a more comprehensive view into the risks we tackle and contribute to the security community's efforts to detect and counter malicious activity on the internet.

Risks related to this Problem Area have varying severity levels and require different types of response. For example, an adversarial network engaging in covert influence operations will be shutdown by disabling the entire network of accounts whereas a single user pushing a misleading message will receive a violation notification. At a user level, we provide explicit warnings to users through the platform user interface or emails, when they are demonstrating violating behaviour or engaging with potential accounts that are demonstrating inauthentic behaviour.

**Limitations:** Since this Problem Area is behaviour driven, it is a dynamic space where human behaviours and adversarial tactics are constantly evolving to find new ways of evading or circumventing enforcement actions. Meta's approach is to constantly learn from past events, industry wide research, and subject-matter experts to refine our enforcement and evolve our processes and systems to address emerging trends and new threats. It is an ongoing effort and we are committed to continually improving to stay ahead.

#### 6.2.2.13 Intellectual Property (IP) Infringement

Meta takes intellectual property rights seriously and believes they are important to promoting expression, creativity, and innovation in our community. Therefore, Meta requires its users to respect copyrights, trademarks, and other legal rights, as published in our Intellectual Property Community Standards.

IP Infringement is associated with the Illegal Content Systemic Risk Area in the DSA. This Problem Area relates to the risk of Facebook being used to adversely impact intellectual property rights, potentially including copyright and trademark infringement. This risk is challenging to manage as enforcement of IP rights is primarily undertaken at the discretion of rights holders, which may mean Meta relies on reporting from rights holders to identify potential cases of infringement. Additionally, threat actors spam user reporting systems with false reports and appeals and they are continuously evolving how they circumvent policies and detection.

**What are we doing to try to prevent and mitigate these risks?**

**2024 Trends:** During the assessment, there were no new trends identified that could potentially change the inherent risk exposure associated with this Problem Area.

**Problem Area Mitigation Overview:** As detailed in Section 6.2.1, we have an extensive ecosystem of controls that work together to manage these Problem Area risks on Facebook. Specifically for IP Infringement, although only rights holders know with complete certainty what content is or is not authorised, if we have a strong basis to believe that something may be infringing, we take action - from removing or blocking the content, to disabling the responsible account or removing it across all of our recommendation surfaces. In order to identify potential violations, we use various automated detection tools that take into consideration a range of different signals such as insights from machine learning models, the presence of certain keywords associated with piracy and counterfeit activity and prior IP violations from problematic accounts. To ensure quick and accurate handling of IP reports, we provide dedicated channels for rights holders to report content they believe infringes their rights, including our online reporting forms available in our Help Centre. We have custom forms dedicated to copyright, trademark and counterfeit issues, which ensure that we receive all the information we need to process an IP report. Rights holders can report different types of content they identify on our platforms, ranging from individual posts, photos, videos or advertisements to an entire profile, account, Page, Group or event. Each report submitted by a rights holder is processed by our IP operations Team, which is a global team of trained professionals who provide coverage in multiple languages. If the report is complete and valid, the team will promptly

[103] Meta Quarterly Adversarial Threat Report Q1 2024

remove the reported content and confirm that action with the rights holder or user that reported it. In December of 2023, globally, **we removed 83.19% of content reported for copyright, 81.95% of content reported for counterfeit, and 58.82% of content removed for trademark**.[104] In December 2023, globally, **we also removed 85.86% of violating content related to copyright and 97.99% of violating content related to counterfeit before it was reported by a rights holder.**[105]

While much of the violating content is proactively removed through Meta's automated systems, Meta strives to help businesses that use our platforms fight against brand impersonation, intellectual property infringement, and infringing content. **Meta has four tools to help rights holders protect their intellectual property at scale.**[106] Our updated Brand Rights Protection Manager platform makes it easier for brands to protect their intellectual property across all our platforms using cross-surface searching, which allows simultaneous searches across different platform areas, including ads, commerce, accounts, and posts and eliminates the need for repetitive search term entries, effectively optimising the process. We added new features to our Rights Manager to help brands manage and protect their copyrighted content at scale such as automatic blocking of matching images, image attribution to rights holders, and bulk actions which allow for enforcement against multiple image reference files at once. Our Intellectual Property Reporting API allows rights holders to automate and streamline the reporting of infringing content by filling out the same fields as Meta's IP reporting forms in a secure and trusted way. We have developed our new Intellectual Property Reporting Centre to improve the process for reporting intellectual property rights violations by allowing rights holders to save account information and reporting history to track and manage cases more efficiently.

All of our IP tools function in a unique way and Meta provides users with guidance to choose the tools that fit their needs in a section of its Business Help Centre dedicated to Intellectual Property Tools. We also have a section dedicated to Intellectual Property in our Help Centre where we provide guidance related to copyright and trademark. Additionally, when we take down content, we provide links to the Help Centre to help educate the user and prevent recidivism.

**Limitations:** Throughout our assessment, we identified an area for improvement as it relates to this Problem Area, which is that threat actors are constantly evolving how they circumvent detection and spamming user reporting systems with false reports and appeals which can lead to overenforcement of non-infringing content.

#### 6.2.2.14 Misinformation

Misinformation is different from other types of speech addressed in our Community Standards because there is no way to articulate a comprehensive list of what is prohibited. With graphic violence or hate speech, for instance, our policies specify the speech we prohibit, and even persons who disagree with those policies can follow them. With misinformation, however, we cannot provide such a line. The world is changing constantly, and what is true one minute may not be true the next minute. Additionally, people have different levels of information about the world around them, and may believe something is true when it is not.

Misinformation is associated with the Deceptive and Misleading, Civic Discourse and Elections, Public Health, and Public Security Systemic Risk Areas in the DSA. This Problem Area relates to the risk of Facebook being used to promote and distribute false or misleading content shared without harmful intent, whereas Disinformation considers harmful intent, such as through covert influence operations and coordinated inauthentic behaviour. See Section 6.2.2.8 for more information on Disinformation.

**What are we doing to try to prevent and mitigate these risks?**

**2024 Trends:** During the assessment, it was identified that the high number of elections in the EU, including the EU Parliamentary Elections, the use of generative AI, and crises in adjacent regions could increase the inherent volume of activity as it relates to this Problem Area. As a result, Meta has dedicated election teams, launched keyword detection to group content related to the EU elections in one place to make it easier for fact-checkers to find, and has a crisis management tool to help decide on trends that should be mitigated when conflicts arise.

**Problem Area Mitigation Overview:** As detailed in Section 6.2.1, we have an extensive ecosystem of controls that work together to manage these Problem Area risks on Facebook. Specifically for Misinformation, in order to manage this nuanced Problem Area, we apply our Remove, Reduce, Inform approach. We remove misinformation or unverifiable rumours that clearly violate our Community Standards. For example, this

[104] https://transparency.meta.com/reports/intellectual-property/notice-and-takedown/facebook/
[105] https://transparency.meta.com/reports/intellectual-property/proactive-enforcement/facebook/
[106] https://www.facebook.com/business/help/611786833293457

includes content that poses risk of imminent harm or violence to people; interferes with people's ability to participate in political processes, including ads that discourage people from voting in elections and information that calls into question the legitimacy of an upcoming or ongoing election or contains premature claims of election victory; and certain highly deceptive manipulated media.

Content that is not already subject to removal under Community Standards may be sent for independent fact-checker review. In many countries, our technology can detect posts that are likely to be misinformation based on various signals, including how people are responding and how fast the content is spreading. It also considers if users flag a piece of content as "false information" and comments on posts that express disbelief. Fact-checkers also identify content to review on their own. We label and reduce such misinformation that does not violate our Community Standards but is still determined false by our third-party fact-checking partners, which in Europe are certified by the EFCSN. Fact-checkers review a piece of content and rate its accuracy. This process occurs independently from Meta and may include, but is not limited to, calling sources, consulting public data, and authenticating images and videos. **Meta has built the largest fact-checking programme in the world, with 29 partners across the EU covering 23 languages and further adding 3 new partners in Bulgaria, France, and Slovakia in 2024.**[107] Further, we onboarded 16 signatories of the Code of Practice on Disinformation (CoP) to report misinformation content directly to Meta for the duration of the EU Parliamentary Elections as part of the Rapid Response System demanded by the code.[108] Additionally, Meta has a Misinformation Repeat Offender (MRO) Programme which limits the distribution of accounts that repeatedly share or publish content that is rated false or altered by fact-checkers for a period of 90 days or longer if the account continues to share misinformation.[109]

To better inform people, Meta also implements proactive mechanisms that connect users to reliable information from trusted experts with the goal of countering misinformation. This is done through centralised hubs like our Information Centre, Climate Science Information Centre or Voting Information Centre. Specifically for the EU Parliamentary Elections, Meta launched an in-app Voter Information Unit and provided Election Day Reminders directing people to local authoritative sources and reminding people to vote. Meta also attaches warning labels to content reviewed by fact-checkers and, if debunked, reduces its distribution in-feed so people are less likely to see it. Meta may also place election-related notifications in user's feeds on Facebook, such as voting day reminders. Between July and December 2023, for example**, over 68 million pieces of content viewed in the EU on Facebook and Instagram had fact-checking labels. When a fact-checked label is placed on a post, 95% of people do not click through to view it**.[110] Meta is also launching a media literacy initiative in the EU to educate people on how to better vet the information found online. Meta also operates an Education Hub that can be accessed at the discretion of users, that includes topics around media literacy and misinformation and provides resources to support teen's online experience.

**Limitations:** There is no society-wide consensus on what constitutes misinformation and how it should be addressed. User perspectives vary on what is false or misleading and similarly may differ as to whether enforcement is appropriate to safeguard information integrity versus the risk of limiting voice. Meta's approach to this is to rely on third party fact-checkers who independently review and rate content, focusing on viral, consequential, and provably false claims. Balancing voice and safety is also a key challenge while handling misinformation that may contribute directly to the risk of imminent physical harm to one or more individuals. However, Meta has processes and policies in place to address these challenges, including voice considerations, misinformation policies, and local expert advisors including civil rights and human rights groups. Another limitation is around fact-checking Stories, which due to their ephemeral nature, may expire before fact-checkers are able to review. However, our matching systems can detect already debunked content in Stories and that helps prevent any mass spread of misinformation.

#### 6.2.2.15 Privacy and Security

Privacy and the protection of personal information, in particular for minors, are fundamentally important values for Meta. At Meta, we do not allow people to post certain types of personal or confidential information and remove content that shares, offers, or solicits personally identifiable information or other private information that could lead to physical or financial harm, as described in our Privacy Violations Community Standards.

Privacy and Security is associated with the Fundamental Rights and Protection of Minors Systemic Risk Areas. This Problem Area relates to the risk of Facebook being used to adversely impact a minor's rights to protection of personal data and respect for private and family life as enshrined in the EU Charter. This can manifest on Facebook when the service is used by threat actors to search for and publish personal

[107] https://www.facebook.com/business/help/997484867366026?id=673052479947730
[108] https://disinfocode.eu/eu-elections-2024/
[109] https://socialmediaarchive.org/record/10/files/US2020_FB%26IG_Elections_External_Codebook.pdf
[110] https://about.fb.com/news/2024/02/how-meta-is-preparing-for-the-eus-2024-parliament-elections/

identifiable and private information without permission (doxxing). In certain cases, this can be challenging to manage because the only way for Meta to become aware is through user reporting. This Problem Area also includes the risk of Facebook potentially processing a minor's data without the proper consent and/or may target minors with ads using unapproved data.

**What are we doing to try to prevent and mitigate these risks?**

**2024 Trends:** During the assessment, there were no new trends identified that could potentially change the inherent risk exposure associated with this Problem Area.

**Problem Area Mitigation Overview:** As detailed in Section 6.2.1, we have an extensive ecosystem of controls that work together to manage these Problem Area risks on Facebook. Specifically for Privacy and Security, since 2019, **we have invested $5.5 billion in our Global Privacy Programme and have grown the teams focused on privacy to more than 3,000 people at the end of 2023.**[111] In the past year, we have updated elements of our Global Privacy Programme, like our Regulatory Readiness Process, strengthened our governance and compliance functions, and leveraged our core technology expertise to address privacy at scale. For example, we have increased the scope of **our Privacy Review Process, which now reviews an average of 1,200 products, features and data practices per month across the company** before they launch to assess and mitigate privacy risks. Additionally, we updated our "Why am I seeing this?" tool to help people understand why they are seeing the ads they do on their Facebook feeds. This provides more transparency about how user activity, both on and off our platforms, may inform the machine learning models we use to shape and deliver ads. We have also redesigned Ad Preferences to allow users to easily manage the ads they see.

Our External Data Misuse Team is dedicated to detecting, investigating and blocking patterns of behaviour associated with scraping. To combat data scraping on our platforms, we implement rate limits and data limits. Rate limits restrict the frequency of interactions with Meta's services in a given amount of time, while data limits control the volume of data accessible to ensure it aligns with normal usage needs. Additionally, Meta's Bug Bounty Programme engages external researchers to identify and report security vulnerabilities, enhancing the protection of user data and platform security. This proactive approach is complemented by ongoing reviews of any additional user profiles to maintain robust privacy and security standards. Additionally, our Privacy Risk Management Programme enables us to identify and assess privacy risks related to how we collect, use, share, and store user data. We leverage this process to identify risk themes, enhance our Privacy Programme, and prepare for future compliance initiatives. As part of our Global Privacy Programme, we have designed safeguards, including processes and technical controls, to address privacy risks and we conduct internal evaluations on both the design and effectiveness of the safeguards for mitigating privacy risk. To help us track and manage remediation of privacy issues, we have established a centralised Privacy Issue Management function that spans the privacy issue management lifecycle from intake and triage, remediation planning, and closure with evidence. We have also established a Privacy Red Team who proactively tests our processes and technology to identify potential privacy risks.

Part of ensuring that everyone understands their role in protecting privacy at Meta is driving continuous privacy learning and education. At Meta, we have developed our foundational required privacy training and also maintain a catalogue of all privacy training deployed across Meta based on topics relevant to people in specific roles. We also deliver ongoing privacy content through internal communication channels, updates from privacy leadership, internal Q&A sessions, and a dedicated Privacy Week. Furthermore, we have built tools to help users secure their information and make the right privacy choices, such as Privacy Checkup which is used by over 10 million users every month, Why Am I Seeing This Ad, and Two-Factor Authentication. Additionally, our Privacy Centre has been designed to help users learn more about our approach to privacy across our apps and technologies so they can make better informed decisions about their privacy.

**Limitations:** Throughout our assessment, we identified areas for improvement as it relates to this Problem Area. Managing privacy, integrity and voice at once is challenging but we continue to enhance our detection and enforcement capabilities to find the right balance. Also, as laws are constantly changing, Meta needs to consistently monitor for changing or new laws. While Meta is able to control what data is shared with third parties, it is difficult to control how data is used once shared. Threat actors may abuse Meta's platforms to obtain data through scraping and malicious links. However, Meta is continuously working to improve and enhance our capabilities to implement further mitigations on Facebook, such as our Network Disruption approach, where we take down each adversarial network of accounts and Pages as a whole, rather than removing them piecemeal.[112]

[111] https://about.fb.com/news/2024/01/investing-in-privacy/
[112] Security | Transparency Centre (meta.com)

#### 6.2.2.16 Restricted Goods and Services

At Meta, we do not allow individuals, manufacturers, and retailers to purchase, sell, raffle, gift, transfer or trade certain goods and services on our platform including firearms, firearm parts, ammunition, explosives, lethal enhancements, non-medical drugs, pharmaceutical drugs, marijuana, endangered species, live non-endangered species, human blood, alcohol/tobacco, weight loss products, historical artefacts, entheogens, or hazardous goods and materials, as described in our Restricted Goods and Services Community Standards and other applicable policies.

Restricted Goods and Services is associated with the Public Health, Public Security, and Protection of Minors Systemic Risk Areas. This Problem Area relates to the risk of Facebook being used to purchase, sell, raffle, gift, transfer, or trade goods and services that could impact public health, potentially including firearms; alcohol, tobacco, prescription products, drugs, and drug paraphernalia; sexual solicitation and prostitution; sexual enhancement products; hazardous goods and materials/explosives; stolen and false goods and services; gambling and games; medical and healthcare product; and documents, currency, and financial instruments. This can be challenging to manage as threat actors continuously use new ways to evade detection, potentially including the use of emojis to solicit restricted goods and services.

**What are we doing to try to prevent and mitigate these risks?**

**2024 Trends:** For this year's risk assessment, we singled out the discrete risk of "Live Non-Endangered Animals and Endangered Species" to account for recent instances of individuals and retailers promoting or coordinating the purchase, sale, raffle, gift, transfer or trade of live non-endangered animals and endangered species on Meta's platforms. Additionally, the following trends were identified: masking various scams as drug sales, rising gambling popularity online, increased risk of exposing minors to content related to alcohol, tobacco, and real money gambling, and threat actors targeting minors with weight loss products and cosmetic procedures. As a result, Meta has put in place several strategic initiatives over the last year to combat these trends.

**Problem Area Mitigation Overview:** As detailed in Section 6.2.1, we have an extensive ecosystem of controls that work together to manage these Problem Area risks on Facebook, as well as proactive detection and enforcement technology to help enforce our Restricted Goods and Services Community Standards. This includes our automated review mechanisms that analyse ads, sponsored Marketplace listings, and on sale posts, as well as other commerce listings, before they go live, and proactively block content that may be selling counterfeits using keyword detection and machine learning models. We further block hashtags where the hashtag name represents or is consistently used to share violating content. Views of violating content that contains restricted goods and services are very infrequent, and we remove much of this content before people see it**. In the first quarter of 2024, globally, 98.3% of violating content we actioned for Drugs and 98.4% for Firearms was detected and actioned on Facebook before users reported it.** Additionally, in the first quarter of 2024, the upper limit for violations of our Restricted Goods and Services policy for Facebook was 0.05%. This means that out of every 10,000 views of content on Facebook, we estimate no more than 5 of those views contained content in violation with our Restricted Goods and Services policy. [113] We also deploy empty search results and interstitials to users searching for restricted goods and services or high severity problems, such as high risk drugs, on the platform to warn users and share resources to learn more information.

For minors, Meta applies age gating restrictions to content related to diet products, cosmetic procedures, real money gambling, alcohol, and tobacco among others and leverages age enforcement infrastructure to make this type of content less visible to them. Meta also has a long-term age appropriate content strategy to reduce teens' exposure to harmful and age-inappropriate content, and has invested in classifiers and infrastructure to support this solution. Additionally, ads are not currently served to minors in the EU. We also provide blueprint training modules for advertisers on different topics, such as alcohol. Advertisers can also sign up for an online course that explains how ads work and the restrictions that apply.

Furthermore, we have teams dedicated to research in this space, including experts that continuously look into trends. With the current rise of discussions around high-risk drugs, we are increasing our human review capacity with reviewers that support machine learning training efforts and specialised workflows for drugs. We also work with external stakeholders to source information on coded keywords used in the drug space to inform our reviewer guidelines. Meta has also been exploring programmes with external stakeholders and other tech companies, such as a cross-industry collaboration on illicit drugs with Snap, including sharing adversarial behaviour signals using Media

[113] https://transparency.meta.com/reports/community-standards-enforcement/regulated-goods/facebook/

Match Service and Cross Problem Multimodal Matching.

**Limitations:** Throughout our assessment, we identified areas for improvement as it relates to this Problem Area. Threat actors consistently seek new ways to circumvent detection and enforcement methods by attempting to sell or solicit restricted goods and services in covert ways, such as using emojis, using slang in private Groups and Pages, signposting leading users to harmful external sites, and posting branded content as organic without paid/partnership labelling to circumvent our ad safeguards. Additionally, due to differing enforcement approaches across our platforms, cross-platform enforcement can be challenging as threat actors may attempt to sell restricted goods and services on one surface, which could result in removal of that user on that platform but not on our other platforms. However, Meta is continuously working to improve and enhance our capabilities to implement further mitigations on Facebook. Furthermore, gambling restrictions and regulatory requirements differ by gambling type in each jurisdiction, which makes it challenging to proactively detect and enforce against across all of our platforms. However, we have robust local law enforcement processes in place which help manage this accordingly.

#### 6.2.2.17 Suicide and Self-Injury

At Meta, we do not allow people to intentionally or unintentionally celebrate or promote suicide, self-injury or eating disorders as described in our Suicide, Self-Injury, and Eating Disorders Community Standards. However, we allow people to discuss these topics because we want Facebook to be a space where people can share their experiences, raise awareness about these issues, and seek support from one another.

Suicide and Self-Injury is associated with the Public Health and Protection of Minors Systemic Risk Areas in the DSA. This Problem Area relates to the risk of Facebook being used to promote, encourage, coordinate, or provide instructions for suicide, self-injury, or disordered eating, potentially including ads that could impact public health. In some instances, this can manifest on our services in the form of gamification which could result in viral trends that impact public health. This also potentially includes depictions of graphic self-injury, suicide attempts, or death by suicide; instructions for extreme weight loss or depictions of body parts with terms associated with disordered eating; or content mocking victims or survivors of suicide, self-injury or disordered eating. This can be difficult to manage as intent can be challenging to assess and we are unable to always intervene/execute interstitials due to our privacy protection safeguards.

**What are we doing to try to prevent and mitigate these risks?**

**2024 Trends:** During the assessment, there were no new trends identified that could potentially change the inherent risk exposure associated with this Problem Area.

**Problem Area Mitigation Overview:** As detailed in Section 6.2.1, we have an extensive ecosystem of controls that work together to manage these Problem Area risks on Facebook. Specifically for Suicide and Self-Injury, over the last year, we have made significant investment in managing suicide and self-injury on our platforms including updating our Suicide and Self-Injury Community Standards to Suicide, Self-Injury, and Eating Disorders Community Standards and included eating disorder as a form of self-injury, improved classifiers, launching personalised demotions of content, and recommendations filtering which has allowed Meta to improve detection and removal or downgrading of content that goes against our Community Standards and other policies. Additionally, Meta's list of keywords has been improved over the last year and 1000+ terms have been added to detect and enforce against potentially violating content on our platforms. Also, interventions have been launched to improve content moderation, which are supported by multiple feedback loops used to train and adapt our classifiers and human review processes. As it relates to minors, Meta has recently gone further than ensuring self-injury content is not recommended for teens, and has now began removing this content from teens' experiences on Instagram and Facebook, as well as other types of age-inappropriate content, even if it's shared by someone they follow.[114]

Our policy is flexible enough to handle both violating content and content that may be sensitive given the nature of this topic. In particular, we have a specific approach toward recovery content. When the content is not graphic or promotional but may still be upsetting, such as depicting healed cuts, we include an interstitial for sensitive content. We also have blocklists and banks of images that are regularly updated to help our detection and enforcement capabilities. We have an escalation pathway specific to this Problem Area called Credible Intent of Suicide (CIS) which sends resources to users who have posted content that is identified as being suicidal or self-harm related where allowed by local law. This helps quickly identify and provide support to users who are at risk for committing suicide or self-injury. As a result, in the

[114] Our Work To Fight Online Predators | Meta (fb.com)

first quarter of 2024, globally, we have seen an increase in actioned content for Suicide and Self Injury due to accuracy improvements in our proactive detection technology and suicide and self-injury content **actioned was 7.1 million, with 99.4% of this content being found and actioned by us before users reported it.**[115]

Furthermore, Meta maintains a repository of in-app mental health resources, including "The World Health Organisation (WHO) Digital Stress Management Guide", which provides easy-to-follow techniques designed to reduce stress and promote mental well-being. Our Emotional Health Hub offers an array of expert mental health tips and education related to suicide, anxiety, depression, and managing well-being. We also offer eating disorder resources in our Safety Centre. Meta's resources have been updated to include more targeted country specific resources, including hotlines for suicide and self-injury and eating disorders. In addition, Meta partners with the Crisis Text Line to support suicide and self-injury crises. We have developed a resource that can be accessed via the Safety Centre called #Chatsafe which helps young users communicate safely online about suicide and self-injury and encourages awareness and reflection on difficult topics. We also have Chatsafe for Educators to help educators better equip young people they have contact with to talk safely on social media about suicide.

Meta collaborates with suicide prevention experts, via its Global Suicide and Self-Injury Expert Advisory Group, to seek input on current research and best practices related to suicide and self-injury. We use these collaborations to inform our safety work as we develop new services and resources to support users who may be experiencing challenges relating to suicide and self-injury.

**Limitations:** Throughout our assessment, we identified areas for continued improvement as it relates to this Problem Area. Managing voice and safety is particularly challenging for this Problem Area as some users post about this type of content to share their experiences, raise awareness about these issues, and seek support from one another. Additionally, suicide and self-Injury content is nuanced and it may be challenging to understand single/individual pieces of content without the context and history at an account level, which may impact our ability to use automation at scale. Furthermore, due to legal restrictions, we cannot use our classifiers to proactively detect in Facebook groups and rely on users' reports.

#### 6.2.2.18 Violence and Incitement

At Meta, we do not allow language that incites or facilitates violence and credible threats to public or personal safety. This includes violent speech targeting a person or group of people on the basis of their protected characteristic(s) or immigration status. We provide further details regarding what content and activity is considered prohibited and the corresponding actions Meta may take against users in our Violence and Incitement Content Community Standards and Violent and Graphic Content Community Standards.

Violence and Incitement is associated with the Civic Discourse and Elections, Public Security and Gender-based Violence Systemic Risk Areas in the DSA. This Problem Area relates to the risk of Facebook being used to incite violence, including against marginalised groups, disseminate content depicting graphic imagery of injured people or animals, promote kidnapping and abduction, disseminate instructions on how to use or make weapons or explosives, and promote or solicit services for hire to kill. Threat actors often speak in coded or veiled languages when disseminating content that may pose a security threat.

**What are we doing to try to prevent and mitigate these risks?**

**2024 Trends:** During the assessment, it was identified that there are evolving trends that impact the risk of Violence Against Marginalised Communities, which include the high number of elections in the EU, such as the EU Parliamentary Elections, the increase of Anti-Semitism and Islamophobia sentiment in the EU, and the increase in anti-immigrant and anti-refugee sentiment. Meta has a number of mechanisms in place to manage increases in activities in the region and prioritisation mechanisms to manage this.

**Problem Area Mitigation Overview:** As detailed in Section 6.2.1, we have an extensive ecosystem of controls that work together to manage these Problem Area risks on Facebook. Specifically for Violence and Incitement, we have built proactive detection and enforcement technology to help enforce our policies, such as Athena, which is a detection tool that provides a view of risks across our services before they become a bigger problem. The tool helps us take a proactive approach by highlighting early warning signs, such as unusually high classifier scores, accounts with multiple recent strikes, or an uptick in violent content, which are used by our Integrity Teams to craft proactive and reactive mitigations in response to these signals. To help understand the unique circumstances that may make some regions more sensitive to this type of content, our Violence and Incitement Team maintains a Temporary High-Risk Location (THRL) list of geographic locations that

[115] https://transparency.meta.com/reports/community-standards-enforcement/suicide-and-self-injury/facebook/

may be designated as having greater risk due to their likelihood of being the target of violence to help prevent Meta's platforms being used to incite violence or exacerbate conflict. Furthermore, to help identify threats of violence, we use our Veiled Threats Assessment Framework to drive governance of veiled threats, which are coded statements where the method of violence or harm is not clearly articulated. These types of threats may be ambiguous to the average reader but clearer to individuals with relevant context. Also, we regularly update our Market-specific Implicit Threat Terms List used in our policies. Consequently, in the first quarter of 2024, globally, we have seen an increase in actioned content for the Violence and Incitement Problem Area due to updates to our proactive detection technology that **improved identification of hostile speech and actioned 8.7 million pieces of content related to violence and incitement globally, with 97.9% of this content actioned proactively before users reported.**[116]

When a user's content is reported as hostile or violent speech and found to be violating, the user is notified of which policy they violated and may be given the ability to edit the post for potential reinstatement. If Meta becomes aware of information giving rise to a suspicion that a threat to the life or safety of a person or persons exists, Meta promptly notifies the applicable authorities and provides relevant information in accordance with our Terms of Service, international standards, and applicable laws. Additionally, we leverage a strike system, where once the threshold for violating multiple policies is met, the user account will be removed. Other enforcement actions include reducing visibility, device blocking, and account restrictions.

To help support users who may be experiencing abuse and/or violence, Meta has developed safety tools, such as the anonymous Domestic Violence Helplines in our Safety Centre where trained experts are available to offer support and specific guidance to create a safety plan. In our Safety Centre under Crisis Support Resources, we provide a global directory of crisis support resources that was compiled in partnership with UN Women, the National Network to End Domestic Violence and the Global Network of Women's Shelters to provide more urgent and expert support.

**Limitations:** Throughout our assessment, we identified areas for continued improvement as it relates to detection. Meta's policies are defined at a global level, which may create challenges for our content moderation mechanisms to understand the language and regional nuances in hostile and violent behaviour. For example, there is a potential risk of over enforcing as it relates to interactions between opposing sports teams when they use certain words such as "fight the enemy team". Additionally, younger users, such as Gen Z, may use terms or phrases that are not yet able to be flagged by detection systems. Furthermore, threat actors circumvent detection and enforcement by using slang and emojis. However, we continue to improve our detection and enforcement capabilities to help manage these limitations accordingly.

#### 6.2.2.19 Voice and Free Expression

At Meta, we are committed to respecting our users' voices and helping them connect and share safely. Our Facebook Community Standards aim to create a place for expression and give people a voice.

Voice and Free Expression is associated with the Civic Discourse and Elections and Fundamental Rights Systemic Risk Areas in the DSA. This Problem Area relates to the risk of Facebook adversely impacting users' fundamental rights, specifically the right to freedom of expression and information as enshrined in the EU Charter. This can manifest on Facebook through overenforcement of non-policy violating content, disproportionate enforcement of policy violating content, language/dialect limitations of human reviewers or classifiers, failure to take down policy violating content and activity that limits or discourages a users' freedom of expression, or because our policy lines err on the side of safety rather than freedom of expression. Additionally, Meta has to evaluate government takedown requests for consistency with our policies and works to prevent government surveillance on our services whilst still alerting the appropriate authorities in legally required situations.

**What are we doing to try to prevent and mitigate these risks?**

**2024 Trends:** During the assessment, it was identified that the high number of elections in the EU, including the EU Parliamentary Elections and crises in adjacent regions could impact this Problem Area but it was not determined to impact inherent risk. As a result, Meta has put in place dedicated election teams to combat the likely increase in adversarial behaviour.

**Problem Area Mitigation Overview:** As detailed in Section 6.2.1, we have an extensive ecosystem of controls that work together to manage

[116] https://transparency.meta.com/reports/community-standards-enforcement/violence-incitement/facebook/

these Problem Area risks on Facebook. Specifically for Voice and Free Expression, the Right to Freedom of Opinion and Expression is central to what we at Meta believe and work to protect. Meta periodically reviews and obtains guidance around protecting voice and freedom of expression from international human rights standards like Article 19 of the International Covenant on Civil and Political Rights (ICCPR), which define when it is appropriate to place restrictions on freedom of expression; we also conduct civil rights analysis on policy developments connected to principles of free speech. Meta maintains processes and systems to understand users' feedback on Meta's approaches for protecting users. These processes and systems include channels for users to provide feedback. We made progress toward our goal to bring the voices of marginalised communities into content policy development. We developed an Inclusivity Framework to ensure our diverse stakeholders are considered in the development of policy as well as to inform Community Standards.

Several of Meta's policies include a freedom of expression element that is taken into consideration by our detection and enforcement mechanisms. Meta allows recovery content such as healed wounds related to suicide and self-injury. As it relates to minors' voices, we have worked with global data protection regulators and organisations like the UN, the Organisation for Economic Co-operation and Development (OECD) and minors' rights groups to create Meta's Best Interests of the Child Framework, which distils the "best interests of the child" standards into six key considerations that product teams consult throughout our product development process, such as create safe, age-appropriate environments for youth and prioritise youth well-being and safety over business goals and interests. Additionally, we have developed extensive operational controls to assess the validity of government takedown requests according to GNI Principles. We do not want our content moderation enforcement to unduly limit freedom of expression. As such, we have developed and continue to improve our AI models to predict whether a piece of content is hate speech or violent and graphic content and our enforcement technologies to determine whether to take an action, such as deleting, demoting, or sending the content to a human review team for further review. By enforcing our policies, we seek to mitigate risks while upholding freedom of expression. When taking action against a user's account and/or content, we provide statements of reason and explanations for enforcement actions which may include the ability to appeal. We also have various appeals processes and frameworks to allow users to challenge us if they disagree with our enforcement decision except for violations with extreme safety concerns, such as child exploitation imagery. When a user appeals a decision taken against their content, we review the content again, using a combination of human review and technology, to determine whether or not it follows our Community Standards and will reinstate or take no action depending on the decision made. After that, if our original decision is not overturned or reversed, there may still be an opportunity for the user to appeal to the Oversight Board, who help us balance free speech and enforce against policy violating content.

**Limitations:** Throughout our assessment, we identified areas for continued improvement as it relates to this Problem Area. Cultural norms, behaviours, and politics are challenging to navigate across regions as there may be different viewpoints and expectations to balance. For example, the European Human Rights Standards can, at times, be stricter than Meta's policies which are globally driven. Another challenge in balancing safety and voice is how threat actors leverage product functionality to infringe on other user's voices, such as adversarial spamming in comments, reporting, harassment, and intimidation. However, we have robust processes in place which help manage these limitations accordingly.

# 7. Risk Mitigation Enhancements

As described in this Report and in line with Article 35 of the DSA, Meta puts in place reasonable, proportionate, and effective mitigation measures to address systemic risks, which includes identifying enhancements to these measures. As part of our journey of continuous improvement, we routinely evaluate our Integrity Ecosystem through several different methods to identify enhancement opportunities, including through our DSA Systemic Risk Assessment Process, monitoring of our extensive ecosystem of controls, user feedback on our enforcement activities, and close collaboration with global experts and industry partners.

In addition to the controls described in this Report, Meta has put in place additional enhancements to its control environment since this year's assessment was concluded. This section provides details on these enhancements.

| Enhancement Name | Enhancement Description |
|---|---|
| **User Reports** | Meta launched a new reporting experience for organic content on Facebook and Instagram and for profiles on Facebook-only to better enable users to indicate high-risk issues. |
| **Recidivism and Cross-Platform Enforcement** | Meta defined policy and implemented cross-platform propagation between several Family of Apps products to more consistently remove threat actors across our different platforms. We also established a measurement for under-enforcement on our platforms. |
| **Recommendation Surfaces** | Meta focused on reducing the prevalence of violating content on search and recommendation surfaces. |
| | Meta harmonised policies for non-recommended content across Facebook and Instagram to improve enforcement consistency. |
| | Meta implemented a comprehensive measurement of risky connections across its surfaces, focusing on mature measurements of connections between Child Safety Actors and their targets and SCAMS and Child Safety disables to enhance surface security. |
| **Prevalence** | Meta launched a system to review ads that are predicted to potentially violate the Ads Standards before the ad is able to go live. We also launched a prioritised review pipeline for teen-relevant content globally on Facebook and Instagram in May 2024. |
| **Payments and Revenue** | Between December 2023 and June 2024, Meta established systematic signal-sharing between Integrity and Financial Compliance teams to ensure user disablement action taken globally by Integrity on user accounts for Human Exploitation, Child Safety, and Terrorism violations is shared with Financial Compliance for financial investigation and action. |
| **Policy Harmonisation** | We are increasing parity between policies and enforcement on organic and ads across Facebook and Instagram globally. |

# 8. Conclusion

Facebook's mission is to give people the power to build community and bring the world closer together. We build technology that helps people connect, find communities, and grow businesses. Facebook also helps people discover and learn about what is going on in the world around them, enable people to share their experiences, ideas, photos and videos, and other activities with audiences ranging from their closest family members and friends to the public at large, and stay connected everywhere by accessing our services.[117] Whilst acknowledging that policy violating content and behaviour risks can occur on Facebook, which may also have wider impacts, we remain committed to identifying, assessing, monitoring and addressing those risks and to providing a trusted and safe environment for our users while respecting their fundamental rights, including freedom of expression.

The purpose of this Report has been to document and share the findings of our second annual Systemic Risk Assessment as required under Articles 34, 35, and 42 of the DSA. Throughout the course of this risk assessment, we defined and documented Facebook's Systemic Risk Landscape depicting 19 Problem Areas either mentioned in the DSA or understood by Meta. Using our deep domain knowledge and robust risk assessment process, we assessed our environment to evaluate 122 risks and corresponding controls. The robust measures we have implemented to identify, manage, and mitigate risk are rigorously tested and highly effective, but we are continuously working to improve and enhance our capabilities. Looking ahead, we will continue to engage with external specialists, experts, and thought leaders to better understand the risks associated with Facebook as the world continues to evolve.

We will continue to strive to be a leader in online trust and safety and contribute to the development of guidance and standards for the industry. We look forward to obtaining feedback and guidance on our Report results and risk assessment approach, and engaging with the European Commission as we continue to improve.

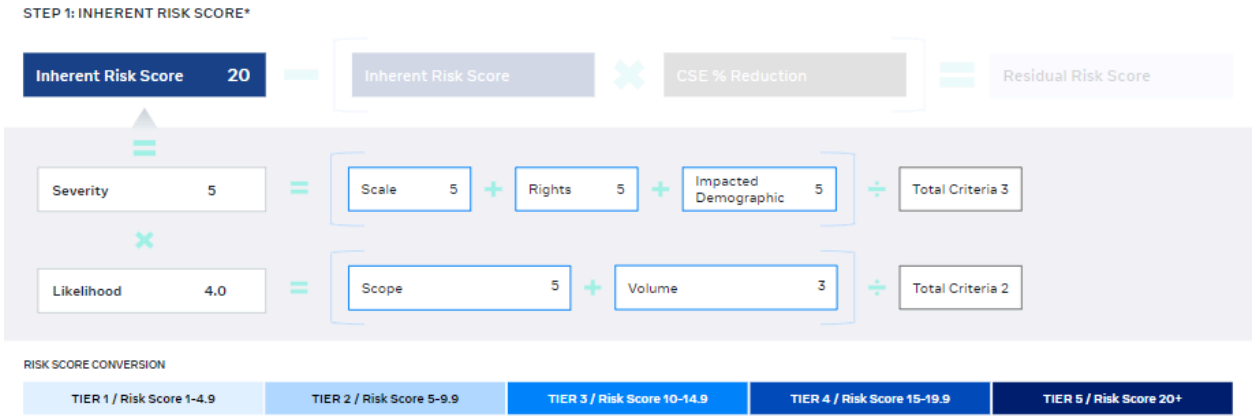
[117] https://d18rn0p25nwr6d.cloudfront.net/CIK-0001326801/c7318154-f6ae-4866-89fa-f0c589f2ee3d.pdf

# 9. Appendix

## 9.1 Meta's Integrity Risk Assessment Methodology: Rubrics

### 9.1.1 Inherent Risk Rubrics

The following measurement approach is used to determine inherent risk.




#### 9.1.1.1 Severity Rubrics

The Severity Rubric is used to measure the level of impact the risk has on users and society.

**SEVERITY RUBRIC**

| | TIER 1 | TIER 2 | TIER 3 | TIER 4 | TIER 5 |
|---|---|---|---|---|---|
| Scale | Isolated to Meta systems or processes | Isolated to individuals | Ramifications to segments or across industries | Broad ramifications to democratic processes/systems, way of life, societal norms, public safety | Severe ramifications to healthcare or finance sector; Country level critical infrastructure impact; Broader society stability; Protection of individuals, including targeted or marginalised groups |
| Nature of Impact | General Meta user experience | Economic impact (e.g., individuals ability to to earn an honest living and/or pursue legitimate business interests) | Impact to democratic systems | Physical and psychological impact | Impact to fundamental rights, such as Human Dignity, Freedom of Expression, Non-Discrimination, Rights of the Child. |
| Impacted Demographic | Repeat offenders, watch listed individuals, and other similar individuals | General users and developers | Elements of society | Society at large | Minors/Children; and/or Marginalised groups (e.g., Women, People with disabilities, elderly) |

#### 9.1.1.2 Likelihood Rubrics

The Likelihood Rubric is used to measure the possibility that a given risk will occur in a specified timeframe.

| LIKELIHOOD RUBRIC | | | | | |
|---|---|---|---|---|---|
| | TIER 1 | TIER 2 | TIER 3 | TIER 4 | TIER 5 |
| Scope | 250K users | <1M users | < 100M users | < 200M users | >200M users |
| Volume | Risk is rarely seen by users on the platform and must be deliberately searched for | Risk is infrequently seen by users on the platform | Risk is sometimes seen by users on the platform | Risk is often seen by users on the platform | Risk is persistently seen by users on the platform and able to be found inadvertently |

The following limitations should be considered when evaluating the likelihood of a risk arising:

- Likelihood is a subjective, qualitative measure and does not guarantee a risk will occur;
- All users are not equally likely to be impacted by the same Problem Area; and
- Volume is assessed from a qualitative perspective, and where available, validated data is used to help provide insight into the relative differences in volume at the risk or Problem Area level. However, it should be noted that there are a number of limitations with this data, including that this is a global data set, the data is at a Problem Area level not a risk level, and there is not data for all Problem Areas.

### 9.1.2 Control Effectiveness Rubrics

The following measurement approach is used to evaluate the effectiveness of controls in place to manage a risk. These questions aim to prompt the evaluator to enable a qualitative assessment.

| | |
|---|---|
| **Purpose** | Is the control detective or preventive, and is it automated or manual? |
| **Appropriateness** | Is the control built in a robust manner to mitigate the risk? |
| **Frequency & Consistency** | Does the control occur with appropriate frequency and is applied in a consistent manner? |
| **Aggregation** | Is the control designed at the correct level of the organisation to mitigate the risk? |
| **Investigation** | Can the control be evaluated and investigated in the future? Is there appropriate documentation? |
| **Dependency** | Does the control have a reliant design and built in a sustainable manner? |
| **Competency** | Does the operator of the control have the right skill level to operate the control? |

#### 9.1.2.1 Design Effectiveness

| DESIGN EFFECTIVENESS | | | | |
|---|---|---|---|---|
| | **TIER 3** | **TIER 2** | **TIER 1** | **TIER 0** |
| **Design Effectiveness** | The control is effectively designed for the control type. The control is operated on a regimented frequency by a responsible party with the optimal skillset or mechanism. Effective controls are regularly tested and reviewed. | The control is designed well for the control type but some improvements are needed. The frequency and occurrence of the control are adequate and reviewed on an ad-hoc basis. | The control design needs significant improvement. The frequency and occurrence of the control are sporadic and not appropriate to mitigate the risk. | The control design is not effective to mitigate the risk and is not reasonable or proportionate. |

#### 9.1.2.2 Operational Effectiveness

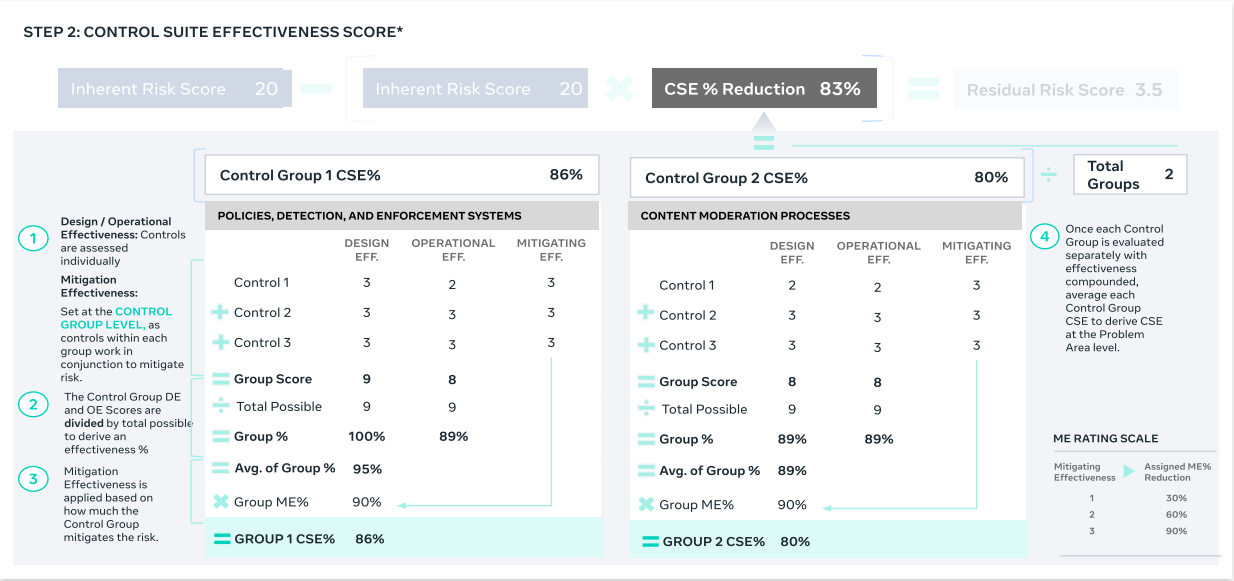
| OPERATIONAL EFFECTIVENESS | | | | |
|---|---|---|---|---|
| | TIER 3 | TIER 2 | TIER 1 | TIER 0 |
| Operational Effectiveness | The implemented control is consistently operating at an appropriate frequency in accordance with how it is designed to mitigate the risk. | The implemented control is mostly operating in accordance with how it is designed to mitigate the risk, but some improvements are needed | The implemented control is rarely operated in accordance with how it is designed to mitigate the risk and significant improvements are needed | The control is not in operation |

#### 9.1.2.3 Mitigation Effectiveness

| MITIGATION EFFECTIVENESS | | | | |
|---|---|---|---|---|
| | TIER 3 | TIER 2 | TIER 1 | TIER 0 |
| Mitigation Effectiveness | The control mitigates the risk effectively and is reasonable, and proportionate. | The control partially mitigates the risk effectively and is reasonable and proportionate. | The control has a limited effect on reducing risk exposure and may be reasonable or proportionate. | The control is not effective at mitigating the risk and is not reasonable or proportionate. |

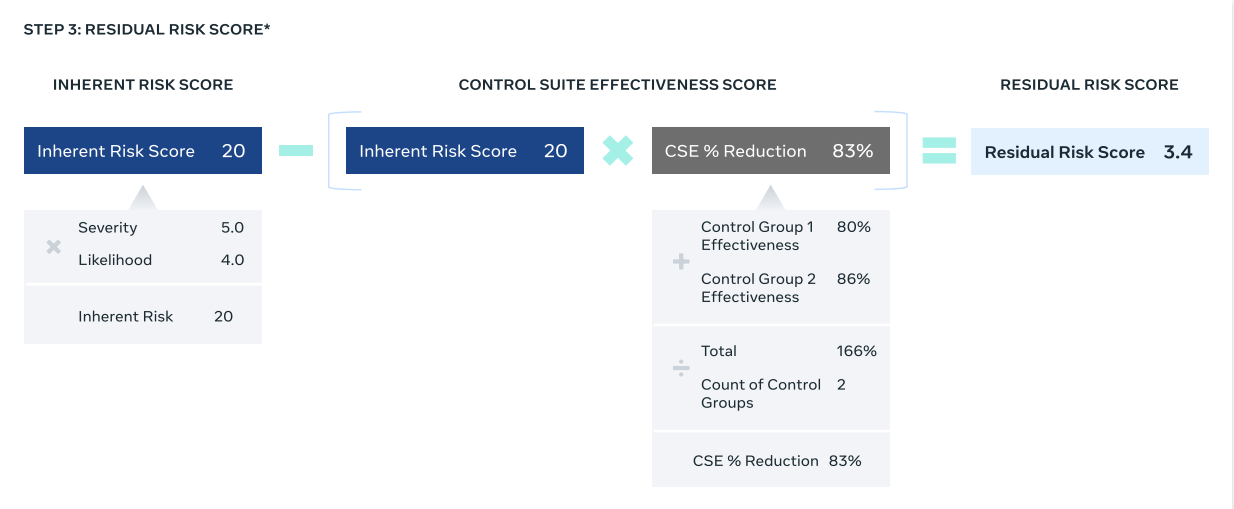
#### 9.1.2.4 Control Suite Effectiveness Calculation

Once the design, operational, and mitigation effectiveness for each control is determined, the following measurement approach is used to calculate the effectiveness of the control suite.




### 9.1.3 Residual Risk Calculation

The following measurement approach is used to determine residual risk.




## 9.2 Principles for ensuring Reasonable, Proportionate, and Effective Mitigation Measures

One way to approach making informed decisions to determine whether to invest in deploying a mitigating measure or enhancement is by considering the principles below. When making such a determination, we consider the impacts on fundamental rights.

| Criteria | Mitigation Measure | Further Details |
|---|---|---|
| **Reasonable** | - Within Meta's control to deploy with limited dependencies on external parties or non-Meta entities<br>- Appropriate, fair, and designed to address integrity risks or issues | Due to the residual risk exposure and/or the extreme criticality of a control in managing a systemic risk, it is appropriate to make investments to adapt, test, reinforce, initiate, adjust, and/or make changes to our systems, processes, and/or activities. |
| **Proportionate** | - Adequate, relevant, suitable and necessary to address specified systemic risks<br>- Not excessive in relation to a declared and specified purpose and residual risk exposure | The investment needed from a financial, technical, and operational perspective is commensurate with the current risk exposure or the risk exposure that will be created if the investment is not made. Additionally, in instances where rights, including fundamental rights, are in tension with a potential mitigation measure, a decision about a mitigation measure is based on the correlative impact a risk could have on users within the EU and society. |
| **Effective** | - Able to prevent, mitigate, or control the residual risk exposure as designed and intended<br>- Able to be monitored in order to measure its effectiveness | The investment needed to adapt, test, reinforce, initiate, adjust, and/or make changes to our systems, processes, and/or activities will effectively reduce the residual risk exposure of a systemic risk. |