Report of Systemic Risk Assessments

2024

Google

Published January 2025 - Google Ireland Limited is voluntarily publishing this Report of Systemic Risk Assessments earlier than required by Article 42(4) of the Digital Services Act.

Table of Contents

Table of Contents 1
1. Introduction 7
Foreword 7
About this Report 8
Scope and Purpose 8
Structure of the Report 8
Our Risk Assessment Methodology 10
Assessing Risk for each VLOSE and VLOP 10
Findings of our 2024 Assessments 12
2. Background 15
Maintaining User Trust and Safety 15
Our Commitments 15
Investing in Systemic Risk Prevention 16
Promoting Trustworthy Content and User Safety 19
One: Protecting Users from Harm 20
Preventing Harm with Safety by Design 20
Building Products that are Private by Design for Everyone 21
Preparing for the Unexpected 21
Designing Appropriate Content Policies 22
Reviewing Content Policies and Practices 23
Balancing Risks and Rights 23
Detecting and Responding to Harmful Content at Scale 24
Handling Government Removal Requests 25
Types of Enforcement Actions 25
Protecting Users with Applied AI 26
Offering Appeals 26
Reporting 27
Child Safety and Generative AI 28
Evaluating Content Across Languages 29
Two: Delivering Reliable Information 31
Surfacing High-Quality Information 31
Using Recommender Systems 31
Fighting Misinformation and Disinformation 32
Supporting Election Integrity 33
Surfacing high-quality information 33
Safeguarding our platforms and disrupting the spread of mis- and disinformation 34
Equipping campaigns and candidates with best-in-class security features and training 35
Helping people navigate AI-generated content 36

Addressing the Risks and Opportunities of Artificial Intelligence 37
Equipping Users 38
Our Approach to the Disclosure of Synthetic Content 40
Three: Partnering to Create a Safer Internet 42
Partnering for Information Quality 42
Consulting with Experts 43
Sharing Tools and Technology 44
Collaborating with Companies and Stakeholders 44
Developing Best Practices 46
Setting High Standards for Advertising 47
3. Methodology 50
Introduction 50
Step One: Classifying Risk 51
Step Two: Identifying Risk 52
Engaging Stakeholders 53
Step Three: Assessing Inherent Risks 55
Step Four: Assessing Preparedness 56
Taking a Human Rights-Based Approach 57
Step Five: Identifying Improvements to Mitigations 58
Step Six: Reporting the Results 58
4. Results of the Assessments 59
Search 60
Description of Service and Associated Risk Profile 60
Systemic Risk Assessment Results and Associated Observations 63
2024 Highlights 63
Removing Content 64
Removing Illegal Content 64
Addressing Violations of Intellectual Property Rights 64
Detecting, Removing, and Reporting CSAM 65
Removing NCEI and ISPI 66
Volume of Content Removed 67
Investing in Search Information Quality 67
Addressing Sensitive, Harmful, and Policy Violative Content 68
Informing Users 69
Providing SafeSearch 69
Tailoring our Content Policies 70
Addressing Civics Mis- and Disinformation 71
Respecting Freedom of Opinion, Expression, Media Pluralism, and Civic Discourse 71
Addressing Mis- and Disinformation 72
Service Design 73
Addressing Unfair Commercial Practices and Fraudulent Content about a Business 73
Respecting Privacy 74

Protecting Children's Rights 75
Obtaining Age Assurance 75
Enabling Parental Control 76
Providing Ads Protections 76
Enabling SafeSearch by Default 77
Built-in safeguards 77
Maps 78
Description of Service and Associated Risk Profile 78
Systemic Risk Assessment Results and Associated Observations 80
2024 Highlights 81
Content Moderation 81
Removing Illegal Content 81
Addressing Content that Violates our Policies 82
Developing Content Policy 82
Enforcing Content Policy 82
Undertaking Automated Detection and Removal 83
Undertaking Human Review 83
Undertaking Enforcement Proactively 84
Posting Restrictions 84
Posting Restrictions for Repeat Violators 84
Assessment Results for Specific Content Risks 85
Protecting Civic Discourse 85
Protecting Consumers and the Freedom to Conduct a Business 85
Respecting Freedom of Opinion, Expression, and Media Pluralism 86
Service Design 87
Respecting Privacy 87
Protecting Users of Maps 87
Protecting Contributors to Maps 87
Addressing Risks Relating to Images and Personal Information on Maps 87
Protecting Children's Rights 88
Enhancing Accessibility 88
Play 90
Description of Service and Associated Risk Profile 90
Systemic Risk Assessment Results and Associated Observations 92
2024 Highlights 92
Content Moderation 93
Removing Illegal Content 93
Addressing Content that Violates our Policies 93
Maintaining Developer Policies 94
Developing Policy 95
Enforcing Policy 96
Addressing Specific Content-Related Risks 97

Preventing Review Bombing and Ensuring Rating and Review Integrity 97
Protecting Civic Discourse 98
Platform Design 99
Knowing Developers and Protecting Users 99
Protecting Privacy 100
Protecting Children's Rights 101
Maintaining Additional Policies for Minors 101
Providing a Teacher-Approved Program 102
Obtaining Age Assurance 102
Enforcing Content Ratings and Content Restrictions 102
Shopping 104
Description of Service and Associated Risk Profile 104
Systemic Risk Assessment Results and Associated Observations 105
2024 Highlights 106
Content Moderation 106
Removing Illegal Content 106
Identifying and Blocking Illegal Products and Services 106
Prohibiting and Detecting Violations of Intellectual Property Rights 106
Addressing Content that Violates our Policies 107
Maintaining Google Shopping Policies 107
Maintaining Guardrails for User-Contributed Content 108
Preventing Unfair Commercial Practices 108
Preventing Fraudulent Business Information 109
Service Design 109
Respecting Privacy 109
Vetting Merchants 110
Monitoring Merchants and Listings 110
Protecting Children's Rights 111
YouTube 112
Description of Service and Associated Risk Profile 112
Systemic Risk Assessment Results and Associated Observations 114
2024 Highlights 115
Content Moderation 115
Removing Illegal Content 115
Two Examples: Terrorist or Violent Extremist Content and Child Sexual Abuse Material (CSAM) 115
Identifying and Removing Violent Extremist Content 116
Detecting, Removing, and Reporting CSAM 116
Prohibiting and Detecting Infringement of Intellectual Property Rights 117
Addressing Content that Violates our Policies 118
Developing Policy 118
Providing EDSA Exceptions 119
Enforcing Policy 119

Undertaking Automated Detection and Removal 119
Maintaining a Priority Flagger Program 120
Enforcing a Three-Strike System for Repeat Violators 120
Dealing with Generative AI content 121
Measuring Success: Violative View Rate 122
Elevating Authoritative Sources 123
Providing Information Panels with Topical Context 124
Addressing Specific Content Risks 124
Addressing Misinformation and Disinformation 124
Addressing Public Health Related Violative Content 125
Addressing Civic-Discourse-Related Violative Content 125
Detecting and Removing Harassment and Bullying in YouTube Comments 127
Prohibiting and Removing Hate Speech 128
Service Design 129
Respecting Privacy 129
Protecting Children's Rights 130
Maintaining Guardrails for Children's Access to Content 130
Addressing Potentially Addictive Behaviour in Children 131
Protecting Children's Data 132
Protecting Children's Safety in YouTube Comments 132
Promoting Equity 133
5. Conclusions 134
Annex A: Full List of Risk Statements 135
Illegal Content, Behaviour, and Products and Services 135
Freedom of Expression and Media Pluralism 135
Privacy and Data Protection 136
Human Dignity 136
Consumer and Business 136
Child Rights 137
Equality and Non-Discrimination 137
Civic Discourse 137
Public Health 138
Annex B: List of Mitigations 139
Background 139
Article 35 Mitigation Types 139
Mitigations Applicable to Multiple Services 141
Google Maps 141
Google Play 142
Google Search 143
Google Shopping 143
YouTube 144
Annex C: List of Consultations 145

Safety expert engagement during 2024 assessment period 145
Selected engagements in support of systemic risk assessments 145

1. Introduction

Foreword

In their first letter to shareholders, our founders described Google's goal as developing services that significantly improve the lives of as many people as possible. Since then, we have sought to expand opportunity by providing people with the information, tools, and services necessary to help them build knowledge, fuel curiosity, and unlock opportunity. This is all in service of our mission to organise the world's information and make it universally accessible and useful.

We have long designed services and policies, built teams, and developed technologies with the wellbeing of users in mind. Billions of people use our services every day, so it is essential that we protect our users with industry-leading security, responsible data practices, and easy-to-use privacy controls. The ability to access information on our services as well as the quality and safety of our services are directly linked to our ability to attract users, which in turn is critical to our continued success as a business.

It is in this spirit that we strive to make our products safe, transparent, and accountable, while ensuring that everyone around the world and in the European Union (EU) continues to benefit from the open web. This is a shared goal requiring a whole-of-society effort across the entire digital ecosystem, including companies, governments, civil society organisations, experts, and those using our services. With billions of users globally and hundreds of millions in the EU, Google has an especially important role to play.

This report sets out the results of the second annual systemic risk assessments that we have undertaken for our designated very large online search engine (VLOSE) and very large online platforms (VLOPs) to meet the requirements of Articles 34 and 35 of the Digital Services Act (DSA). These assessments represent continuous improvement in our approach to systemic risk and benefit from nearly two years of the DSA's requirements being in effect.

The results of these assessments illustrate that our mitigations are well-matched to the evolving systemic risk landscape, such as the expanding use of digital services during elections, the growing public availability of generative artificial intelligence (AI) tools, and shifting geopolitical conflicts. They help confirm that many longstanding Google commitments—such as safety by design, enforcing content policies that protect users from harm and respect freedom of expression, and providing user reporting channels—address risks to people and society in the EU.

We believe that the notion of "systemic" risk rightly conveys the need for collaborative approaches that reach across sectors, platforms, and stakeholders. In this context, these 2024 systemic risk assessments are noteworthy for their emphasis on new and growing collaborative efforts, such as our role establishing the Frontier Model Forum, joining the Coalition for Content Provenance and Authenticity (C2PA) as a steering committee member, joining the Global Anti-Scam Alliance, and our commitment to the Safety by Design Generative AI Principles to prevent AI-generated child sexual abuse and exploitation.

We welcome the opportunity to review our analysis and discuss how we can all benefit from an open web that is safe, transparent, and accountable.

About this Report

Scope and Purpose

This report is issued by Google Ireland Limited. The report and the appendices meet the requirement under Article 42(4) of the DSA that the providers of very large online search engines (VLOSEs) or very large online platforms (VLOPs) make available to the Digital Services Coordinator of establishment and the European Commission a report setting out: (a) the results of the systemic risk assessment undertaken to meet the requirements of Article 34 of the DSA; (b) the mitigation measures consistent with Article 35(1) of the DSA; and (c) information about the consultations conducted in support of the risk assessments and design of the risk mitigation measures.

Article 34 of the DSA requires VLOSEs and VLOPs to identify, analyse, and assess enumerated systemic risks in the EU stemming from the design or functioning of their services and related systems, while Article 35 requires providers of VLOSEs and VLOPs to put in place reasonable, proportionate, and effective mitigation measures to address systemic risks identified in the Article 34 risk assessment.

This is our 2024 (i.e., second) annual report of systemic risk assessments and covers the period July 1st, 2023 - June 30th, 2024. In scope for this report are Google's designated VLOSE (Google Search) and VLOPs (Google Maps, Google Play, Google Shopping, and YouTube). Separately, we also publish VLOSE and VLOP Transparency Reports pursuant to Articles 15, 24, and 42(1) of the DSA.

Structure of the Report

This report has five sections:

- **Background**: We describe how Google uses service design and content moderation to create and maintain services that balance maximising the benefits they provide with minimising potential negative externalities.
- **Methodology:** We outline the six-step methodology used to conduct the systemic risk assessments.
- **Results:** We share the results of the systemic risk assessments for each of our VLOPs (Google Maps, Google Play, Google Shopping, and YouTube) and our VLOSE (Google Search). Each section includes:
    - Discussion of the identification and assessment of the most important inherent and residual risks.

- Description and assessment of our long standing content policies, safety- and private-by-design practices, and other measures designed to mitigate systemic risk.
- Mitigation enhancements that represent additional commitments by Google to further address systemic risk in the EU and, in many cases, globally. Taken in combination with our existing measures, these enhancements help ensure that our mitigations are reasonable, proportionate, and effective, and address the evolving nature of systemic risk.

Throughout each VLOP and VLOSE section, where relevant, we describe how the internal and external factors articulated in Article 34(2) of the DSA and regional or linguistic considerations had an impact on the assessment of risks or mitigations.

- **Conclusion:** We provide observations on the future of systemic risk assessments at Google, in the EU, and beyond.
- **Appendices:** We outline more details about the systemic risk assessment.
  - A complete list of risk statements for each VLOSE and VLOP.
  - A list of the mitigations consistent with Article 35 of the DSA.
  - A list of consultations used in support of the risk assessment and the design of risk mitigation measures.

Our Risk Assessment Methodology

Article 34 of the DSA requires VLOSEs and VLOPs to identify, analyse, and assess systemic risks in the EU stemming from the design or functioning of their services and their related systems or from the use of their services. Absent regulatory guidance, we developed our systemic risk assessment methodology by combining the systemic risk assessment requirements of the DSA with proven risk assessment methodologies, such as those used to assess enterprise risk, human rights risk, compliance risk, and systemic risk assessments in other sectors. We used the same risk assessment methodology in our 2024 assessments as we used in our 2023 assessments, with targeted enhancements that reflect our culture of continuous improvement and build feedback, learning, and growth into our risk assessment process.

Assessing Risk for each VLOSE and VLOP

Step One: Classification. We established 42 "risk statements" across the four categories of systemic risk in Article 34 of the DSA. The risk statements describe the potential adverse impacts for each risk category and provide the focus for each systemic risk assessment.

Step Two: Identification. We identified the risk drivers that may lead to inherent risk for each risk statement and pinpointed the quantitative and qualitative insights needed to assess systemic risk.

Step Three: Assessment of Inherent Risks. We assessed each risk statement to determine the potential severity of the adverse impacts that could arise from that risk and the probability or frequency of occurrence. Combined, these elements produce an estimate of the inherent risk—i.e., the risk absent our risk reduction efforts. In practice, the inherent risk does not reflect actual risk on the service because we launch services with risk mitigations.

Step Four: Assessment of Preparedness. We reviewed the mitigations (e.g., product features, policies, controls, enforcement practices, and other measures) we have in place to address each risk and assessed our level of preparedness, resulting in an estimate of residual risk for each risk statement—i.e., the risk after our mitigation efforts. This evaluation considers the extent to which those mitigations address the probability of occurrence and the potential adverse impact associated with each risk. This assessment of preparedness included, but was not limited to, our progress implementing the additional mitigations we identified in the 2023 assessments consistent with Article 35 of the DSA.

Step Five: Mitigations. We used the results of the risk assessment to identify where additional or enhancements to mitigations are appropriate. We identified these additional measures to ensure that there are reasonable, proportionate, and effective mitigations in place to address systemic risks, consistent with Article 35 of the DSA.

Step Six: Reporting. We disclose the results of the systemic risk assessments in this report, including a discussion of the most important inherent and residual risks and our efforts to address them. We will publish this report (subject to removal of confidential information) in due course, consistent with the requirements of Articles 35 and 42 of the DSA.

We discuss this methodology in more detail in Section 3.

Findings of our 2024 Assessments

Building on our 2023 assessments, we concluded that our mitigation measures address the highest inherent risks and are well-tailored to the purposes of the Google services we assessed. We describe these mitigations throughout the report.

Many themes arising in our 2024 systemic risk assessments align with the findings of our 2023 assessments. For example, we continue to find that service type and the corresponding risk profile is a primary factor in determining the greatest inherent risks, that risk from highly motivated bad actors is a primary cause of concern, and that we invest in programs to address our most significant inherent risks.

Changes in external factors, such as the wider social, political, and technological context of the past year, resulted in some revisions to our inherent risks during the assessment period. For example, generative AI presents powerful benefits to users and society, but the wide public availability of generative AI tools has the potential to lower barriers to the creation of large volumes of content and influences several inherent risks, including fraud, scams, mis- and disinformation, and child sexual abuse material (CSAM). The 2024 assessment period also included the outbreak of the Israel-Hamas war and the continuation of war in Ukraine, both of which influenced our risk profile in the EU. Additionally, there were more elections worldwide in a single year than ever before.

This evolution in inherent risk is to be expected given the changing external context, the ever-expanding tactics of adversarial bad actors, and the transformational capability of new technologies. Our 2024 systemic risk assessments demonstrate that our experience, tools, and methods allow us to adapt quickly to address evolving risks and respond rapidly to unfolding events.

During the assessment period we engaged with external stakeholders (such as civil society organisations, academics, regulators, and consumer organisations) to discuss how we address inherent risk, as encouraged by Recital 90 of the DSA. Topics included our approach to election integrity, the enforcement of content policy in the EU given heightened risks of hate speech, and concerns that generative AI tools may be used to create content that violates the law or our policies. We incorporated these topics in the 2024 assessments.

Our 2024 assessments found lower residual risk (i.e., risk after our mitigation efforts) for some risks compared to our 2023 assessments. In this report we expand upon the following highlights for both cross-Google and service-specific progress:

- **Election Integrity:** We played our part in whole-of-society efforts to achieve high standards of civic discourse by deploying a three-part strategy of (1) surfacing high-quality information, (2) safeguarding our services from abuse that undermines democratic participation, and (3) equipping political campaigns and elected officials with best-in-class security tools and training. Key areas of progress included a “prebunking” approach to countering mis- and disinformation on Search, and updating our Ads Political Content Policy to require prominent disclosure when Election Ads contain synthetic or digitally altered content that inauthentically depicts real or realistic-looking people or events. Read more in Supporting Elections Integrity.

- **Content Provenance:** We supported the development of technical standards for certifying the source and history of media content by joining the Coalition for Content Provenance and Authenticity (C2PA) as a steering committee member, participating in the Partnership on AI's efforts to advance responsible practices for synthetic media, and pioneering our own cutting edge technology for imperceptible but detectable digital watermarking via SynthID. We have also introduced a new tool in YouTube Creator Studio requiring creators to disclose when realistic content is made with altered or synthetic media. Read more in Our Approach to the Disclosure of Synthetic Content.

- **AI Enhancements:** We are pursuing opportunities to use AI to prevent, detect, and respond to illegal and harmful content at scale, such as continuing to invest in the quality and the accuracy of our moderation systems across languages, including languages spoken in the EU. Using large language model technology (LLMs), we are now able to build and train models capable of finding specific kinds of abuse in our services faster than ever before. Read more in Protecting Users with Applied AI.

- **Scams and Fraud:** We joined the Global Anti-Scam Alliance, a large network of organisations committed to protecting consumers from the activities of online fraudsters. We are joining with over 100 organisations, including governments, law enforcement, consumer protection, financial authorities, and internet companies, to address the proliferation of online scams and fraud globally. We also put in place enhanced protections against scams and fraud, such as expanding Limited Ad Serving and onboarding new organisations into our Priority Flagger Program to notify us of potentially harmful scams and fraud issues. Read more in Promoting Trustworthy Content and User Safety.

- **Child Safety:** We further enhanced our child safety protection measures, such as enhancing CSAM deterrents and introducing an age-indeterminate video classifier on Search to improve our ability to detect possible CSAM when the subject's age is ambiguous. We announced our commitment to the Safety by Design Generative AI Principles to prevent the creation, dissemination, and promotion of AI-generated child sexual abuse and exploitation, and worked with the Tech Coalition to launch Lantern, the first cross-platform signal sharing program for companies to strengthen how they enforce their child safety policies. Read more in Child Safety and Generative AI.

- **Search:** We introduced additional mitigations to address non-consensual explicit images (NCEI) and involuntary synthetic pornographic images (ISPI). We reduced residual risk of content promoting practices harmful to health with enhanced mitigations that improve our ability to accurately detect a range of personal crisis searches, such as suicide, substance abuse, self-harm, and eating disorders. Read more in Search.

- **Maps:** We enhanced our appeals channels for potentially erroneous user-generated content removals and feature restrictions, updated our personal information policy that directs users to request content removal if they believe personal information has been posted without their consent, and introduced additional accessibility features. Read more in Maps.

- **Play:** We introduced new developer validation and pre-publication app testing requirements to address risks such as unintentional data sharing, scams, malware, and phishing. We also updated our developer policies for topics such as health apps, photo/video permissions, manipulated media, generative AI, and child safety, and introduced new channels for users to flag reviews that may violate policy and appeal removals that they may disagree with. Read more in Play.

- **Shopping:** We enhanced our methods to review and verify merchant identity-related signals to address the risk of disinformation about businesses. We also implemented infrastructural changes aimed at reducing the time required for detection and enforcement, enabling prompt identification and enforcement action. Read more in Shopping.

- **YouTube:** We implemented various measures to manage risks related to generative AI content including a new labelling tool for creators and a complaints process for AI-generated or other synthetic content that looks or sounds like an individual. We also updated our Strikes System to introduce new, optional training that provides users who have received a warning with more information on specific Community Guidelines policies. If a user completes a training, their warning will expire after 90 days. Read more in YouTube.

We will continue to monitor external changes that influence our inherent risks and establish reasonable, proportionate, and effective measures to address risks. This will include measures that we can take alone and whole-of-society efforts that are best pursued collaboratively with other companies and stakeholders.

2. Background

Maintaining User Trust and Safety

We have a longstanding commitment to examining and addressing the impacts our services can have on societal risks. We have built teams, service protections, tools, and partnerships to address risks arising from the increasing use of the internet by society and risks that may result from the use of our services. We begin with an overview of key teams at Google that work to promote user safety and combat potential harm, then detail our approach to preventing risk at scale.

Our Commitments

Our approach to maintaining user trust and safety on a global scale stems from our focus on how our services affect people and societies. This is reflected in several policy frameworks:

- Our Human Rights Policy and White Paper, which set out our commitment to respecting human rights and upholding the standards established in the United Nations Guiding Principles on Business and Human Rights (UNGPs).
- Our Responsible AI Principles, which describe our commitment to developing technology responsibly and work to establish specific application areas we will not pursue.
- Our Information Quality and Content Report, which outlines the key considerations that guide our product, policy, and enforcement decisions.
- Our Transparency Center, which outlines the policies that help keep users safe from harm and abuse, our reporting and feedback channels, as well as information about how we develop and enforce those policies.
- Our Privacy and Terms Center, which sets out our Privacy Policy, Terms of Service, Privacy and Security Principles, and other relevant guides and resources.
- Our Safety Center, which sets out how we help keep everyone safe online, including as regards content safety, family safety, and cybersecurity.
- Our EU Elections with Google Hub, which provides resources to help campaigns connect with voters and manage their security and digital presence.

We provide regular updates on The Keyword, our official blog for product and technology announcements, news, and stories.

Investing in Systemic Risk Prevention

Each of our services seeks to help users while keeping them safe from potential harms. Within each VLOP and VLOSE are well-developed functions that refine and enforce content policies, and design and maintain features aimed at avoiding and/or mitigating risks to our users. We also invest in another layer of systemic risk mitigation via central cross-service teams that lead our efforts to mitigate specific types of systemic risk associated with our services. These teams include:

- **Trust and Safety:** Our Trust and Safety teams include professional experience in content moderation or sensitive workflows, knowledge in the relevant content matter, and linguistic expertise. The linguistic expertise required varies depending on the specific workflow of a product or service, the type of content, and languages that content is available in. Some products or services require native proficiency in global supported languages. Others may use translation tools, and some videos or images do not require any language proficiency to review (e.g., some static imagery may be subject to our policies even when not accompanied by any written or spoken words). Our teams also monitor emerging trends to address new harm vectors before they can become a larger issue.
    - **Google Trust and Safety:** We pioneered the now industry-wide practice of investing in Trust and Safety specialists who are trained to analyse bad actors, abusive practices, content issues, and the effectiveness of existing policies. Today, our Trust and Safety teams consist of experts, specialists, and engineers working to keep people safe online by using the latest technology (including AI and LLMs) to enforce our policies and moderate content. These teams partner with external experts and teams across Google to carry out our mission to keep people safe online and protect our services and products from abuse.
    - **YouTube Trust and Safety:** YouTube has built its own Trust and Safety team, with expertise in addressing the unique content challenges that arise on an open video-first service. Like Google's company-wide Trust and Safety organisation, YouTube Trust and Safety partners with members of our legal, operations, public policy, product management, and engineering teams to develop innovative ways to combat potentially harmful content. Hundreds of hours of new content are uploaded to YouTube every minute, and we use a combination of people and automated systems to detect problematic content at scale.
- **Kids and Families:** Our Kids and Families program includes a Kids and Family Steering Committee, which brings together executives and leaders from relevant services. It also includes a central team tasked with managing minors' accounts, creating age-appropriate experiences across our services, and advancing child safety protections. The program and its staff have built on years of input from experts and research insights to build tools and features that empower kids and teens while also giving families the ability to exercise choice over their children's relationship with technology. The results are products, features, and policies such as YouTube Kids, Assistant for Families, Family Link, and Google Play Families Policies.
- **Human Rights:** The Human Rights program is a central function responsible for ensuring that we are meeting our human rights commitments across all functions, products, and services. The program advances company-wide strategy on civil and human rights, advises product teams on potential civil

and human rights impacts, conducts human rights due diligence, and engages external experts and stakeholders.

- **Privacy, Safety, and Security:** The Privacy, Safety, and Security (PSS) organisation combats digital threats to users and is committed to keeping the internet as a whole protected. We do this because we are an internet company, and our fate is tied to the fate of the internet. So we do not just design solutions to protect our users, we eliminate entire classes of threats from being effective on our services and products and across the internet.

  PSS comprises industry-leading experts focused on protecting users and data, improving governance and assurance practices related to security, and increasing our technical and operational capabilities. PSS develops and implements automatic protections from bad actors in the data and security space across our services. Our PSS efforts include the following focus points:

  - **Privacy:** Our Privacy program teams drive strategy for and provide leadership on Google's privacy priorities. The central Privacy program teams are responsible for administering privacy policies, training, and documentation that ensure that our products and services protect the privacy of our users. We have also embedded privacy teams and specialists in product areas to ensure that privacy goals are part of product work, and to ensure that we maintain a consistent and high standard of privacy protection and support across the company. Central and product privacy specialists coordinate across the company in working groups that focus on privacy issues that are relevant to particular products or sectors and track best practices and developments relating to particular policy topics. Our privacy subject matter experts also oversee privacy review processes to verify that our services and products vigilantly protect the privacy of our users.

  - **User Protection:** Within our User Protection framework, our Google Threat Intelligence organisation (composed of Google's Threat Analysis Group (TAG) and Mandiant Intelligence teams) is responsible for countering threats from government-backed attackers, coordinated information operations, and serious cybercrime networks. TAG actively monitors threat actors and studies the evolution of their tactics and techniques, using research to continuously improve the safety and security of our products, improve Google's defences, and protect users.

    TAG shares intelligence with our industry peers and publicly releases information about the operations it disrupts via public bulletins highlighting the group's work. For example, TAG has been closely tracking and disrupting campaigns targeting individuals and organisations in Ukraine, and frequently publishes reports on Russian threat actors. The group also works closely with product teams to detect and remove malicious ads, videos, or channels that may be spreading disinformation, malware, or other types of cyber threats.

    TAG and the team at Mandiant Intelligence help identify, monitor, and tackle emerging threats to elections, ranging from coordinated influence operations to cyber espionage campaigns. They meet regularly with experts in industry, academia, and elsewhere to share threat information and suspected election interference, and help organisations build holistic election security programs and harden their defences.

Complementing the work of TAG, our Account and Device Integrity (ADI) team within User Protection keeps users safe by ensuring products interact with legitimate users and devices. ADI's technology works to ensure that accounts and devices have access to Google products and services in ways that are proportional to their demonstrated integrity. In addition, ADI's offering limits opportunities for accounts to be created, compromised, or operated at scale to abuse our products or violate the privacy and security of people who use our services.

- **Civics:** Our dedicated Civics team works across our services, addressing threats to democratic participation in partnership with Trust and Safety specialists. The Civics team oversees products, initiatives, and promotional efforts that aim to safeguard the integrity of elections-related information and provide users with candidate information from authoritative sources. These teams also provide 24/7 support to triage emergent issues during elections.

- **Health:** People frequently come to Google services with health-related questions. Because of the ties between these queries and our users' health and wellbeing, we have prioritised building products to empower people with accurate, actionable health information. To implement this goal, we have recruited experts with decades of experience in health care, public health, and life sciences who help us translate clinical knowledge into product impact. Many of our product areas have policies prohibiting content that contradicts well-established medical consensus, and our Clinical Team helps enforcement teams calibrate medical claims and ensure we are not exposing users to harmful medical mis- and disinformation.

These teams and experts are some of the key groups that partner with other teams across Google to assess and mitigate systemic risks. Their work helps us make good on our commitments to protect users from harm, deliver reliable information, and partner to create a safer internet.

Promoting Trustworthy Content and User Safety

Three core concepts guide how we provide access to trustworthy information and content while keeping users protected.

- **Protect users from harm**. We keep users and society safe through built-in protections utilising the latest technology that enables us to prevent, detect, and respond to illegal and harmful content.
- **Deliver reliable information**. We enable confidence by delivering reliable information and best-in-class tools that give additional context and put users in charge of evaluating content.
- **Partner to create a safer internet**. We scale our industry-leading practices to help keep users safe online through proactive partnership with experts and organisations to both inform and share our resources and technologies.

While we pursue these principles in all of our endeavours, we also recognise that working towards user trust and safety requires constant adaptation to changing social context, evolving threats, and new techniques employed by bad actors. We can never bring the threat of systemic risks to zero, but these principles guide our efforts to constantly increase trust and safety across all of our services.

One: Protecting Users from Harm

We work hard to keep users and society safe through built-in protections that enable us to prevent, detect, and respond to illegal and harmful content.

Preventing Harm with Safety by Design

Our first line of defence is the set of safety features we build into our products to protect user data and prevent abuse.

We present risk assessment results for our four VLOPs and our VLOSE as one report because many of the most effective protections we offer to users are implemented at the Google account level, and these protections are effective across our different service offerings. These account design features protect users whether they are browsing Google Search or downloading books on Google Play. And because we scale privacy and security solutions across all our services, we are able to minimise the number of times our services collect user data and the number of places we store that data.

Clear account settings options, robust account verification, and a secure sign-in process are fundamental to user safety and data security. Strong protections around these processes help guard user data from bad actors, empowering users and their family members to interact with our services the way that they wish. We invest in protecting these processes because they are the primary entry points for many risks. That's why we have developed features like Google's 2-Step verification, which requires a second layer of verification after a user enters a password, and helps guard against compromised passwords.

During 2023, we began rolling out passkeys across Google Accounts as an easier and more secure way to sign in to apps and websites, and a major step towards a "passwordless" future. Passkeys let users sign in to apps and sites the same way they unlock their devices: with a fingerprint, a face scan, or a screen lock PIN. Unlike passwords, passkeys are resistant to online attacks like phishing, making them more secure than solutions like SMS one-time codes.

We also apply protections for signed-in and signed-out users who we believe are minors, and have engineered easy management of ads preferences and privacy settings through the My Ads Center. These protections, and many others, are designed as an integral part of our services, making it simple and quick for our users to benefit from advanced security infrastructure.

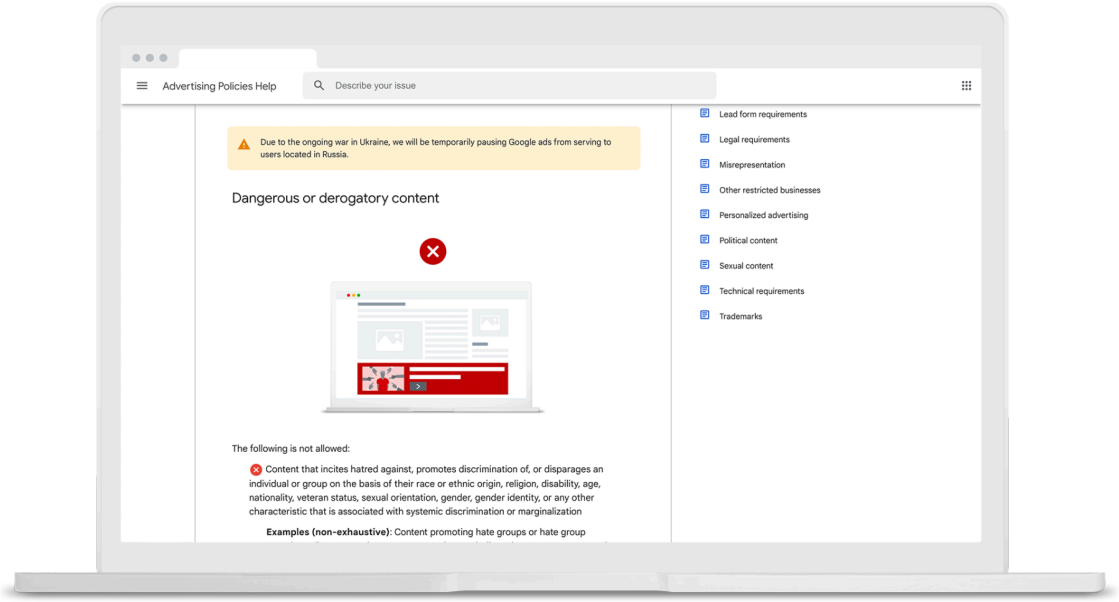
We also build services that consider safety at the outset and incorporate safety considerations into service design. For example, on Play, we reduce the risk of "review bombing" and sham ratings by using a percentage of the most recent reviews, not the average of all ratings, to determine the overall rating for an app.

Building Products that are Private by Design for Everyone

When people use Google, they trust us with their information, so we work hard to protect that information, as outlined in the Google Privacy Policy. We set clear expectations with the Google Terms of Service to help define our relationship with users when they interact with our services. Protecting user privacy and security is a responsibility that comes with creating products and services that are accessible for all. We have developed a set of Privacy and Security Principles to guide our products and processes to keep sensitive data private, safe, and secure.

Preparing for the Unexpected

Other protections focus on less likely but potentially serious events. These policies, like many others described in this report, permit us to more nimbly respond to unexpected events. For example, Google Ads' sensitive events framework is designed to prevent ads that potentially profit from or exploit a sensitive event, such as a natural disaster, public health emergency, act of terrorism, conflict, or act of mass violence. We disallow ads that seek to profit from a tragic event with no discernible benefit to users, engage in price gouging that restricts access to vital supplies, or use keywords related to a sensitive event to drive traffic. We enforced our sensitive event framework in response to both the war in Ukraine and the Israel-Hamas war by prohibiting ads that exploit, dismiss, or condone the wars.



Ads prohibited under our Sensitive Events policy

Designing Appropriate Content Policies

We design content policies across our services to protect users from harm and observe a high standard of quality and reliability for advertisers, publishers, and content creators. Our content policies, which are publicly available, articulate the purpose and intended use of each service to which they apply. They explain what types of content are allowed to be created, uploaded, sent, shared, and monetised, and the process by which a piece of content, or the user responsible for it, may be removed from a given service. We regularly update our content policies as our services evolve and new threats arise. You can find them here for Google Search, Google Maps, Google Play, Google Shopping, and YouTube. Additionally, these services may present ads, for which there are distinct Google Ads content policies.

We carefully tailor the rules about allowable content on each service according to the core purpose of that service and available levers to enforce the rules. While our content policies share the same goal to keep all users safe while respecting the right to freedom of expression by only making necessary and proportionate removals, they may differ across Google for several reasons:

- **Hosted versus indexed content:** When we are hosting content versus when we are indexing content.
- **Public versus private content:** When we enable public dissemination versus when privacy expectations are higher.
- **Recommended content:** When there are features that organise or suggest content to users.
- **Direct versus indirect services:** When we own a service versus when we only provide the infrastructure for someone else.
- **Monetisation:** When a feature enables monetisation or facilitates transactions.

For example, Search is intended to facilitate the exploration of a broad range of information from a wide variety of sources on the open web. Search's objective is to maximise access to information, and we remove web results from Search in only very limited and clearly defined circumstances. When listings and other information are presented as Search features (like featured snippets), however, users may interpret the information as having greater quality or credibility, and we apply more restrictive policies.

Our advertising services have policies that restrict certain types of harmful content because we do not believe the digital advertising ecosystem should profit from the sale of harmful content or experiences. Similarly, because Maps is designed to be a source of reliable information about places and experiences, its policies place a greater emphasis on accuracy, authenticity, and relevance.

Additionally, YouTube values freedom of expression and is built on the premise of openness. Its policies aim to support the interest of its creators and their incredible array of diverse voices and perspectives. YouTube is committed to protecting its community from harmful content, while giving creators the freedom to share a broad range of experiences and perspectives through video. Because YouTube hosts and serves user-generated content, it has unique content policies.

Reviewing Content Policies and Practices

Promoting high-quality content and responding to harmful content is a dynamic challenge that requires constant adaptation. Data, law, world events, experts, and user feedback all inform our policy development by helping us to identify emerging harms and gaps in our existing policies. This is part of a four-phase policy development process that each VLOP and VLOSE engages in with respect to all of its policies:

- **Identify emerging trends and novel safety concerns:** Our expert teams assess emerging issues and consider external feedback.
- **Gather examples and identify common themes:** We review evidence of harm, identify common themes, and view similar examples.
- **Draft policy standards and enforcement guidance:** We assess impact, determine enforcement mechanisms, and consult experts.
- **Assess effectiveness and continuously review:** We launch, test, and refine policy, assess effectiveness, and review enforcement actions and appeals.

For example, YouTube regularly reviews its policies to make sure that they are effective at preventing real-world harm, and to ensure they properly address changes occurring both on and off our service. YouTube works directly with civil society organisations, academics, and relevant experts with varying viewpoints and from different countries to inform this policy review. Much of YouTube's work on content policies, which we call the YouTube Community Guidelines, focuses on analysing, assessing, and addressing emerging issues before they reach, or become widespread on, YouTube. Similarly, as risks change and evolve, so do our content policies for Maps (e.g., fake engagement, misrepresentation, and misinformation policies), Play (e.g., user-generated content policies), Search (e.g., highly personal information), and Shopping (e.g., vehicle ads).

Balancing Risks and Rights

Fundamental rights are interdependent. The fulfilment of one right (e.g., freedom of expression) may facilitate the fulfilment of other rights (e.g., civic participation and democracy) or come at the expense of others (e.g., freedom from discrimination).

As a result, fundamental rights are sometimes in tension with each other.[1] For example, the pursuit of child safety may limit adult users' rights and present risks to different rights held by children, such as their rights to participation, privacy, and freedom of expression and information. We address these tensions through various means, such as providing parents or guardians with

[1] This is recognised in the DSA. Recital 153 of the DSA states that "in situations where the relevant fundamental rights conflict, a fair balance between the rights concerned, in accordance with the principle of proportionality" should be achieved.

controls that allow them to supervise minors' access to content, and giving users extensive controls over their privacy settings.

When efforts to protect or advance one right may result in the limitation of another right, our approach is to identify and implement sensible mitigation measures to address potential adverse impacts to both rights. This balancing involves considering appropriate and proportionate mitigation techniques, such as protecting freedom of expression via appeals mechanisms or raising authoritative content to address other lower quality content that may exist on the service.

Throughout the VLOSE- and VLOP-specific sections of this report, we explain why one risk may take precedence over another in certain circumstances, describe how the nature and purpose of the service being assessed inform these choices, and set out the reasonable and proportionate mitigations we believe strike the right balance.

Detecting and Responding to Harmful Content at Scale

In every country in which we operate, different laws govern what is considered permissible expression. To address these nuances, we have teams and systematic processes to develop and deploy localised policies and enforcement practices. When users report content they believe violates the law on our services, we carefully review whether to block, limit, or remove access to it.

The enforcement of content policy involves both human review and an array of technologies, including automated systems that use machine learning technology, working together to achieve high levels of accuracy when reviewing content. We design models and train classifiers[2] to identify potentially violative content, use machine learning to constantly improve those classifiers, take automated actions when that content violates our policies, and enqueue content for review by specialist teams when we have lower confidence in fully automated techniques. These human content moderators help confirm whether machine-identified content should be removed, and we use the results of the human review to further train our classifiers and improve their ability to detect evolving violative content.

This collaborative approach helps improve the accuracy of our models over time, as models continuously learn and adapt based on human feedback. And it also means our enforcement systems can manage the significant scale of content that's available on our services, while still rendering nuanced decisions on whether a piece of content violates our policies. Examples of automated systems and humans working in combination are provided in each of the VLOSE and VLOP sections that follow.

[2] A classifier is an algorithm that identifies and sorts content into different categories. For example, a classifier may identify content likely to violate a specific Google policy.

Handling Government Removal Requests

Courts and government agencies around the world regularly request that we remove user-generated content from our services. We were the first company to publish (in 2010) a formal transparency report about such requests. You can read more about our process and the volume of requests we receive in the Government Requests to Remove Content segment of our latest Transparency Report.

In addition, we use automated systems and reporting channels to detect content that may violate our content policies (such as YouTube's Community Guidelines) and remove that content if it is found to violate them. Some content that violates our policies may also be illegal.

We maintain a robust process to receive, evaluate, and act on government removal requests. We review these requests closely to confirm that they are supported by local laws and international norms of human rights and to determine whether we should remove content as a matter of national law or our platform-wide policies. Consistent with our commitment to the Global Network Initiative Principles, we assess the legitimacy and completeness of government requests, which must be in writing, made through appropriate channels, as specific as possible about the content to be removed, and clear in their explanation of how the content is illegal.

In some narrow cases, to protect the rights of users, we do not act on orders that appear illegitimate or inapplicable. For example, we examine the legitimacy of documents we receive, and if we determine that a court order is forged, we won't comply with it. In other cases, we do not need to take action because the content has already been removed by the uploader.

Types of Enforcement Actions

We take a wide range of enforcement actions on our services to support information and content quality and maintain a trusted experience for all. Enforcement actions, like policies, differ from service to service and are tailored to the purpose of each service, based on what is reasonable, proportionate, and effective, taking into consideration the appropriate approach to freedom of expression for each service. Action may be taken on the content, such as limiting monetisation opportunities, restricting access to content via interstitials,[3] or removing content from the service entirely; action may also be taken at the account level if the issue is serious enough, such as temporarily pausing an account owner's ability to access a service or, in the case of serious violations, disabling an account entirely.

[3] In this context, an interstitial is an element (e.g., a blurred image or a warning) that appears before the desired content is displayed, such as before shocking, disturbing, or graphic content.

Protecting Users with Applied AI

We are pursuing opportunities to use LLMs to prevent, detect, and respond to illegal and harmful content at scale.

Using LLMs, we are now able to build and train models capable of finding specific kinds of abuse in our services faster than ever before. This is especially valuable for new and emerging abuse areas, such as a new narrative or derogatory term, where we can now quickly prototype a model to detect potential abuse and automatically route it to our teams for enforcement. We frequently evaluate both the accuracy (i.e., precision) and completeness (i.e., recall) of both our automated and human review systems.

LLMs can rapidly review and interpret content at a high volume, while also capturing important nuances within that content, and these advanced reasoning capabilities have already resulted in larger-scale and more precise enforcement decisions on some of our Ads policies. One example is our policy against Unreliable Financial Claims, which includes ads promoting get-rich-quick schemes. The bad actors behind these types of ads adjust their tactics and tailor ads around new financial services or products, such as investment advice or digital currencies, to scam users. While traditional machine learning models are trained to detect these policy violations, LLMs are more capable of addressing the fast-paced and ever-changing nature of financial trends, differentiating between legitimate and fake services, and quickly scaling our automated enforcement systems to combat scams. This has helped our teams become even more nimble in confronting emerging threats of all kinds.

We are still testing these new techniques and deployment remains at an early stage, but they have demonstrated impressive results so far. This effectiveness results from both the power of Google's LLMs and the quality of our abuse experts who provide "seed" intelligence and tune the models to ensure the results are reliable. We anticipate this type of innovation significantly advancing our efforts to protect our users at scale across Google products, especially from new and emerging risks. We also believe that these advances will both improve our ability to identify problematic content and reduce the need for human moderators to review disturbing content.

Offering Appeals

We heavily invest in the training of automated tools and human content reviewers to increase accuracy. However, sometimes we make errors and remove content on our services that does not actually violate our policies. In many cases, appeals channels are an appropriate way to fulfil our commitment to freedom of expression and to the UN Guiding Principles on Business and Human Rights by providing a check against incorrect removal and ensuring that content creators have redress.

Our appeals process for removals aims to ensure due process, efficiency, and transparency for users appealing our enforcement decisions, without facilitating abuse by bad actors. Different services provide different methods to appeal, and while users often can access appeals forms via their violation notification, we present a non-exhaustive list of appeals forms on our website. Our VLOP services allow users to appeal enforcement actions they believe may have occurred in error. We address incoming appeal requests as quickly as possible to clarify requirements and help users understand the actions taken on their account.

Whether an appeal is valid requires a case-by-case determination. In some cases, consideration of other equities counsels against providing appeals, such as those involving repeat or abusive violators, ancillary content, or egregious conduct.

We seek to ensure that these mechanisms are accessible and work to learn from appeals outcomes, including making our content moderation efforts more accurate. Insights gained from these appeals processes also inform policy changes to prevent future adverse impacts. Sections of this report specific to each VLOP will describe our appeals mechanisms and where we are expanding them to mitigate risks to freedom of expression.

Reporting

In addition to our own review and legal removal requests and user flags of illegal content (described above), we offer a variety of mechanisms for users to report and request removal of policy violating content. For example, Maps users can flag content that may violate our policies or profiles of users who are contributing false information, uploading offensive content, or taking other abusive actions. On YouTube, users can flag videos that may violate our policies. Trained content moderators then review credible flags and take appropriate action, which may result in content being removed, age-restricted, geo-restricted, or left up.

While the option to report will be found in the service itself or within its policies, we have also created a dedicated page to help users find ways to report harmful content on several of our services.

Our approach to flagging also involves partnering with other organisations. Our Priority Flagger Program provides channels for participating organisations to notify us of potentially harmful issues on certain of our products and services that violate our policies. We use a dedicated intake channel to expedite review of potential policy violations. We also participate in ongoing discussions and feedback about Google and YouTube content policies. This program is most suitable for organisations such as non-governmental organisations (NGOs) and government agencies with an identified expertise in recognising and fighting harm online in at least one policy area.

Our Public Interest Framework guides policy and enforcement decision-making by safeguarding against improper content removals. The framework helps us consider (1) how content, if not removed, could adversely impact the rights of an individual, community, or society as a whole, or (2) whether allowing the content is in the public interest because it furthers the understanding of social, political, cultural, civic, or economic affairs.

Child Safety and Generative AI

Generative AI may be used by bad actors to create new outputs that exacerbate some existing child safety risks and introduce new risks. There are three risks we are particularly focused on addressing:

- AI-generated CSAM or computer-generated imagery depicting child sexual abuse, including new material that has never been seen before, or the editing of either benign or abusive images of real children.
- The sexualisation of children across a range of modalities, such as graphic sexual stories involving children, or images that may not be illegal but objectify and sexualise children.
- The use of generative AI to support other child sexual abuse behaviours, such as providing text instructions on how to carry out abuse, supporting offenders to groom or sextort children, or promoting or normalising sexual interest in children.

We have policies addressing each of these risks. The key is to ensure that we can detect this sort of content effectively at scale even when it is AI-generated and can work in collaboration with others to stay ahead of these risks.

Our child safety experts rigorously test our generative AI products before launch, and we also seek to address generative AI-created CSAM appearing elsewhere across the web. Our work to detect, remove, and report CSAM has always included violative content involving actual minors, modified imagery of an identifiable minor engaging in sexually explicit conduct, and computer-generated imagery that is indistinguishable from an actual minor engaging in such conduct.

We are working to ensure that our Child Safety Toolkit performs well with AI-generated CSAM. We are also proactively engaging with child safety experts from industry, academia, government, and civil society, such as at our recent "Growing Up in the Digital Age" summit hosted at the Google Safety Engineering Center (GSEC) in Dublin.

We are also committed to the Safety by Design Generative AI Principles, developed by Thorn and All Tech is Human, to help make it as difficult as possible for bad actors to misuse generative AI to produce content that depicts or represents the sexual abuse of children. The mitigations set out in these principles complement our existing work to prevent the creation, dissemination, and promotion of AI-generated child sexual abuse and exploitation.

During the assessment period, Google.org committed €5M to non-profit organisations with the goal of helping teens understand AI so they can use it safely, as part of Google.org's commitment to support AI training and skills in Europe. We will continue to support organisations who focus on

reaching teens who are most underserved as well as their support networks, such as educators and parents.

Evaluating Content Across Languages

During the assessment period many external stakeholders asked about our ability to undertake effective content moderation across languages spoken in the EU, including the effectiveness of both automated systems and human reviewers. Given the important role played by automated systems, we continue to assess the risk that algorithms may be less well trained in some languages, dialects, and vernaculars than others.

We deploy automated systems that detect violating content and behaviour at scale; however, human operators are often required to review, validate, and train these automated systems because humans can evaluate content or other signals in ways that might be difficult for current automated systems, such as understanding nuance, context, and slang.

Taken together, Google services maintain Trust and Safety coverage, including human content moderators, across official EU languages, as well as many other languages commonly spoken in the EU. Our EU DSA Biannual VLOSE/VLOP Transparency Report provides data for our human resources evaluating content across the official EU Member State languages, segmented by service.

One important element of this assessment is the review of how significant advances in machine translation assist with review of content at scale. We use both human content moderators and machine learning to constantly train these classifiers and improve their ability to accurately detect such content across different languages, dialects, and vernaculars. Google strives for a high degree of accuracy for automated content moderation across EU languages and reports accuracy broken down by language where available. For instance, from March 2023 to August 2023, the accuracy level for all automated content moderation decisions for Maps was 90%. Given the operational challenge of having content moderators available 24/7 for less widely used languages, use of these tools enables us to undertake moderation of content at scale more rapidly, consistently, and effectively, while relying on human review for close calls.

Our systemic risk assessments found some residual risk remaining for the performance of automated systems across languages, dialects, and vernaculars for all VLOP and VLOSE services reviewed. Google-wide advances (such as continuous improvement in machine translation or the development of LLMs that can learn quickly across multiple languages) will over time support the

work of different Google services as each develops custom models, thresholds, and confidence levels tailored to their own policy enforcement needs.

During the assessment period, we refreshed our models to improve the accuracy of our moderation systems across languages, improving our general translation quality between English and German, French, Italian, Portuguese, Dutch, Polish, Turkish, Arabic, Russian, and Spanish.

Going forward, we will continue to test the performance of classifiers to identify differences in performance across languages and pursue continuous improvement, including a rolling program to identify priority languages for investment in enhancing translation and content moderation quality.

We will also continue to keep pace with developments in local contexts—including how language and terminology may evolve with potential for higher-risk events, such as elections—and use human content moderators and native speakers to improve the quality of automated systems, including classifiers.

Two: Delivering Reliable Information

Providing access to high-quality information to all users is core to our mission. We also provide users with best-in-class tools that give additional context that help them evaluate content.

Surfacing High-Quality Information

The world wide web holds an unprecedented, and growing, volume of information that is not ordered or easily navigable. But automated systems operating at scale to sort, organise, and deliver relevant information can help users find the needles in the world's largest haystack.

Algorithms power our services by prioritising relevant information in search results, making app recommendations on Play, and providing relevant product listings in Shopping. Our algorithms sort through hundreds of billions of pieces of content to find the most relevant and useful results.

Algorithms enable us to advance quality and relevance while reducing systemic risk to users and society. Search uses signals such as meaning, relevance, quality, usability, and context to help determine which results are returned and prioritised for each query. Our systems use these and hundreds of other signals to prioritise the results that seem most helpful, in particular content that seems to demonstrate expertise, experience, authoritativeness, and trustworthiness. These signals are especially important for what we call "Your Money or Your Life" (YMYL) topics, defined as those that may significantly impact or affect the health, financial stability, or safety of individual people, or the welfare of society.

To help us test and improve our Search algorithms we continue to put all possible changes to Search through a rigorous evaluation process to analyse metrics and decide whether to implement a proposed change. We work with external Search Quality Raters to evaluate if our Search systems are generating helpful results that demonstrate experience, expertise, authoritativeness, and trustworthiness.

This overall approach is summarised in How Search Works.

Using Recommender Systems

Some risk factors under Article 34(2), such as the use of recommender systems, may increase or decrease risk. Poorly designed or controlled recommender systems may increase the risk that harmful content goes viral. But properly functioning recommender systems should decrease risk by increasing the visibility of high-quality and trustworthy content and by promoting a diversity of topics and sources for users to explore.

Recommender systems are an essential tool as we navigate the inherent tensions that come with respecting countervailing fundamental rights while fulfilling our mission to organise the world's

information and make it universally accessible and useful. We aim to make all of our recommendations useful, inclusive, and empowering.

Using recommender systems to order the presentation of content, including by elevating high-quality and trustworthy content, is often a more proportionate approach to addressing harmful content risk than removing content altogether, which can present risks to freedom of expression and information.

Fighting Misinformation and Disinformation

We continue to invest heavily in elevating authoritative sources and countering mis- and disinformation, particularly as it relates to people's finances, health, livelihood, or civic participation and to sensitive events. Mis- and disinformation can manifest itself in different ways on different services across the open web, such as misleading pages attempting to monetise their content with our Ads services, health mis- and disinformation videos on YouTube, or websites spreading mis- and disinformation appearing in search results. Other examples of mis- and disinformation practices include fraud, deceptive behaviour (such as the use of deep fakes), impersonation, misrepresentation of ownership, and medical mis- and disinformation. We take action to prevent the spread of this type of content at scale.

For example, during a breaking news cycle, speculation, and mis- and disinformation can outrun facts while legitimate news outlets are still investigating. Bad actors may publish content with the intent to mislead, or to attract attention and traffic on the basis of unverified information. To defend against these risks, YouTube and Search have systems designed to promote authoritative content.

We have long recognised the importance of multi-stakeholder approaches to mis-and disinformation, including the EU's 2018 Code of Practice on Disinformation and a Strengthened Code that Google signed in June 2022. As part of the Strengthened Code, we have committed to providing the European Commission with reports detailing how we have implemented our commitments. Our commitment to the Strengthened Code applies to Search, YouTube, and Google Ads, and you can read more about it in the Search and YouTube sections of this report.

Our commitment to fighting mis- and disinformation guided our reaction to the COVID-19 crisis. We enacted policy revisions, stepped up enforcement, and raised the visibility of authoritative content through features such as Health Knowledge Panels and structured search results designed to make trusted information easy to access.

We are continuing to monitor the threat landscape in Eastern Europe, disrupt coordinated influence operations from threat actors, and remove content and YouTube channels that violate our Community Guidelines and Terms of Service. Our breaking news and top news shelves on the YouTube homepage continue to receive tens of millions of views in Ukraine and, as the largest video-sharing service in Russia, YouTube continues to provide Russian citizens uncensored news and information.

Supporting Election Integrity

We are committed to playing our part in whole-of-society efforts to achieve high standards of civic discourse and election integrity. We are focused on the following three priorities:

- **Surfacing high-quality information:** Supporting the democratic process by connecting people to reliable, trustworthy, and high-quality information.
- **Safeguarding our services and disrupting the spread of mis- and disinformation:** Developing and enforcing policies that safeguard our services from abuse that undermines democratic participation and trust in the civic process.
- **Equipping campaigns:** Providing political campaigns and elected officials with best-in-class security tools and training for election-related security challenges.

All three priorities include a focus on risks and opportunities presented by AI to the information landscape related to elections.

This strategy builds on work we do around elections in other countries and regions, and the commitments we made in the EU Code of Practice on Disinformation. We believe our approach to reinforcing internal processes that mitigate election-related risks aligns well with the recommendations made in the Communication from the European Commission set out in Guidelines for providers of Very Large Online Platforms and Very Large Online Search Engines on the mitigation of systemic risks for electoral processes pursuant to the Digital Services Act.

Surfacing high-quality information

During the assessment period, we worked to ensure that election-related information was high-quality.

For example, people searching for topics like “how to vote” or “how to register” found details such as ID requirements, registration, voting deadlines, information on voting abroad, and guidance for different means of voting, like in person or via mail. This was available in 22 languages and included a country selector for the 12 million people who are eligible to vote in a different country than the one they live in. From the beginning of April 2024 to the end of the EU Parliamentary elections, these features were viewed over 55 million times.

During the 2024 European Parliamentary elections, YouTube made it easy to find election content from authoritative sources in search results and on prominent news shelves. Election information panels made it easier for people to register to vote, find out how to vote, learn more about the candidates, and access election results. In total, these dedicated election features appeared to YouTube viewers across the EU over 1.1 billion times.

All advertisers running election ads in the EU on our services are required to go through a verification process and have an in-ad disclosure that clearly shows who paid for the ad. These ads are published in our Political Ads Transparency Report, where anyone can look up information such as how much was spent and where it was shown. We also limit how advertisers can target election ads.

Keeping advertisements free from abuse is an important safeguard for the democratic process, so it is important to note that our existing ads policies outline a variety of other advertisements that are never allowed, including manipulated media intended to deceive, defraud, or mislead others, ads with unreliable claims, and ads that are exploitative during sensitive events, such as civil emergencies, natural disasters, or conflict.

Our Google Trends Elections Hub for the European Parliamentary Election provided EU-wide Search Trends and country-level data for Germany, France, Poland, and Spain. The hub featured real-time charts that provided an overview of how Search interest in the parties, candidates, and political topics evolved over time, based on Google Trends data.

Safeguarding our platforms and disrupting the spread of mis- and disinformation

We continue to enhance our enforcement systems and invest in trust and safety operations to better secure our services and prevent abuse, including policies around demonstrably false claims that could undermine democratic processes.

To help enforce our policies, our AI models are enhancing our abuse-fighting efforts with faster and more adaptable enforcement systems that enable us to remain nimble and act even more quickly when new threats emerge. You can read more in Protecting Users with Applied AI.

However, fighting mis- and disinformation online requires effort across the whole ecosystem. We contributed €25 million to help launch the European Media and Information Fund, an effort designed to strengthen media literacy and fight misinformation across Europe. The Fund has since backed 70 projects across 24 countries, including fact-checking during elections and critical events and improving the media literacy of harder-to-reach populations. We also support the Global Fact Check Fund and numerous civil society, research, and media literacy efforts from partners, including TechSoup Europe, the Civic Resilience Initiative, Baltic Centre for Media Excellence, and the Central European Digital Media Observatory (CEDMO).

We made a €1.5 million contribution to the European Fact-Checking Standards Network (EFCSN), a newly created association representing European fact checking organisations, to launch Elections24Check, a coalition of 40+ news and fact-checking organisations working together to fact-check the European Parliamentary Elections. Elections24Check created a comprehensive database of election-related disinformation, claims, and narratives to support research and fact-checking around the world.

Our approach to disrupting the spread of mis- and disinformation in Europe has been enhanced by "prebunking," which shows audiences how to spot common manipulation techniques so they can better recognise mis- and disinformation online. Our prebunking initiative was developed with local experts and partners to tackle key manipulation techniques identified in our research: scapegoating, discrediting, and decontextualisation. The initiative included short video ads in France, Germany, Italy, Belgium and Poland, which reached over 120 million people across those countries and were translated into all EU languages, as well as Arabic, Russian, and Turkish.

The prebunking initiative also highlighted programs from partners, along with Google and YouTube products and features that help people investigate what they see online, including fact-checking features

on Search like About This Result and YouTube's Hit Pause media literacy resources. The European Parliament, European Digital Media Observatory (EDMO), and the EFCSN highlighted information and media literacy resources to help people tackle disinformation via a landing page that was available in all EU languages.

Implementation of these collaborative efforts was informed by a convening of experts across government, academia, civil society, and industry at our election-focused Fighting Misinformation Online summit in Brussels, organised with our partners the European University Institute and the Calouste Gulbenkian Foundation. This convening discussed key topics including media literacy, civic engagement, mis- and disinformation in the context of AI, and the importance of working together.

Mis- and disinformation during the EU parliamentary elections was focused on undermining trust in institutions and governments at the national level, with some targeting of the EU itself, though we observed only limited volumes of content around political violence. We took action in instances where content, including deepfake content, violated our policies. For example, as of June 9, YouTube had terminated over 1,000 channels and removed over 140 EU election-related videos in 2024 for violating YouTube's Community Guidelines, including policies around manipulated content and misattributed footage.

Equipping campaigns and candidates with best-in-class security features and training

Elections come with increased cybersecurity risks, so we have been helping high-risk users—such as campaigns and election officials—to improve their security in light of existing and emerging threats and informing them on how to use our products and services.

We offer free services like our Advanced Protection Program and Project Shield, which provides unlimited protection against Distributed Denial of Service (DDoS) attacks. We also partner with Possible, The International Foundation for Electoral Systems (IFES), and Deutschland sicher im Netz (DSIN) to scale account security training and provide security tools including Titan Security Keys, which defend against phishing attacks and prevent bad actors from accessing Google Accounts.

Google Threat Intelligence also helps identify, monitor, and tackle emerging threats, such as coordinated influence operations and cyber espionage campaigns against high-risk entities. We report on actions taken in our quarterly TAG bulletin and meet regularly with government officials and others in the industry to share threat information and suspected election interference. Mandiant also helps organisations build holistic election security programs and harden their defences with comprehensive solutions, services, and tools, including proactive exposure management, proactive intelligence threat hunts, cyber crisis communication services, and threat intelligence tracking of information operations.

For example, in May 2024, working with Google's Threat Analysis Group, YouTube terminated 21 channels as part of our ongoing investigation into disinformation campaigns linked to Russia. The channels shared content in various EU languages with narratives critical of the domestic conditions in EU countries and US/EU financial aid for the war in Ukraine. By the time polls closed, as part of our proactive coverage for the European Parliamentary elections, YouTube had terminated 240 channels for being part of coordinated influence operations targeting the EU.

We also launched an EU-specific hub at euelections.withgoogle with resources and training to help campaigns connect with voters and manage their security and digital presence.

Helping people navigate AI-generated content

Like any emerging technology, AI presents new opportunities as well as challenges. For example, generative AI makes it easier than ever to create new content, but can also raise questions about the trustworthiness of information. There are three risks we are particularly focused on in the context of elections:

- Use of generative AI to deceive the public about matters of civic and public concern in a way that could harm overall discourse about elections, though not directly interfering with it.
- Use of generative AI tools as a force multiplier for bad actors (state-backed or otherwise) to identify new attack vectors or make existing attack vectors more efficient/scalable.
- Use of generative AI tools to interfere with the democratic process, either by disrupting the process itself, or by disrupting voter intent—for example, disinformation about an election, such as a fake image of a long line at a polling station.

One important step we took to help people navigate AI-generated content was in the realm of political advertising. We expanded our political content policies to require all election advertisers to prominently disclose when their election ads include synthetic content that inauthentically depicts real or realistic-looking people or events, including image, video, and audio content, becoming the first company to do so. This disclosure must be clear and conspicuous and must be in a location where it is likely to be noticed by users. Our ads policies already prohibit the use of manipulated media, like deepfakes or doctored content, to mislead people.

We also supported the cross-industry Tech Accord to Combat the Deceptive Use of AI in 2024 Elections, a set of commitments to deploy technology countering harmful AI-generated content meant to deceive voters and work collaboratively on tools to detect and address online distribution of harmful AI content. We joined the Coalition for Content Provenance and Authenticity (C2PA) coalition and standard, a cross-industry effort to help provide more transparency and context for people on AI-generated content.

Other parts of Google's approach are described elsewhere in this report, including content labels on YouTube to indicate altered or synthetic content, providing users with additional context (such as About This Image in Search), tackling spammy, low-quality content on Search, and SynthID, a tool from Google DeepMind that directly embeds a digital watermark into images, text, audio, and video generated with Google's AI tools.

Addressing the Risks and Opportunities of Artificial Intelligence

In 2017 we announced our intention to be an "AI-first company", and we wholeheartedly believe AI has the potential to transform our societies for the good. We continue to develop artificial intelligence tools to help solve some of society's biggest challenges. AI is embedded in many of our services, such as on Maps where we are cutting carbon emissions by reducing stop-and-go traffic.

AI also presents important challenges that must be addressed clearly, thoughtfully, and affirmatively. In 2018 we set out our AI Principles and accompanying framework for responsible AI innovation that describe our commitment to developing technology responsibly and the specific application areas we will not pursue.

The recent momentum behind large-scale machine-learning models (including generative AI) has sparked additional dialogue around the impacts to society of AI and surfaced concerns, which we discuss in more detail throughout this report.

The opportunities and challenges presented by large-scale machine-learning models require global, multi-stakeholder, and collaborative approaches. For this reason we are a founder or active participant in several new initiatives, such as:

- The Frontier Model Forum, a new industry body focused on ensuring safe and responsible development of frontier AI models.
- The Partnership on AI's (PAI) Responsible Practices for Synthetic Media: A Framework for Collective Action, an initiative to foster best practices in the development, creation, and sharing of media created with generative AI, and PAI's Guidance for Safe Foundation Model Deployment.
- The White House's Office of Science and Technology Policy initiative to ensure safe, secure, and trustworthy AI.
- The Coalition for Content Provenance and Authenticity (C2PA), an effort to develop technical standards for certifying the source and history (or provenance) of media content.

In these systemic risk assessments we considered the risks to our services presented by the use of generative AI by bad actors, such as influence campaigns, scams, phishing, cyberattacks, and spam. Over time, it is possible that the development of large-scale machine-learning models will alter the scale and possible severity of some risks, especially those relating to the generation of illegal or harmful content (such as child abuse and exploitation content, terrorist and violent extremist content, and hate speech); misinformation and disinformation relating to elections, civic

discourse, and democratic participation; and digital threats such as account hijackings, phishing attempts, or malware.

We believe our existing mitigations perform well, but we must keep pace with the latest developments in AI technology. Additional mitigations (see Our Approach to the Disclosure of Synthetic Content below) include integrating watermarking, metadata, and other innovative techniques into our latest generative AI models and an About this image tool within Search to give users context about where an image first appeared online and whether a SynthID watermark or IPTC metadata is detected. However, large-scale machine-learning models are evolving rapidly, and we expect to assess the impact of these developments across all relevant risks in our future systemic risk assessments.

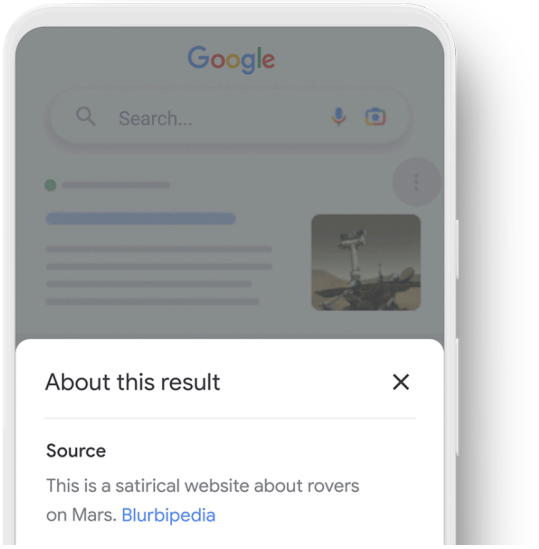
We have made a set of voluntary commitments—developed jointly with other leading AI companies and the White House's Office of Science and Technology Policy—to promote the safe, secure, and transparent development and use of AI technology. These commitments will support efforts by the G7, the OECD, and national governments to maximise AI's benefits and minimise its risks.

More detail about our overall approach to addressing the risks and opportunities of AI is described in our recent paper End-to-end responsibility: A lifecycle approach to AI. Building on our previous efforts, this paper describes our four-phased AI Responsibility Lifecycle process (Research, Design, Govern, Share) that guides responsible AI development at Google. In the paper we share insights into how we developed this process, with recent examples and practical tips for implementation.

Our Secure AI Framework (SAIF) is also designed to address top-of-mind concerns for security professionals, such as AI/ML model risk management, security, and privacy. This helps to ensure that when AI models are implemented, they are secure-by-default.

Equipping Users

In addition to raising the visibility of high-quality information and fighting mis- and disinformation, we aim to equip users with the tools they need to evaluate information they come across on our services.



The 'About this result' feature in Google Search allows users to learn more about the information they are seeing

For example, our "About this result" feature in Search allows users to learn more about where the information they are seeing is coming from and how our systems determined it would be useful for their query. This feature is available in all languages where Search is available. With this context, users can make more informed decisions about the sites they may want to visit and what results will be most helpful to them. Similarly, My Ad Center gives users greater control of the ads they see on Google services—like Search and YouTube—by providing options for customising ads, managing privacy settings, and influencing how we determine what ads to show.

Be Internet Awesome is a Google-created educational program that teaches kids the fundamentals of digital citizenship and safety so they can explore the online world with confidence, such as how to communicate responsibly, discerning between what's real and what's fake, and safeguarding valuable information.

On Google Play, the Teacher Approved program identifies apps approved by teachers and children's education specialists, and then offers a description of the apps' quality attributes. This information helps families easily review the apps and make informed choices about whether they want their children using an app or game.

Informed users are able to make better use of and see more benefit from our services, a win for both our users and us.

Our Approach to the Disclosure of Synthetic Content

Generative AI is a type of machine learning model that can help generate new content, including text, images, audio, video, and code. While it is important to note that generative AI content is not inherently a systemic risk, we know that AI will introduce new risks and will require new approaches.

Computer-generated content, also known as "synthetic content," is becoming harder to distinguish from content that has not been created by an AI system. This creates new challenges for the trustworthiness of content across the web, and we are committed to helping users understand and evaluate what they find online.

For example, we have launched SynthID, a tool for watermarking and identifying AI-generated content. This technology embeds a digital watermark directly into the images, text, audio, and video generated with Google's AI tools in a way that is imperceptible to the human eye but detectable for identification. We are not exposing SynthID labels to users, but the ability to identify computer-generated content is one method for helping prevent the spread of mis- and disinformation.

In addition, any content published by Lyria—our AI music generation model used by YouTube—will be watermarked with SynthID. Similar to how it handles images, SynthID embeds a watermark into AI-generated audio content that is inaudible to the human ear and doesn't compromise the listening experience. This watermark is designed to maintain detectability even when the audio content undergoes modifications such as noise additions, compression, or speeding up and slowing down the track.

However, synthetic content is also created using non-Google products, and for this reason it is important for Google to participate in collective efforts to shape AI responsibly and address shared challenges.

In February 2024 we joined the Coalition for Content Provenance and Authenticity (C2PA) as a steering committee member. The C2PA is a cross-industry effort to address the prevalence of misleading information online through the development of technical standards for certifying the source and history (or provenance) of media content. We will support its work by helping to develop its technical standard and support the development and adoption of Content Credentials, a new type of tamper-resistant metadata that provides an interoperable method to share information about how content was made and edited over time.

We continue to participate in the Partnership on AI's Responsible Practices for Synthetic Media: A Framework for Collective Action, an initiative to foster best practices in the development, creation, and sharing of media created with generative AI.

A variety of other resources also help users distinguish synthetic content, such as About this image, Fact Check Explorer, and Search with an Image. YouTube now requires creators to disclose to viewers when realistic content—content a viewer could easily mistake for a real person, place, scene, or event—is made with altered or synthetic media, including generative AI and the alteration is not inconsequential.

Despite these efforts, it is important to acknowledge that bad actors can and do seek to remove or hide watermarking when they have reason to believe it will restrict their misuse of a generative AI product. Strategies to address the new risks arising from generative AI should be multi-faceted and not limited to efforts focused on the disclosure of synthetic content.

Three: Partnering to Create a Safer Internet

We recognise that the systemic risks associated with VLOSEs and VLOPs are not unique to Google and cannot be addressed by Google alone. We scale our industry-leading practices to help keep users safe online through proactive partnership with experts and organisations to both inform and share our resources and technologies.

Partnering for Information Quality

To effectively combat mis- and disinformation, technology companies already collaborate with academics, policymakers, publishers, and civil society organisations who possess the expertise that helps inform effective methods to address the issue at scale.

During 2023 we initiated long-term partnerships across Central and Eastern Europe, a region considered highly vulnerable to disinformation and propaganda due to its geographic proximity to the war in Ukraine. In the Baltics, we have entered into a long-term partnership with the Civic Resilience Initiative and the Baltic Center for Media Excellence. These two established and well-respected organisations have received €1.3 million in funding from Google to build on their impactful work towards increasing media literacy, building further resilience, and actively tackling disinformation in Lithuania, Latvia, and Estonia. We are partnering with the Charles University in Prague, the main research centre of the Central European Digital Media Observatory (CEDMO) project, and providing €1 million in funding for CEDMO to further expand its research into information disorders (such as misinformation, disinformation, or clickbait) and work to increase the level of media and digital literacy in Poland, Czechia, and Slovakia.

This year, with many countries in Europe lowering the age of voting to 16 years old, Google.org is awarding a $1 million grant to ThinkYoung to fund youth-led hackathons across Europe, empowering young voters to combat disinformation and develop solutions with a focus on underserved communities.

Google.org also launched a €15M open call for European nonprofits, civic entities, academic institutions, and social enterprises to help scale initiatives promoting democratic resilience in the region, including through the use of advanced technology and AI. This builds on Google.org's longstanding commitment to youth media literacy and online safety, with support for more than 60 organisations in this field since 2018.

We worked closely with civil society to tackle misinformation ahead of the 2024 European elections by partnering with a diverse range of organisations across the EU. These partnerships helped develop and disseminate prebunking initiatives designed to increase media literacy among voters. This included organisations like Libraries without Borders, Debating Europe, BBC Media Action, AFP, European Parliament, EDMO, and the European Fact-Checking Standards Network (EFCSN).

Also during 2024 we made a €1.5 million contribution to the EFCSN, a newly created association representing European fact-checking organisations, to launch Elections24Check, a coalition of more than 40 news and fact-checking organisations working together to fact-check the European Parliamentary Elections. Elections24Check created a comprehensive database of election related disinformation, claims, and narratives to support research and fact-checking around the world.

Consulting with Experts

We scale our industry-leading practices to help keep users safe online through proactive partnership with experts and organisations to both shape and share our resources and technologies.

In 2023, we launched the Trust and Safety Research Awards, providing unrestricted grants to support research efforts across areas of interest related to trust and safety in technology. We granted a total of $600,000 to nine applicants—including six where the principal investigators represented universities in Europe—and in 2024 we expanded the grant value to $1 million.

Also in 2024, we hosted research workshops and roundtables with nearly 200 researchers from academia and civil society, covering topics of safety by design, child safety, and the use of AI in content moderation.

The Google Safety Engineering Center (GSEC) in Dublin is a regional hub for Google experts working to tackle the spread of illegal and harmful content, and a place where we can share this work with policymakers, researchers, and regulators. Between July 2023 and June 2024, GSEC Dublin held approximately 50 public and private engagements to share our experience of managing content risk and hear from experts across a wide range of topics, including misinformation, ads safety, election integrity, the use of AI in content, and fighting child sexual abuse and exploitation online. For example:

- As part of our ongoing support for the people of Ukraine, GSEC Dublin conducted several Fighting Misinformation Online roundtables and summits with local governments, NGOs, and fact-checking organisations across Central and Eastern Europe. We shared Google and YouTube's approach to mis- and disinformation, and gained real-time insights from over 100 local experts and organisations.

- In October 2023, we conducted the Fighting Misinformation Online Summit, connecting Members of the European Parliament, Google.org grantees, European civil society groups, and others to discuss progress in using LLMs in content moderation, announce our Trust and Safety Research Awards and awardees, and hold workshops on relevant concerns, such as elections, the Israel-Hamas war, and combating mis- and disinformation.

- In March 2024, we hosted a Fighting Misinformation Online: Elections 2024 summit in Brussels, where delegates working across the misinformation landscape came together to discuss the most pressing and urgent issues.

- In April 2024, Google, YouTube, and partners convened over 300 civil society groups, academics, and policymakers from across Central and Eastern Europe in Warsaw for Fighting Misinformation Online: Election Integrity in the Age of AI. The full-day event, hosted by Polish public news channel TVP Info, featured almost 50 speakers from 19 countries. Discussions focused on collaborative strategies to strengthen ecosystem efforts to minimise misinformation ahead of 2024 elections and other critical moments, as well as Google and YouTube efforts to tackle misinformation online.

- Our Fighting Misinformation Online event series aims to tackle mis- and disinformation by bringing together stakeholders from industries and nations in Europe to collaborate, share knowledge, and debate. The series, hosted by the European University Institute, Calouste Gulbenkian Foundation,

Google, and YouTube, launched with the first summit in October 2021, when we announced Google's inaugural contribution of €25 million to the launch of the European Media and Information Fund.

- In March 2024, GSEC Dublin convened experts from industry, academia, government and civil society at a Growing up in the Digital Age summit to continue the dialogue around the protection of young people online. Key themes included strengthening privacy and safety, promoting healthy digital habits, and helping teens and families navigate new technologies confidently and safely.

YouTube regularly updates its family product experiences and policies in consultation with experts in children's media, child development, digital learning, and citizenship from a range of academic, non-profit, and clinical backgrounds. A key channel for this consultation is YouTube's Youth and Families Advisory Committee, a collection of independent experts who provide advice on the policies and services YouTube offers to young people and families.

YouTube also sponsored the National Academy of Medicine to convene an independent advisory group to develop principles and attributes to guide digital services companies in identifying and elevating credible sources of health information in their channels. The outcome of this project was a peer-reviewed discussion paper and the use of these principles when providing content from reliable health sources on Google.

Sharing Tools and Technology

We also share tools to help organisations protect platforms and users, including safety APIs across child safety, security (such as cyber attacks, malware, and phishing), and information quality (such as mis- and disinformation, toxic discourse, and explicit content).

For example, our Child Safety Toolkit consists of two APIs: the Content Safety API (which classifies previously unseen images of potential child sexual abuse and exploitation) and CSAI Match (which matches known abusive video segments). We offer these APIs to qualifying partners free of charge. Our partners use these technologies to process billions of files each year, allowing them to evaluate millions of images and videos for abusive behaviour and prioritise the most concerning content for review.

Perspective API (which uses machine learning to identify "toxic" comments, making it easier to host better conversations online) and Harassment Manager (an open-source codebase that allows users to document and manage abuse targeted at them on social media) help journalists, activists, politicians, and other public figures document and manage abusive comments on their sites.

Collaborating with Companies and Stakeholders

Many of the risks reviewed in this systemic risk assessment cannot be addressed by a single company acting alone, so we have established and continue to fund and participate in a mix of multi-company and multi-stakeholder efforts that take system-wide approaches to the most intractable problems. This includes sharing signals of illegal and harmful content, collaborating with civil society to gain deeper insights into risk, and sharing best practices across companies.

- **Global Internet Forum to Counter Terrorism (GIFCT):** In 2017, YouTube co-founded GIFCT with a group of companies dedicated to disrupting terrorist abuse of members' digital platforms. GIFCT provides a formal structure to accelerate and strengthen our work and present a united front against the online dissemination of terrorist content, such as by identifying and sharing signals of terrorist and violent extremist activity via the GIFCT hash sharing database.

  During 2023 we worked with GIFCT, Tech Against Terrorism, and the Christchurch Call to launch Altitude, a new free and open-source tool to help small- and medium-sized online platforms protect their communities from terrorist and violent extremist content. Altitude provides online platforms a single elevated view of potential terrorist and violent extremist content on their sites, helping them to triage and remove the content. In 2024, we formally handed Altitude over to Tech Against Terrorism, which will continue its development and maintenance.

- **Tech Coalition (TC):** In 2006, we joined the Tech Coalition, teaming up with other tech industry companies to develop technical solutions that disrupt the ability to use the Internet to exploit children or distribute CSAM. For example, we have been one of two members to test a system to increase the chances of detecting CSAM videos through hash matching, while our child safety experts also chair or actively participate in half a dozen key working groups of the Tech Coalition.

  During 2023, we were among the first wave of tech platforms to join Tech Coalition's Lantern, the first cross-platform signal sharing program for companies to strengthen how they enforce their child safety policies. Until now, no consistent procedure existed for companies to collaborate against predatory actors evading detection across services, and Lantern fills this gap.

- **Global Network Initiative (GNI):** We were a founding member of the GNI in 2008, and since then we have worked closely with civil society, academics, investors, and industry peers to protect and advance freedom of expression and privacy globally, especially when faced with demands from governments that conflict with international human rights standards. GNI brings together academics, civil society, companies, and investors around thematic workstreams, including enabling shared learning on key trends and emerging developments in the technology sector.

- **Global Anti-Scam Alliance (GASA):** During 2024 we joined the Global Anti-Scam Alliance, a large network of over 100 organisations committed to protecting consumers from the activities of online fraudsters. We will link up with hundreds of organisations, including governments, law enforcement, consumer protection, financial authorities, and internet companies, to address the proliferation of online scams and fraud globally.

- **Coalition for Content Provenance and Authenticity (C2PA):** During 2024 we joined the Coalition for Content Provenance and Authenticity as a steering committee member. The C2PA is a cross-industry effort to address the prevalence of misleading information online through the development of technical standards for certifying the source and history (or provenance) of media content.

Developing Best Practices

We actively participate in efforts to develop best practices that advance responsible and effective approaches to risk assessment across the industry, as well as to develop the field of trust and safety more broadly.

- **Digital Trust and Safety Partnership (DTSP):** In 2021 we co-founded the DTSP alongside nine other companies. The DTSP is committed to developing industry best practices, verified through internal and independent third-party assessments, to ensure consumer trust and safety when using digital services.

- **Partnership on AI (PAI):** In 2016 we were a co-founder of the Partnership on AI, a non-profit partnership of academic, civil society, industry, and media organisations helping AI advance positive outcomes for people and society. We are also a member of PAI's Responsible Practices for Synthetic Media: A Framework for Collective Action, which is fostering expertise and best practices for responsibility in the development, creation, and sharing of media created with generative AI.

- **World Economic Forum:** We have been an active participant in the Global Coalition for Digital Safety to accelerate public-private cooperation to tackle harmful content online, exchange best practices for new online safety regulation, take coordinated action to reduce the risk of online harms, and drive forward collaboration on programs to enhance digital media literacy. We have also participated in the AI Governance Alliance to foster inclusive, ethical, and sustainable AI across industries.

- **Trust and Safety Professional Association:** We are a founding supporter of the Trust and Safety Professional Association, a non-partisan membership association that supports the global community of professionals who develop and enforce principles, policies, and practices that define acceptable behaviour and content online and/or facilitated by digital technologies.

- **Frontier Model Forum:** During 2023 we partnered with Anthropic, Microsoft, and OpenAI to launch the Frontier Model Forum, an industry body focused on ensuring safe and responsible development of frontier AI models. The Frontier Model Forum draws on the technical and operational expertise of its member companies to benefit the entire AI ecosystem, advancing AI safety research and supporting efforts to develop AI applications to meet society's most-pressing needs.

- **MLCommons:** We participate in this engineering consortium to measure and improve the accuracy, safety, speed, efficiency, and safety of AI technologies, helping companies and universities around the world build better AI systems that will benefit society. For example, we contributed to the proof of concept for the first AI Safety Benchmark.

- **UN B-Tech Project:** We participate in this UN Human Rights project that provides authoritative guidance and resources for implementing the UN Guiding Principles on Business and Human Rights in the technology industry. Launched in 2019 after consultations with civil society, businesses, states, and other experts, the B-Tech project has four strategic focus areas: (1) human rights risks in business models; (2) human rights due diligence and end use; (3) accountability and remedy; and (4) regulatory and policy responses to the human rights impacts of technology.

Setting High Standards for Advertising

We strive to create a healthy, trustworthy, and transparent digital advertising ecosystem that supports users and advertisers. Our advertising policies are designed to ensure a safe and positive experience for our users. This may include prohibiting content that is harmful to users and the overall advertising ecosystem.

Our advertising policies and review processes help address risks across our VLOSE and our VLOPs and cover four broad areas:

- **Prohibited content:** Content that cannot be advertised, such as counterfeit goods, dangerous products, and inappropriate content.
- **Prohibited practices:** Practices that advertisers may not engage in, including those related to misrepresentation and data collection and use. We also prohibit ads that would abuse the Google ad network by trying to circumvent or bypass our ad review processes.
- **Restricted content and features:** Topics that are sometimes legally or culturally sensitive, such as alcohol, gambling and games, healthcare and medicines, financial services, and political content. This content can be promoted, but on a limited basis.
- **Editorial and technical:** Editorial standards for ads and destination requirements for websites and apps.

We set a high standard of quality and reliability for advertisers. We have processes in place to identify bad ads on our services and to monitor violations on an ongoing basis.

We do not permit advertisers to run personalised ads on content designated as "made for kids," and we maintain a separate Ads & made for kids content policy. Advertisers may not run personalised ads, make use of any third-party trackers, or otherwise attempt to collect personal information from minors or on content designated as made for kids. Advertising that is intended for children or is on content designated as made for kids must not be deceptive, unfair, or inappropriate for its intended audience.

This review covers the content of ads, including the ad content (e.g., text, images, video, audio of ad), targeting (e.g., search keywords), and destination (e.g., advertiser's web page). Ads that do not comply with Google policies are disapproved and are not able to run until the policy violation is fixed and the ad is reviewed again. Users can also report ads that they find inappropriate or which they believe may violate the law or Google policies.

Accounts may be suspended if we detect an egregious violation. In 2023, we removed over 5.5 billion ads, restricted[4] over 6.9 billion ads, and suspended over 12.7 million advertiser accounts worldwide, nearly double the total from the previous year.

For repeat violations of an Ads policy, we issue strikes to the Google Ads account and penalties progressively increase with each subsequent strike leading up to account suspension. To address the risk of over-enforcement, advertisers can appeal potentially erroneous ad reviews, strikes, and suspensions.

In addition, in the EU, our Google Ads Transparency Center provides a searchable hub of all served ads and is designed to give users more information about the ads they see on Google services. In the Ads Transparency Center, users can see the ads an advertiser has run, find out which ads were shown in a certain region, and learn more about the advertiser.

We are also committed to delivering ads responsibly in ways that respect user privacy, which we seek to achieve by applying the following five privacy principles to our ads business:

1. We never sell your personal information to anyone. This includes for ads purposes.

2. We are transparent about what data we collect and why. We clearly label ads and sponsored content on our services and make it easy for users to understand why specific ads are shown, what information is used, and how users can control their Google ad experience.

3. We make it easy for users to control their personal information. My Ad Center allows users to customise their ad experiences on Google services. Ads personalisation can be turned off altogether, and activity data tied to an account can be permanently deleted at any time.

4. We reduce the data we use to further protect users' privacy. We never use sensitive information like health, race, religion, or sexual orientation to tailor ads, and never use the content users create and store in apps like Drive, Gmail, and Photos for ads purposes. We do not allow ads personalisation for users where we know that they are under 18.

5. We protect users by building products that are secure by default. We verify advertisers globally and work to detect bad actors and limit their attempts to misrepresent themselves.

During the assessment period we also announced an update to our Ads Political Content Policy to require all verified election advertisers to prominently disclose when their ads contain synthetic or digitally altered content that inauthentically depicts real or realistic-looking people or events. Their disclosure must be clear and conspicuous, and placed in a location where it is likely to be noticed by users. Other changes made to our Ads policies during the assessment period are available here.

[4] Restricted ads are legally or culturally sensitive and can run only in limited contexts.

To learn more about our commitment to maintaining a responsible advertising service, see our Ads Safety Report.

3. Methodology

Introduction

Article 34 of the DSA requires providers of VLOSEs and VLOPs to identify, analyse, and assess systemic risks in the EU stemming from the design or functioning of their services and their related systems or from the use made of their services. The DSA requires that these systemic risk assessments are undertaken annually and prior to deploying functionalities that are likely to have a critical impact on systemic risks.

We have undertaken a separate systemic risk assessment for each Google service designated as a VLOSE (Google Search) or VLOP (Google Maps, Google Play, Google Shopping, and YouTube).

The DSA enumerates four categories of systemic risks to be addressed:

A. The dissemination of illegal content.

B. Any actual or foreseeable negative effects for the exercise of fundamental rights enshrined in the Charter of Fundamental Rights of the European Union (EU Charter), in particular human dignity, privacy, data protection, freedom of expression and information, non-discrimination, rights of the child, and consumer protection.

C. Any actual or foreseeable negative effects on civic discourse and electoral processes, and public security.

D. Any actual or foreseeable negative effects in relation to gender-based violence, the protection of public health and minors, and serious negative consequences to the person's physical and mental wellbeing.

Absent regulatory guidance, we developed our systemic risk assessment methodology by combining the specific systemic risk assessment requirements of the DSA with proven risk assessment methodologies, such as those used to assess enterprise risk, human rights risk, compliance risk, and systemic risk assessments in other sectors. In the 2024 assessments we incorporated several enhancements (described below) in pursuit of continuous improvement.

To help ensure that our methodology was sound and executed well, we retained the services of two consultancies with expertise in risk assessments of different kinds. Each reviewed and contributed to the development of our risk assessment, bringing points of view from their respective fields. Teams from Business for Social Responsibility (BSR), with extensive experience in the field of human rights assessments, and KPMG, with expertise in systemic risk assessments in the financial, energy, and pharmaceutical sectors, contributed to the development of our systemic risk assessment methodology as well as the execution of the assessment.

Importantly, we designed our systemic risk assessment methodology to identify and prioritise risk to people and society, rather than risk to business objectives.

Our methodology has six steps:

Our systemic risk assessment methodology

Step One: Classifying Risk

In this step we established a list of 42 “risk statements" across the four categories of systemic risk. The risk statements are plain-language articulations of the potential adverse impacts for each risk category and provide the core focus for each VLOSE and VLOP systemic risk assessment.

By using specific risk statements, we were able to assess related risks that may require different mitigations, or competing risks that need to be balanced against each other. Risk statements are commonly used in risk assessments to clarify the scope of the risk assessment and focus risk assessor responses on specific exposures.

We relied on insights from a range of internal and external sources to generate these risk statements, including human rights due diligence, outputs from external stakeholder engagement (e.g., surveys, dialogue, roundtables), literature review, and discussions with relevant teams, staff, and subject matter experts at Google.

While the 2024 risk statements are substantially the same as 2023, some enhancements were made to reduce duplication and ensure our risk statements continue to capture all the systemic risks present on each platform. The complete list of 42 risk statements can be found in Annex A.

We also ensured full coverage by reviewing the risk statements against all articles in the EU Charter and cross-referencing against rights contained in international human rights instruments.

We created 39 risk statements to apply to each VLOSE and VLOP, and added a small number of VLOP-specific risk statements where unique service features warranted it. Specifically, because Play is a service that hosts other apps, we created three additional risk statements to determine whether those app offerings were adequately diverse to serve many demographic groups.

Step Two: Identifying Risk

In this step we identified the risk drivers that may lead to inherent risk for each risk statement relevant to each VLOSE and VLOP, and pinpointed the quantitative and qualitative insights supporting the assessment of systemic risk. This included establishing clarity on the purpose, function, and features of each VLOSE and VLOP, including the volume and type of content and the service's potential contribution to the virality of content.

The systemic risk assessment included a mix of quantitative factors (e.g., the number of actions taken against illegal or policy violative content) and qualitative factors that required professional judgement (e.g., the remediability of a privacy violation or the severity of hate speech).

Engaging Stakeholders

Recital 90 of the DSA sets out the expectation that providers of VLOSEs and VLOPs engage with external stakeholders (such as representatives of the recipients of the service, representatives of groups potentially impacted by their services, independent experts, and civil society organisations) when undertaking risk assessments and designing mitigation measures.

We have long engaged external stakeholders to provide expertise related to emerging and evolving issues that intersect with our services. This input is important to the business, and helps inform our decision-making, our due diligence efforts regarding human rights obligations, and our design of mitigation measures.

Consistent with Recital 90, we systematically catalogued and synthesised insights relevant for each risk statement from these engagements, as well as engagements focused specifically on the systemic risk assessments required by the DSA:

- **Google's Government Affairs and Public Policy** engagements with independent experts and civil society organisations for due diligence, decision-making, and strategy.
- **Google's Human Rights Program** stakeholder engagement on human rights issues with civil society and human rights organisations, government officials, and others.
- Engagements through the **Google Safety Engineering Center** (GSEC) led by **Google's Trust and Safety** team with policymakers, researchers, and regulators with an interest in Google's content policy and its enforcement. During the 2024 assessment period, GSEC held several convenings that informed the risk assessments, including on the topics of growing up in the digital age, misinformation and elections, and AI.
- Foundational research led by the **Google Trust and Safety Research** organisation on ecosystem-wide technology risks and policy issues, including information generated via direct user feedback and insights. The team's portfolio includes reports that encompass time series studies, deep investigations, and secondary research for rapid decision-making.
- Google's participation in relevant multi-stakeholder and multi-company efforts, such as the **Global Network Initiative** (GNI), the **Global Internet Forum to Counter Terrorism** (GIFCT), **Partnership on AI**, the **Tech Coalition**, the **Family Online Safety Institute**, the **Content Authenticity Initiative**, the **Global Anti-Scam Alliance**, the **Frontier Model Forum**, the **World Economic Forum AI Governance Alliance**, **MLCommons**, and the **Digital Trust and Safety Partnership** (DTSP).

- In June 2024, the DTSP and GNI hosted the EU Rights and Risks Forum at Google's Brussels' office to bring together representatives from VLOPs and VLOSEs, as well as civil society and academic experts from across Europe and other jurisdictions to discuss systemic risk assessments as provided for in the DSA. Over the course of two days of panels and workshops, participants explored DSA risk assessments and their potential impact on fundamental rights. The Forum was designed to inform approaches to risk assessment and stakeholder engagement but not to constitute the totality of any company's assessment or engagement. Google covered the costs of hosting the conference and travel for staff and civil society participants, with additional travel support provided by TikTok. This forum built upon two multi-stakeholder convenings hosted by the GNI and DTSP during 2023 to discuss methodological questions, such as the definition of systemic risk and risk assessment methodology, and substantive issues, such as fundamental rights, illegal content, civic discourse, and gender-based violence.

- Engagements that **content policy teams** within VLOSEs and VLOPs have with civil society organisations, academics, and relevant third party experts to inform the review, development, and enforcement of content policy and get ahead of emerging issues.

- YouTube's **Youth and Families Advisory Committee**, a collection of independent experts that provide advice on the policies and services YouTube offers to young people and families.

- **User engagements** facilitated by marketing functions and specific product teams to test service features or understand user sentiments about Google and its services.

Taken collectively, the insights gleaned from external stakeholder engagements like these help ensure that the risk statements developed for the assessment appropriately address the categories of systemic risk identified in the DSA and inform our assessment of inherent and residual risk. The insights also informed the development of additional mitigations consistent with Article 35 of the DSA.

The perspectives of external stakeholders were formally recorded alongside each risk statement and contributed to the overall mix of information that was used to assess systemic risk and design mitigations. The insights gained tended to be qualitative rather than quantitative in nature and were especially useful for assessing remediability and preparedness.

Many of the issues raised by external stakeholders are ongoing issues, such as concerns relating to freedom of expression, content policy enforcement, and removing illegal content. However, some of the themes more specific to this assessment period included:

- Elevated scam and fraud risk, including scams arising from the use of generative AI.

- Mis- and disinformation, including heath, conflict, and elections mis- and disinformation, and the impact of generative AI on this mis- and disinformation.

- Image-based abuse, including content created using generative AI tools (such as CSAM, NCEI, and ISPI) and gender-based/LGBTQIA+ harassment, hate, and bullying.
- Enforcement of content policy during times of conflict.
- Information and content moderation quality across geography and language.
- Cross-platform and collaborative approaches to systemic risk.
- Use of generative AI to support improved content moderation.

Step Three: Assessing Inherent Risks

In this step we assessed each risk statement according to the potential severity of the negative effects that the risk could cause and the probability or frequency of the risk's occurrence. Combined, these elements produce an estimate of the ***inherent*** risk—the risk absent our risk reduction efforts. This step is necessarily a theoretical and high-level estimate because we have long been dedicated to mitigating all the systemic risks identified in the DSA. That estimate was then used in the later steps as the foundation to review how well we address each risk. This enabled us to understand the ***inherent*** systemic risks that could stem from the design, functioning, and use of VLOSE and VLOP services, as well as from potential misuses by others.

This step included two important elements:

First, we considered whether the following factors set out in the DSA would impact the risk positively, negatively, or both:[5]

A. Design of recommender systems and any other relevant algorithmic system

B. Content moderation systems

C. Applicable terms and conditions and their enforcement (e.g., content policy)

D. Systems for selecting and presenting advertisements

E. Data-related practices of the provider

F. Intentional manipulation of the service, including by inauthentic use or automated exploitation of the service

G. Amplification and potentially rapid and wide dissemination of illegal and policy-violating content

We also considered whether linguistic or regional differences could affect the risk or any of the above factors.

[5] *See* Article 34(2) of the DSA.

Second, we used the quantitative and qualitative metrics and insights pinpointed in Step Two (Identification) to assess each risk statement according to the following objective criteria:[6]

- **Severity**, meaning the potential consequences of the risk for people and societies, as defined by two criteria:
    - **Scope:** The number of users and/or persons who could be primarily affected by the risk—for example, we considered whether the risk would impact all users of the service or only a subset, and whether the risk would impact non-users as well as users of the service.
    - **Remediability:** The potential to reverse the impact of the risk were it to occur—for example, we considered the adverse impacts on physical, mental, or financial wellbeing, and whether a post-hoc remedy could restore those affected to their condition prior to the impact.
- **Probability:** The likelihood and frequency of the risk—for example, the prevalence and potential virality of violative content, the volume of cases or data involved, or the number of successful appeals. The assessment of probability in particular drew upon quantitative data contained in our various Transparency Reports, with the 2024 assessment benefiting from quantitative data published in our EU DSA Biannual VLOSE/VLOP Transparency Report.

In line with human rights guidance and risk assessment best practices, we used a weighting system so that severity rather than probability would be the predominant factor, meaning that "high severity/low probability" risks received higher prioritisation than "low severity/high probability" risks.

These inherent risks do not actually manifest in our services because each of our services take steps (as described below) to mitigate these inherent risks.

The inherent risk associated with specific risk statements may have altered between the 2023 and 2024 assessments owing to changes to the external context, such as the evolving geopolitical environment, growth in availability of AI tools, and new insights from external stakeholders. To reflect these changes, we refined some risk statements and undertook a calibration exercise to maintain consistency in our analysis of inherent risk across our services.

Step Four: Assessing Preparedness

In this step, we reviewed the mitigations (e.g., policies, controls, enforcement practices, and other measures) we have in place to address each risk and assessed the level of our preparedness, resulting in an estimate of residual risk (i.e., the risk after our mitigation efforts) for each risk statement.

To achieve this estimate, we identified controls and other measures that contribute to our preparedness including (1) the existence and coverage of design decisions, features, policies, processes, metrics, accountability, and formal controls, and (2) other relevant measures, such as participation in industry and

[6] Recital 79 of the DSA states: "In determining the significance of potential negative effects and impacts, providers should consider the severity of the potential impact and the probability of all such systemic risks. For example, they could assess whether the potential negative impact can affect a large number of persons, its potential irreversibility, or how difficult it is to remedy and restore the situation prevailing prior to the potential impact."

multi-stakeholder efforts to address risks. Ultimately, we considered the extent to which the combination of mitigations prevents or significantly addresses adverse impacts of the risk.

Many of the factors the DSA directs to be considered, and which we incorporated in our assessment of inherent risk (such as recommender systems, content systems, terms and conditions, and systems for selecting and presenting advertisements) are also important measures for addressing risk, so we also considered them in our determination of preparedness.

The residual risk associated with specific risk statements may have altered between the 2023 and 2024 assessments owing to changes in inherent risk (described above) or changes to preparedness, such as implementing the additional mitigations we identified in the 2023 assessments consistent with Article 35 of the DSA, other new mitigations established during the assessment period, and continuous improvement.

A discussion of the most important residual risks for each VLOSE and VLOP is found in the results section below.

Taking a Human Rights-Based Approach

We have long been committed to respecting the rights enshrined in the Universal Declaration of Human Rights and its implementing treaties, and to undertaking human rights due diligence (including human rights assessments) using methods based on the United Nations Guiding Principles on Business and Human Rights (UNGPs).

The systemic risk assessment requirement of the DSA shares a common goal with the ongoing human rights due diligence processes we undertake to fulfil our commitment to upholding the UNGPs.

For example, the DSA requirement that a systemic risk assessment consider actual or foreseeable negative effects for the exercise of fundamental rights enshrined in the EU Charter is very similar to the UNGPs expectation that companies assess any actual or potential adverse human rights impacts using internationally recognised human rights as a reference point. Other elements of systemic risk assessment (such as impacts on civic discourse, electoral processes, public security, gender-based violence, public health, and physical and mental wellbeing) are also clearly relevant to ongoing human rights due diligence.

While we designed our systemic risk assessment methodology to meet the requirements of the DSA, we were able to build upon our prior experience undertaking ongoing human rights due diligence based on the UNGPs. This included the generation of risk statements, which was informed by our prior ongoing human rights due diligence, and the creation of assessment criteria, which were based on the notions of severity and likelihood used during human rights due diligence.

Step Five: Identifying Improvements to Mitigations

In this step, we used the results of the risk assessment to identify where additional or enhancements to mitigations are needed. We identified these measures to ensure that there were reasonable, proportionate, and effective mitigations in place to address the specific systemic risks we identified, consistent with Article 35 of the DSA. To ensure successful execution, these mitigations are being tracked and monitored through established, formal governance processes.

This step concluded the systemic risk assessment and mitigations process. We calibrated the results across our VLOPs and VLOSE to ensure the consistent application of the methodology, and they were approved by central Google stakeholders as well as risk owners and leadership teams from each VLOP and VLOSE. Our Independent Compliance Function ensured that the risks were properly identified and reported, and that identified risk mitigations were reasonable, proportionate, and effective.

A complete list of the mitigation enhancements we established consistent with Article 35 of the DSA in our 2024 assessments can be found in Annex B.

Step Six: Reporting the Results

We disclose the results of the systemic risk assessments in this report. We will publish these reports (subject to removal of confidential information) in due course, consistent with the requirements of Articles 35 and 42 of the DSA.

Some information in this report is confidential or security-sensitive. This includes specific discussion of vulnerabilities, or details of security and programs that could be abused by bad actors. We reserve the right to remove this information from the publicly available version of this report, as contemplated by Article 42(5) of the DSA.

4. Results of the Assessments

Dedicated sections below contain the 2024 systemic risk assessment results for each VLOSE (Google Search) and VLOP (Google Maps, Google Play, Google Shopping, and YouTube). Those VLOP and VLOSE sections each describe:

- The service and its associated systemic risk profile based on its use.
- A summary of assessment results, emphasizing elevated inherent and residual risks, and describing any significant changes compared to the 2023 assessment results.
- The existing mitigations such as content and service design choices that address systemic risk, with mitigation enhancements introduced consistent with Article 35 of the DSA listed in Annex B. Taken in combination, existing and enhanced mitigations are intended to be reasonable, proportionate, and effective for the risk being addressed.

Four important observations emerged across the five systemic risk assessments.

First, the purpose of a service is a primary factor in determining the greatest inherent risks. For example, services prioritising broad access to information (such as Search) had lower risks to freedom of expression and higher risks associated with potentially harmful content; services oriented toward a narrower purpose (such as Maps) had higher risks to freedom of expression and lower risks associated with potentially harmful content. Product and content policies are tailored to allow or disallow certain types of content and conduct based on this purpose.

Second, the highest evaluations of preparedness (i.e., our existing mitigations) generally correlated with the most significant inherent risks, confirming that we are appropriately allocating resources to the most significant risks.

Third, despite some increases in inherent risk between 2023 and 2024, our existing and enhanced mitigations have resulted in lower residual risk overall. Important factors influencing increased inherent risk include the determination of bad actors to manipulate recent elections, the growing public availability of generative AI tools, and worsening geopolitical conflict globally. Important factors influencing reduced residual risk include substantial efforts to protect election integrity, new and enhanced user reporting and appeals channels, and the use of AI enhancements to improve the effectiveness of our content moderation.

Fourth, and despite our existing and additional measures, risk from highly motivated bad actors continues to be of concern in connection with mis- and disinformation related to civics, public health, and fraudulent business, as well as external digital threats such as fraud, malware, scams, and malicious sharing of private information. Notable shared characteristics of these areas include the wider availability of increasingly sophisticated technology, the determined nature of highly motivated bad actors, and the importance of industry-wide and multi-stakeholder efforts to address the challenges.

Google Search

Search

Description of Service and Associated Risk Profile

Google's mission to organise the world's information and make it universally accessible and useful starts with Google Search. Search continuously sorts through hundreds of billions of web pages and other content in our Search Index to connect users to the most relevant and helpful search results for their queries. You can read more in our description of How Search Works.

Search plays an essential role in supporting the enjoyment, realisation, and fulfilment of the right to freedom of opinion and expression. Over 375 million users in the EU[7] exercise their right to seek and receive information[8] through Search, and web publishers are better able to express themselves and reach interested audiences through search results.

We remove pages from search results when we have a legal obligation to do so and maintain clearly defined policies that apply to content surfaced anywhere within Search. However, where there is clear user intent to find certain content, returning responsive results that some may find objectionable, offensive, or problematic is not just tolerable, but the right outcome, ensuring users' access to information they seek. When a user wants to know where on the web a particular piece of content can be found, the user should be able to construct a query that seeks it out and Search should return responsive information with links to relevant sources, subject to any legal obligations. Content appearing in response to sufficiently clear queries indicates that Search is working as intended. Failure to deliver this content would harm both the rights of the speaker to freedom of expression and the rights of the user to seek and receive information.

This approach is consistent with our understanding of the DSA, which acknowledges the important distinction between search engines and hosting services[9] and states that VLOPs and VLOSEs should pay particular consideration to the impact on freedom of expression when mitigating content risks and avoid unnecessary restrictions on the use of their service.[10]

To protect the right to freedom of expression, it is therefore essential that any restrictions we implement be reasonable and proportionate. We first outlined our approach to freedom of expression in 2007, and while

[7] Average monthly counts based on distinct signed-in accounts of recipients.
[8] Article 19 of the Universal Declaration of Human Rights states: "Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers"; Article 11 of the EU Charter: "Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers."
[9] For example, Recital 19 recognises "the different nature of the activities" between caching services and hosting services.
[10] Recital 86 states that mitigations should "avoid unnecessary restrictions on the use of their service, taking due account of potential negative effects on those fundamental rights…providers should give particular consideration to the impact on freedom of expression."

we have refreshed and refined our principles, our philosophy on this issue has remained largely consistent since then.

Content policies for Search are designed to minimise restrictions on freedom of expression and promote access to information. This design means that risks associated with potentially illegal or "legal but harmful" content will always be present with Search because content may still be discoverable if it is available on the internet. When returning search results, we take action to avoid surfacing egregious content, such as CSAM, or content that violates our policies relating to highly personal information, including doxxing and non-consensual explicit imagery.

We take protective measures to avoid showing shocking or harmful results when a user is not deliberately looking for such content, and provide tools such as SafeSearch to limit unwanted explicit results. This includes turning SafeSearch on by default for known minors, and applying the SafeSearch explicit-image blurring by default for all new users. These measures help address risks relating to content that may be objectionable, offensive, or problematic, especially to those who are not seeking it out.

Our approach is informed by several important factors.

First, search results should not unexpectedly present content that may be objectionable, offensive, or problematic. We deploy a range of measures such as ranking algorithms, quality testing, and content policies (described below) to ensure that results are relevant, helpful, and trustworthy. We acknowledge the risks of problematic content and provide users with relevant contextual information where appropriate.

Second, we are cognisant of the unique concerns around protecting children who use Search, and have implemented tools like SafeSearch and Family Link, described further below, to address those concerns.

Third, our approach distinguishes between core web results (such as links to external pages) and certain other features of Search (such as Autocomplete, Featured Snippets, and Discover). To keep information accessible, we remove content from the core web results that are relevant to a query only in limited circumstances: this includes blocking CSAM, spam, highly personal information, and content that is subject to valid legal complaints or site owner requests. By contrast, Search features offer additional value—such as providing extra context, helping users formulate queries, or creating a personalised feed—and we understand that users may perceive this content to have higher credibility because of how it is presented. Here we apply content policies that cover a wider variety of issues, including barring harassing, hateful, and violent content. We carefully consider what appears in Search features because our presentation can emphasize and highlight the content in a manner beyond the simpler ordered list we use to display core web results.

Lastly, and critical to understanding the nature of systemic risk on Search, search engines do not have the same relationship with users and user-generated content as website owners and hosting providers, including online services. Search engines cannot remove content from the web. Only website owners and hosting providers can remove content or moderate illegal and harmful content on their sites.

Placing broad restrictions on the types of content that can be accessed through search engines would interfere with the right to freedom of opinion and expression, including the right to seek, receive and impart information, and the ability to access and hear different views. That's why we remove content from search results only in very limited circumstances, including legal removals, violations of our web search

spam policies, or violations of our narrowly-scoped policies that address highly personal information where it is rarely, if ever, in the public interest to display.

Systemic Risk Assessment Results and Associated Observations

In this systemic risk assessment, we considered risks associated with Search and features that appear on Search. We assessed 39 different risk statements[11] for inherent risk (i.e., risk absent any action taken by Google), preparedness (i.e., the cumulative measures currently in place to mitigate the risk), and residual risk (i.e., risk after mitigation by Google). Residual risk serves as a guide for where further investment may be warranted. The full list of risk statements is found in Annex A to this report.

One important theme for Search is the ongoing presence of residual risks relating to the availability of potentially illegal or harmful content in search results for those seeking this content. This arises from our chosen emphasis on maximising access to information and awareness of the risk of over-broad restrictions to freedom of expression and information; we believe this result to be appropriate given the nature and purpose of Search and the DSA's goals relating to freedom of expression and information.

In the next three sections ("Removing Content"; "Investing in Search Information Quality"; "Service Design") we address each of the categories of systemic risk articulated in Article 34(1) of the DSA. We emphasize where the assessment showed elevated inherent or residual risks and describe what Search is doing and plans to do to mitigate them. We also highlight external factors and improvements to our mitigation measures that changed the levels of inherent or residual risk between 2023 and 2024.

As explained above, fundamental rights are closely interconnected and there is a high degree of dependency between different risk statements. The improvement or deprivation of one fundamental right can advance or adversely affect the fulfilment of other fundamental rights, while the controls and measures to address one risk statement may address other risk statements too. With this in mind, we have grouped risks and mitigations together based on how these risks manifest for Search and how they are addressed. This allows for efficient explanation of Search's existing mitigating practices, as well as improvements consistent with Article 35 of the DSA.

2024 Highlights

- Introduced additional mitigations to address non-consensual explicit images (NCEI) and involuntary synthetic pornographic images (ISPI)
- Surfaced high-quality information during the EU Parliamentary Elections
- Reduced residual risk of content promoting practices harmful to health with enhanced mitigations

[11] *See* Methodology Step One: Classification.

Removing Content

Removing Illegal Content

We remove pages from search results when we have a legal obligation to do so. In many cases, content that is manifestly illegal also violates our policies, so we remove it before we receive a legal order to do so. For example, CSAM is illegal regardless of the context in which it appears, so we use automated methods such as hash matching to swiftly detect and remove CSAM from search results, and this significantly lowers the level of residual risk. This is described in more detail below.

Other types of potentially illegal content (such as terrorist and violent extremist content, hate speech, or non-consensual explicit images) either have no standard definition or require contextual understanding, such as whether the subject of the content consents to its availability online; whether the content has educational value, appears as part of a documentary, or represents artistic expression; or whether the content is being distributed for a particular bad purpose (such as to promote terrorist groups). Deciding whether content is illegal is not always a determination that Google is equipped to make, and we balance taking action against content with respect for the right to freedom of expression and information. Based on the mitigation measures described below, we assess risks relating to potentially illegal content to have lower levels of residual risk. However, risks relating to potentially illegal online activity (such as sharing of non-consensual explicit images) are more complex to address and require notification to Google by affected persons to determine their illegality, and we have assessed these to have more elevated levels of residual risk.

We hold ourselves to a high standard when it comes to our legal requirements to remove content from search results. We encourage people and authorities to alert us to content they believe violates the law, and we make every effort to respond appropriately to legal notices.[12]

Addressing Violations of Intellectual Property Rights

Search responds to clear and specific notices of alleged copyright infringement and delists content and URLs that violate applicable copyright law from search results.

However, a search engine cannot automatically confirm whether a given page has a licence to host content, so we depend on reports from copyright owners. To initiate the process to delist content from search results, a copyright owner who believes that a URL points to infringing content sends us a takedown notice for allegedly infringing material. When we receive a takedown notice, our teams and automated systems carefully review it for completeness and validity. If the notice is complete and we find no other issues, we delist the URL from search results.

Because of our established frameworks for understanding and mitigating risks associated with intellectual property violations (including copyright), as well as our internal processes for legal review and removal after claims of infringement, our assessment did not find elevated residual risk with respect to these concerns. You can read more in our Copyright Help Center and the Content Delistings Due to Copyright section of the Google Transparency Report. The latter provides data for the number of URLs requested to be delisted, the

[12] *See supra at* Handling Government Removal Requests.

number of unique individuals or entities that have claimed an exclusive right to content specified in copyright delisting requests, and the reporting organisations, specified domains, and copyright owners who have submitted or been cited in the most requests.

Detecting, Removing, and Reporting CSAM

The presence of CSAM on a page is illegal in most jurisdictions regardless of context and causes clear harm to victims, so we develop ways to automatically identify that content and prevent it from showing in our results.

We invest heavily in fighting CSAM and exploitation online, and use our proprietary technology to deter, detect, remove, and report it on our services. We identify and report CSAM with trained specialist teams and tools, including machine learning classifiers and hash matching technology which creates a "hash", or unique digital fingerprint, for an image or a video so it can be compared with hashes of known CSAM. When we detect content that appears to be CSAM, we report it to the National Center for Missing and Exploited Children (NCMEC), which liaises with law enforcement agencies around the world.

For many years, we have been working on automated systems that allow us to identify never-before-seen CSAM imagery so it can be reviewed and, if confirmed as CSAM, removed from search results and reported as quickly as possible. In addition to consistently applying it to eliminate CSAM from Search, this technology also powers the Content Safety API, which we developed to help partner organisations classify and prioritise potential abuse content for review. The Content Safety API is one part of our child safety toolkit—alongside CSAI Match, YouTube's proprietary technology for combating child sexual abuse imagery (CSAI) videos online. Every month, our partners use the toolkit to process over 4 billion images and videos, helping them identify problematic content faster and with more precision so they can report it to the authorities. When we help our online partners identify more abusive content, the entire internet benefits.

It's our policy to block search results that lead to child sexual abuse imagery or material that appears to sexually victimise, endanger, or otherwise exploit children. We are constantly updating our algorithms to combat these evolving threats.

We apply extra protections to searches that our systems identify as likely seeking CSAM content. We filter out explicit sexual results if the search query seems to be seeking CSAM. For queries seeking adult explicit content, Search won't return imagery that appears to include children to help break the association between children and sexual content. In many countries, users who enter queries clearly related to CSAM are shown a prominent warning that child sexual abuse imagery is illegal, with information on how to report this content to trusted organisations like the Point de Contact in France and FSM in Germany. When these warnings are shown, users are less likely to continue looking for this material. Our evaluations of the effect of the prominent warning show a 20-27% increase in queries with no interactions and a 15% reduction in CSAM-seeking follow-on queries.

During 2023, we reported and removed nearly 850,000 URLs for CSAM from search results by using automated and manual methods. This is in addition to our efforts to deter CSAM-seeking queries as noted above. You can read more about the scale of our efforts to combat online CSAM in our transparency report.

External stakeholders have raised the question of whether generative AI may exacerbate CSAM risks by lowering barriers to the creation of new CSAM, including the editing of otherwise benign images of children. We did not see an increase in CSAM during the assessment period and we believe that our existing methods to detect, remove, and report human-generated CSAM are as effective at protecting against computer-generated CSAM. However, we will continue to monitor this important risk, and you can read more about our approach to guard against computer-generated CSAM, including our commitment to the Safety by Design Generative AI Principles, in Child Safety and Generative AI.

During the assessment period, and consistent with Article 35 of the DSA, we introduced an age indeterminate video classifier to improve our ability to detect possible CSAM when the subject's age is ambiguous, regardless of query intent. We also streamlined our notice and takedown process to help simplify the process for third parties like NCMEC and the Internet Watch Foundation (IWF) to report CSAM.

Removing NCEI and ISPI

Globally, we have policies to remove both non-consensual explicit images (NCEI) and involuntary synthetic pornographic images (ISPI) upon receiving a request that meets certain requirements. Online sharing of this type of material can be extremely distressing to the subjects. In some contexts and jurisdictions, this content is not only offensive and harmful, but also illegal to post or distribute.

For people who wish to remove NCEI and ISPI depicting them from Search, we provide a process to request removal of links to the content from search results pursuant to our policies against this type of content. People can also submit a separate legal removal request if they believe the content violates particular laws, such as copyright laws or local laws prohibiting the non-consensual sharing of explicit images.

Once NCEI has been reported, if it meets removal requirements and Search has removed the content from results, then we also begin to block duplicates and filter explicit results on queries that return results similar to the previously reported NCEI content. Further, if we process a high volume of personal information removals involving a site with exploitative removal practices (i.e., sites that require payment to remove content), we demote other content from the site in our results. We also look to see if the same pattern of behaviour is happening with other sites and, if so, apply demotions to content on those sites. We may apply similar demotion practices for sites that receive a high volume of doxxing content removals. We also maintain automatic protections designed to prevent non-consensual explicit personal images from ranking highly in response to queries involving people's names.

While Search has a robust set of policies and tactics to mitigate the risk of this content appearing in search results, especially for users who are not looking for explicit content, the volume and virality of NCEI represents a higher level of inherent risk. We are well prepared to address this risk, but highly motivated bad actors and the difficulty of proactively detecting NCEI and ISPI means that there is always room to improve our mitigations.

During the assessment period, and consistent with Article 35 of the DSA, we enhanced our removals process making it easier for people to remove NCEI and updated our ranking systems to reduce the prominence of this content in Search. These updates were developed based on feedback from experts and victim-survivors.

As generative imagery technology has continued to improve in recent years, there has been a concerning increase in generated images and videos that portray people in sexually explicit contexts, distributed on the web without their consent. This risk has been raised by external stakeholders, and we've also observed the increase in reports of non-consensual explicit content when monitoring our own systems. We will continue to monitor this risk, and believe that our existing methods to detect, remove, and report human-generated NCEI and ISPI are as effective at protecting against computer-generated content.

Volume of Content Removed

Our EU DSA Biannual VLOSE/VLOP Transparency Report discloses the number of actions we took on Search, segmented by type of illegal content or violation of terms and conditions.

For URL-level removals (i.e., the number of individual URLs removed due to legal or policy violations) the most common reason for removal was intellectual property infringement (over 800 million), followed by protection of minors (over 10 million), data protection and privacy violations (over 100,000), and non-consensual behaviour (over 70,000). During the reporting period, >99.99% of all fully automated removal decisions on Web Search that impacted users based in the EU were unchanged while <0.01% were reinstated as a result of a counter notice.

For URL-level filtering (i.e., the number of times individual URLs were filtered algorithmically from Discover feeds based on content policies), the most common reasons for action were sexualised content (over 30 million) and violence (over 19 million). We use precision metrics to measure our effectiveness, defined as the ratio of true positive instances (i.e., correct automated decisions) as a proportion of both true and false positives, and these range from 73% - 90% for racy, commercial, violent, and profane content and over 99% for spam.

Taken in combination, these numbers provide an illustration of inherent risks for illegal and policy violative content prior to the removal actions taken by Search.

Investing in Search Information Quality

The systemic risk assessment reviewed several risks relating to a wide variety of harmful content such as hate speech, violent extremist content, gender-based violence, content inciting violence, content promoting practices harmful to health, and content that constitutes harassment and bullying. We deploy a wide array of measures to address the risks related to harmful content, including Search ranking (such as surfacing credible and high-quality content over lower-quality content in web results) and content policy enforcement, especially in Search features.

However, Search has indexed hundreds of billions of web pages, images, videos, and other content, so search results might occasionally contain material that some find objectionable, offensive, or problematic. Content that Search has no legal obligation to remove and is not prohibited by our policies remains available in search results for users who express an intent to explore that content, even if indicators suggest it is of relatively low-quality or potentially harmful. While we believe our approach to be reasonable, proportionate, and effective in the context of a search engine service, the continued availability of this content results in elevated levels of residual risk for several risk statements relating to harmful content in

the fundamental rights and public health dimensions of the systemic risk assessment. Below we describe some of the mitigating measures we take.

Our automated systems are our first line of defence to limit the appearance of harmful content in search results for the most common queries, but we may also have trained experts who manually review and remove content that violates our Search features content policies. During the assessment period, and consistent with Article 35 of the DSA, we launched a framework for demoting unexpected harmful content (i.e., content that users were not searching for), including sensitive symbols and other offensive images.

Addressing Sensitive, Harmful, and Policy Violative Content

We use ranking systems to sort through hundreds of billions of web pages and other content in our Search index to present the most relevant and useful results in a fraction of a second. Our ranking systems are central to addressing systemic risks relevant to Search.

Search ranks and prioritises content using signals that align with meaning, relevance, quality, usability, and context. Our approach is to raise the ranking of the highest quality information, rather than removing low-quality information. Our emphasis on the ranking rather than availability of content allows us to address the risk of harm in a proportionate manner and reduce risks to freedom of expression and information.

Our ranking systems are especially designed to surface high-quality content for what we call "Your Money or Your Life" (YMYL) topics, defined as those that may significantly impact the person who is directly viewing or using the content, other people who are affected by the person who viewed the content, or groups of people or society affected by the actions of people who viewed the content. YMYL topics can directly and significantly impact people's health, financial stability or safety, or the welfare or wellbeing of society—for example, pages that offer financial advice or information regarding investment, taxes, retirement plans, loans, banking, insurance, or which facilitate purchases or online money transfers. You can read more about our more notable ranking systems in our guide to Google Search ranking systems.

To help us test and improve our Search algorithms we put all possible changes to Search through a rigorous evaluation process to analyse metrics and decide whether to implement a proposed change. We work with external Search Quality Raters to evaluate the quality of these automated ranking systems based on the expertise, experience, authoritativeness, and trustworthiness of content. This approach to testing our ranking systems is explained more in How insights from people around the world make Google Search better and An overview of our rater guidelines for Search.

During the assessment period we made further progress in our efforts to address the risk of content that promotes practices harmful to health by improving our ability to accurately detect a range of personal crisis searches, such as suicide, substance abuse, self harm, and eating disorders, and to recognise the intent behind the search. This has meaningfully improved our ability to connect users with contact information for national hotlines and more reliably show trustworthy and actionable information during times of personal crisis.

Informing Users

Our About this result and About this page tools enable users to learn more about the source of a search result—such as a description of the source and how others describe them—so that users can make more informed decisions about the sites they visit and the results that are most useful to them. Our About this page tool provides additional information about the source and topic of a web page, including perspectives on the same topic from high-quality sources.

We also recently introduced an About this image tool that can be used from Search, Circle to Search, or Google Lens, to provide context about images users encounter online. The About this image tool helps users get more information about how other sites (such as news and fact-checking sites) use and describe the image, an image's metadata (such as the creator of the image and some information about how that image may be created) and whether the image was generated using AI, provided it contains Google DeepMind's SynthID watermark embedded within its pixels.

The growth of disinformation and misinformation and the emergence of new technologies require our ranking methods to continually adapt to evolving risks. In addition to tools empowering users to discern the trustworthiness of information for themselves, Search continues to invest in methods to prioritise the ranking of the most relevant and reliable information available and ensure effectiveness of our algorithms across languages, countries, cultures, and contexts.

Providing SafeSearch

Keeping people safe on Search also means helping them steer clear of unexpected, shocking results. One way we tackle this is with SafeSearch settings, which help detect and manage access to explicit content like pornography and graphic violence in Search. We also offer options for parents and schools to lock this on Search for supervised minors.

To help protect people from inadvertently encountering explicit imagery on Search we provide a SafeSearch setting that blurs explicit images in search results. This setting is enabled by default for users that do not have SafeSearch Filter turned on. Users have the option to enable SafeSearch "Filter" (blocking any explicit content) or turn the SafeSearch protections "Off" at any time.

For Google Accounts for people under 18, we take additional steps to help minors make choices to avoid results that may be shocking or harmful. SafeSearch Filter is enabled for declared minors under the age of 13 (or the applicable age of consent in the relevant country) when signed in to an account managed with Family Link. Additionally, SafeSearch is turned on automatically when our systems indicate that a user may be under 18.

When SafeSearch is "Off," users find all relevant results for their search, even if they are explicit, but our SafeSearch signals still apply to suppress irrelevant explicit content when the user does not appear to be seeking it out. Our safety algorithms improve hundreds of millions of searches globally on a daily basis across web, image, and video modes.

Supervised users are unable to change their SafeSearch setting—for example, for child and student accounts, parents and schools can lock SafeSearch, while parental controls on an operating system or

antivirus software may override an individual's SafeSearch setting. Parents can use Family Link to set up supervision on a child's account, with SafeSearch Filter turned on automatically and locked so that the child cannot change the setting. You can read more in Manage Search on your child's Google Account.

Tailoring our Content Policies

In Search, we take a multi-tiered approach to content policies to balance the need to protect freedom of expression with providing users with high-quality information.

Search policies apply to content surfaced anywhere within Search, which includes web results (i.e., web pages, images, videos, news content or other material that Google finds from across the web). Search's policies cover essential content restrictions such as CSAM, spam, and valid legal requests. We maintain the following three categories of content policies.

Search policies include a highly personal information policy under which we remove certain personal information that creates significant risks of identity theft, financial fraud, or other specific harms such as doxxing content, explicit personal images, and involuntary fake pornography. These policies were developed following an extensive stakeholder consultation to help inform how we balance taking action to protect user privacy and safety with the right to freedom of expression, and were enhanced in 2022 to include the removal of additional personal information (such as contact information, confidential login credentials, and confidential biometric data) from Search in cases that do not involve doxxing.

Search features policies apply to many of our search features, such as Autocomplete, Featured Snippets, and Google Discover. The presentation of these features emphasizes and highlights the content differently than our relevance-based web results.

Search feature-specific policies explain how certain search features work, and set forth any additional feature-specific restrictions. Examples include prohibiting predictions about medically hazardous health claims on Autocomplete and applying a higher quality threshold for recommending content for YMYL topics on Discover.

Addressing Civics Mis- and Disinformation

Search aims to enable users' informed participation in democracy by providing high-quality information that is accurate, up-to-date, and protected during elections. One way we ensure reliable information is returned to users in the elections context is through our use of classifiers to identify elections-related queries, so that our systems know that returning high-quality information is especially important. As part of this effort, Search has developed a number of features aimed at ensuring we show users trustworthy elections-related content from reliable third parties. These features are activated during elections and in response to elections-related queries to mitigate the risk of low-quality content and ensure we return organised search results pages that include comprehensive authoritative information in all EU Member States during their national elections.

During 2024 we collaborated with the European Parliament to provide high-quality information about the European Union Parliamentary elections. For example, people searching for topics like "how to vote" found information related to ID requirements, registration, voting deadlines, voting abroad, and guidance for

different means of voting such as in person or via mail. This was available in 22 languages and included a country selector for the 12 million people who are eligible to vote in a different country than the one they live in. More details about Google's overall approach to the 2024 European Union Parliamentary elections, including Search, is available in Supporting Elections Integrity.

In addition to developing Search features with trustworthy third parties, we also utilise Search Quality Raters, as described above, and elections-specific classifiers to ensure search results surface high-quality information about key persons or entities.

Search information quality processes, including ranking and our robust policies that allow us to remove violative content, consistently perform well in preventing mis- and disinformation from surfacing to users. For example, in 2021, the Leibniz Institute for the Social Sciences conducted a comparative algorithm audit of presence of conspiracy-related information in top search results across five search engines: Google, DuckDuckGo, Yahoo, Bing, and Yandex. Their research found that "all search engines except Google consistently displayed conspiracy-promoting results and returned links to conspiracy-dedicated websites in their top results, although the share of such content varied across queries."[13]

Respecting Freedom of Opinion, Expression, Media Pluralism, and Civic Discourse

We reviewed several risks relating to content removal, users making informed decisions about what to view, and media pluralism (e.g., the plurality, polarisation, and diversity of perspectives available). Consistent with our findings in 2023, Search's approach to making information universally accessible reduces the level of inherent risk and results in low residual risk overall. These results are consistent with Search methods and initiatives like the Search Quality Evaluator Guidelines and Google News Initiative. Search endeavours to play an essential enabling role for the realisation, enjoyment, and fulfilment of these rights.

We believe, and studies have shown, that Search returns relevant and helpful sources with "no evidence of ideological bias" when users are looking for news.[14] Search is not designed to favour or disfavour any particular publications based on ideology. Instead, our systems look at signals such as relevance, prominence, freshness, authoritativeness, or trustworthiness to determine the most helpful, relevant content to show users. Search does not take the political viewpoint of a web page into account when ranking. In the Google Search Quality Evaluator Guidelines, Search instructs evaluators that "[r]atings should not be based on your personal opinions, preferences, religious beliefs, or political views."

Reputable studies consistently find that Search is fair. For example, two studies by The Economist evaluated claims of bias in Google News results and found no evidence of ideological bias, concluding that "Google rewards reputable reporting, not left wing politics".[15] An extensive study by academics at Stanford University drew a similar conclusion. Over a six-month period, researchers reviewed search results appearing on the first page for every candidate running for federal office in the 2018 U.S. general election. Four million URLs were scraped from Search and audited, and the researchers found that search results did not exclude sources from either the left or the right of the political spectrum.[16]

[13] Urmana A, Makhortykhb M, Ulloac R, Kulshresthad J (2021) Where the Earth is flat and 9/11 is an inside job: A comparative algorithm audit of conspiratorial information in web search results: Leibniz Institute for the Social Sciences.
[14] *The Economist* (2019) Google rewards reputable reporting, not left-wing politics.
[15] *Id.*
[16] Danae Metaxa, Joon Sung Park, James Landay, Jeff Hancock (2019) Search Media and Elections: A Longitudinal Investigation of Political Search Results, Proceedings of the ACM on Human-Computer Interaction, Volume 3 Issue CSCW. Article No. : 129.

Search is also best-in-class in displaying diverse results, which are indicative of strong support for media pluralism. In 2022, the Hamburg University of Applied Sciences published a paper reporting the findings of a study examining the difference between results retrieved by four major web search engines. Researchers compared the top 10 results from Google, Bing, DuckDuckGo, and Metager, using 3,537 queries generated from Germany and the US. The findings of the study showed that "Google displays more unique domains in the top results than its competitors, and Wikipedia and news websites are the most popular sources overall."[17]

Addressing Mis- and Disinformation

We believe that elevating authoritative information and combating mis- and disinformation are of utmost importance to systemic risks relevant for Search. These efforts are especially relevant to issues such as public health, elections, and civic engagement. While the efforts described above all go to combat mis- and disinformation appearing on Search and across our services, there are other efforts that are critical to our holistic approach.

We implement a multi-faceted approach to address the complex challenges and risks raised by mis- and disinformation across our services. While our ranking systems seek to connect people with authoritative sources and are described elsewhere in this report, we are cognisant that these are complex issues that no single actor is able to solve on their own.

For this reason, we have welcomed the multi-stakeholder approach, including the EU's 2018 Code of Practice on Disinformation and a Strengthened Code that we signed in June 2022. As part of the Strengthened Code, we have committed to providing the European Commission with reports detailing how we have implemented our Commitments under the Code.

Our baseline report under the Code highlighted the breadth of our work across EU Member States to detect and counter a range of threats to the integrity of our services, empower users, and work with a variety of stakeholders. The report also provided information about the quantitative impacts of our work at the Member State level.

Following this baseline report, we expect to publish subsequent versions of this report biannually. In addition, we expect to remain a committed and productive member of the Code of Practice's Permanent Task-force.

Service Design

Addressing Unfair Commercial Practices and Fraudulent Content about a Business

We take many actions to mitigate the risks of unfair commercial practices and fraudulent content about businesses, such as prioritising the highest-quality results as part of the ranking process described above, removing policy-violating content from Search features, and removing fraudulent content subject to legal removal requests. However, this type of content will continue to be returned in search results when a user

[17] Nurce Yagci, Sebastian Sünkler, Helana Häußler, Dirk Lewandowski (2022) A Comparative of Source Distribution and Result Overlap in Web Search Engines: Hamburg University of Applied Sciences.

seeks it out, provided it is not the subject of a valid legal removal request or prohibited by Google policies (e.g., spam), and is still available on the internet. For this reason, the assessment found that some elevated residual risk of fraudulent business information appearing on Search remains.

We employ a higher standard and a different approach to address unfair commercial practices and fraudulent content about a business in the advertising context. Our Ads Policies, which apply to ads on Search and our VLOPs, have several policies relevant to mitigating this risk, such as policies prohibiting misrepresentation (e.g., phishing, obscuring charges associated with financial services, misleading claims regarding weight loss or financial gain) and policies prohibiting the sale or promotion of counterfeit goods, dangerous products and services, and products or services enabling dishonest behaviour (e.g., hacking software, fake documents, or academic cheating).

Google Ads does not allow ads that deceive users by excluding relevant product information, such as billing details or charges, interest rates, fees, and penalties, or by providing misleading information about products, services, or businesses. This includes impersonating brands or businesses, concealing or misrepresenting a business identity, and implying endorsement by another individual, organisation, product, or service without their knowledge or consent. For egregious violations (those so serious that they are unlawful or pose significant harm to our users), we will suspend Google Ads accounts upon detection and without prior warning, and not allow the advertiser to advertise with us again, unless an appeal brings to light compelling grounds for a different outcome.

During the assessment period we updated our spam policies to better address new and evolving abusive practices that lead to unoriginal and low-quality content showing up on Search. We have long had a policy against using automation to generate low-quality or unoriginal content at scale with the goal of manipulating search rankings, but we have strengthened our policy to allow us to take action on more types of content with little to no value created at scale, like pages that pretend to have answers to popular searches but fail to deliver helpful content. We also now consider (1) very low-value third-party content produced primarily for ranking purposes and (2) expired domains that are purchased and repurposed with the primary intention of boosting search ranking of low-quality or unoriginal content to be spam.

Respecting Privacy

Privacy is an enabling right, furthering rights such as freedom of expression, association, opinion, religion, movement, and bodily security.[18] Once violated, the right to privacy can be challenging to remediate because private or highly personal content can remain in circulation on the internet. Given the role of Search in surfacing information from nearly anywhere on the open web, privacy risks in the context of Search can be important inherent risks.

Search has addressed these inherent risks by (1) ensuring responsible stewardship of user data by refraining from selling user data, constantly refining data collection practices, and providing users with easy-to-use data settings; (2) providing avenues for highly personal information to be removed from Google (described above) and respecting the "right to be forgotten"; and (3) complying with requirements under applicable data protection and privacy laws, including minimising the data being collected, purpose limitation, and providing transparency to users.

[18] UN Special Rapporteur on the right to privacy.

To be responsible stewards of user data, we take a private by design approach: we encrypt every search, build controls so that users can choose the privacy settings that are right for them, and never sell personal information to anyone. Search also offers privacy controls so that users can decide what to save to their Google Account and can turn on auto-delete to automatically delete data on an ongoing basis.

An example of this private by design approach took place during the assessment period when we launched Circle to Search, a new way to search anything on your phone with a simple gesture and without switching apps. We put trust and safety at the centre of Circle to Search engaging privacy experts—including academics, civil society organisations, researchers, and privacy and safety professionals—in a series of structured workshops during the product development process. Some of the priorities we implemented following this process included data minimisation (i.e., only using the selected area of the screen to begin a search), an easy process to delete history, and not saving images from Circle to Search to a user's search history. We do not use biometrics matching to inform Circle to search results—for example, a Circle to Search for a public figure would return the same picture in different news articles, but a Circle to Search for a picture of a friend would not surface other pictures of that friend from the web.

Since 2014 we have been responding to requests to delist content under European privacy law, which provides individuals with the right to ask search engines like Search to delist certain results for queries on the basis of a person's name if the links in question are "inadequate, irrelevant or no longer relevant, or excessive." We evaluate each request on a case-by-case basis, and may not delist content where there is an overriding public interest in the information remaining available in search results. Our requests to delist content under European privacy law report provides information and data about the volume of requests, the URLs delisted, the individuals submitting requests, and the content of websites and URLs identified in requests. Since 2014, we have received around 1.5 million requests to remove around 5.6 million URLs. We take our responsibility to ensure compliance with European privacy law seriously while being committed to providing access to information, and carefully balance these commitments when assessing each request. As a result, we have refused to delist around 50% of the requests we have received to date; a large majority of those refusals are sustained when challenged in court or before data protection agencies.

These measures are typically sufficient to reduce many privacy risks to much lower levels of residual risk. However, our assessment concluded that some elevated residual risks remain for Search, most notably the unintentional or malicious sharing of private or highly personal information in search results. This information can be challenging to verify and requires that we be informed of and verify a privacy violation before removing content from search results.

Protecting Children's Rights

Our services provide vital opportunities for learning, communication, and social interaction, and can be formative for a child's cognitive and social development. However, these opportunities are accompanied by risks to which children are particularly vulnerable given their unique stages of development, nascent digital literacy, and evolving cognitive abilities and decision-making skills. It is important for us to address these risks with mitigations, such as user guidance and parental controls, that help children navigate their online experiences now and over the course of their lifetimes.

The systemic risk assessment reviewed several risks relating to children's rights, such as behavioural addictions in children, use of children's data for ads targeting, and unnecessary or disproportionate

limitations on children's access to Search. We found the highest inherent risk to be the risk that children under a defined minimum age access services that they should not be able to, and may be exposed to harmful, hateful, or age-inappropriate content or conduct. Based on the mitigation measures described below, we concluded that Search (and our VLOPs) are taking actions that significantly reduce residual risks for children's rights.

The following protections apply horizontally across all our services, and thus pertain to Search and our four VLOPs. These protections will be described here and cross-referenced in the VLOP sections. Our efforts in this space must balance adults' rights to access services and information with a reasonable level of privacy, and the need to protect children from accessing services and information that are not appropriate for their age.

Obtaining Age Assurance

We require users to manually enter their date of birth (without pre-populated options, referred to as a neutral age-screen) during Google Account sign up to help determine which users are likely under the age of 18 so that we can apply heightened privacy, content, and safety protections. To reduce the burden on our users and in accordance with data minimisation principles, these processes are carried out at the Google Account level so that the results can then be used in connection with all signed-in services (including Search) that are accessed by the user.

Depending on what birth date a user provides at the time of Google Account creation, we apply different protections.

- If a user provides an age that is under the minimum age to have a Google Account in their country, we require approval from a parent/guardian before continuing with account creation, and the account must be supervised until the user attains the minimum age (see further information on Family Link below).

- If a user provides an age under 18, we apply a number of default protections to the account, and we disable access to age-restricted features and settings across some of our services. Parents/guardians who manage their child's account through Family Link may choose to change some of these default settings if a different approach works best for their family.

We independently assess whether or not a user is likely an adult, both for users who sign into their accounts and those who do not. We use a variety of signals, such as the types of sites a user has searched for or the categories of videos a user has watched on YouTube, as well as indicators like the longevity of an account. For example, searches for mortgage lending sites or tax assistance might be signals that the user is likely an adult. Once our model has sufficient signals about a user's age, it sends a signal to our services to automatically set appropriate default settings and protections, such as by turning SafeSearch filtering on for those under 18. This approach does not involve collecting additional information from users.

Enabling Parental Control

Family Link parental controls are available in the Family Link app and also via web browsers. Parents/guardians of minors under the applicable minimum age can create Google Accounts for their

children and must manage those accounts using Family Link parental controls. Family Link parental controls are also available for parents/guardians to supervise minors over the applicable minimum age, but consent from these minors is required before supervision may be enabled.

Family Link helps parents/guardians stay informed about and manage their child's experience on compatible Android and ChromeOS devices. For example, Family Link empowers parents/guardians to set digital ground rules for their family by managing the apps their child can use, keeping an eye on screen time, or setting a bedtime and daily limits for their child's device. On iOS devices, parents can give consent for their child's Google Account sign-in and manage some YouTube and Search settings. These controls help parents/guardians manage their child's experience in ways that make sense for their family.

Providing Ads Protections

We prohibit age-sensitive ad categories from serving to users under 18, including ads that feature adult or sexually suggestive content, alcohol, or gambling and games. We also prohibit the display of personalised ads based on age, gender, or interests to any users we determine to be under the age of 18. Ads shown to these users must meet our under-18 ads policies, and may only be served based on non-personalised contextual information, such as the content on the current site a user is visiting.

Enabling SafeSearch by Default

Using age-appropriate default settings is one way that we incorporate "safety by design" into our products. Specific to Search, SafeSearch filtering is enabled by default for Google accounts for children younger than 18, and parents and schools have the option to lock it on for supervised minors. As described above, SafeSearch filtering blocks explicit content (like sexual activity and graphic violence) from search results across images, videos, and websites—when the filter setting is on, explicit results will be filtered even when they might be relevant for the query.

Built-in safeguards

These age-appropriate settings are also used in built-in safeguards applied to all users to address content which may be harmful to children or other vulnerable users. The controls are designed to help make Search a place where the whole family can confidently search for new information. For example, in many markets, a search for information on suicide, sexual assault, substance abuse, and domestic violence will return contact information for national hotlines alongside the most relevant and helpful results.

Google Maps

Maps

Description of Service and Associated Risk Profile

Google Maps is a service that helps users navigate and explore the world. The service also includes accurate and reliable information about places, business, and experiences, and helps businesses build an online presence, engage with customers, and grow their business. Some of the key elements that make the service compelling include satellite/aerial views, digital street maps, information about places and business, 360° interactive panoramic views of streets, real-time traffic conditions, and route planning for driving, walking, cycling, public transportation, and flying. The information about places and businesses includes some user-generated content, such as content from consumer users and merchant users (including ratings, reviews, and photography) and content provided by the merchant users interested in listing or advertising their business on Maps.

Maps is free, available in over 100 languages, and used by over 285 million users in the EU every month.[19] You can read more in Google Maps Help and our Maps 101 blog series.

The primary purpose of Maps is to help users navigate from place to place and explore the world, with elements such as images, reviews, and information about places being in service of that goal. This systemic risk assessment validated that the most important risks are not intrinsic to the primary purpose of Maps—helping users to get from A to B—but associated with the various features designed to enhance the user experience when fulfilling this purpose.

The service emphasizes being a source of reliable information and a reflection of genuine user experiences. For this reason we lean towards user-generated content policies that are designed to maximise the quality, accuracy, and authenticity of information for consumer and merchant user contributions. We go to great lengths to make sure content published by our consumer and merchant users is helpful and reflects the real world, recognising that this means accepting some attendant limitations to freedom of expression. Content has a low likelihood of going viral because of the way Maps is designed, thereby reducing inherent risks associated with illegal and policy violative content.

Risks relating to conducting a business (e.g., unfair commercial practices, such as paying, incentivising, or encouraging the posting of positive or negative reviews that do not represent a genuine experience) are important to address given the role of Maps in connecting potential customers with businesses (e.g., helping users find a restaurant or auto repair shop that meets their needs).

[19] Average monthly counts based on distinct signed-in accounts of recipients.

Finally, the locational nature of the Maps service, combined with the existence of user-generated content, makes it important to address privacy-related risks (e.g., data practices and risks to personal information, such as unintentional or malicious sharing of private or personal information), as we describe in the “Service Design” section below. The privacy of our users is of utmost importance to us, and while we welcome content that illustrates our world, it's critical to do so in a manner that respects users’ right to privacy.

Systemic Risk Assessment Results and Associated Observations

We assessed 39 different risk statements[20] for inherent risk (i.e., risk absent any action taken by Google), preparedness (i.e., the cumulative measures currently in place to mitigate the risk) and residual risk (i.e., risk after mitigation by Google). Residual risk serves as a guide for where further investment may be warranted. The full list of risk statements is found in Annex A to this report.

This systemic risk assessment surfaced important themes relating to the inherent and residual risk. Overall, we found that the highest assessments of Maps' existing mitigations were correlated with the highest inherent risks, confirming that we have been prioritising action on the most significant risks.

The two most significant inherent risk themes for Maps are directly related to the nature of Maps: risks associated with information about businesses shown on Maps (e.g., fake reviews, unfair commercial practices, disinformation) and risks associated with the locational nature of Maps (e.g., privacy-related risks). However, we have long recognised these risks and our efforts to address them result in much lower residual risks.

We also found higher inherent risks relating to illegal and harmful content, but concluded that these have much lower residual risk given emphasis on the quality, accuracy, and authenticity of information, our substantial efforts to enforce content policy, and the low likelihood of content going viral on Maps.

Article 34(1) of the DSA encompasses a range of systemic risks that are interconnected and cannot be dealt with in isolation; our policies and practices for Maps often address multiple risks at the same time. To provide a comprehensive understanding of Maps' existing mitigating practices and align with Article 35 of the DSA, we have categorised specific manifestations of systemic risks into groups for efficient explanation of existing mitigations and discussion of planned improvements.

In the next two sections we consider content on Maps, including the development and enforcement of content policies ("Content Moderation"), and explore service design choices that target risks associated with Maps' functionality, including privacy ("Service Design").

Taken in combination, these two sections address each of the four systemic risk categories outlined in Article 34(1) of the DSA. We emphasize areas where the assessment has identified elevated inherent or residual risks, elucidating the measures already implemented by Maps to tackle these risks, as well as any future plans to address systemic risks, as appropriate. We highlight where changes in the external context or improvements to our mitigation measures cause significant modifications to inherent or residual risk between the 2023 and 2024 systemic risk assessments.

[20] *See* Methodology Step One: Classification.

2024 Highlights

- New appeals channels for potentially erroneous user-generated content removals and feature restrictions
- Updated our personal information policy that directs users to request content removal if they believe personal information has been posted without their consent
- Additional accessibility features

Content Moderation

In this section we show how Maps has designed and enforced its user-generated content policies to address the systemic risks articulated in Article 34(1) of the DSA. We detail some risks in the assessment with elevated inherent and residual risk, and describe what Maps is doing and plans to do about those systemic risks.

Maps is a local discovery and navigation service with a focus on providing topical and authoritative information and genuine experiences, rather than being a forum for dialogue. For this reason, we maintain content policies to guard places and businesses from off-topic content and fake engagement, especially when there's potential for this type of content to lead to harmful and targeted abuse.

Removing Illegal Content

Maps has clear policies in place prohibiting illegal content on the service through its policies related to restricted, dangerous, and illegal user-contributed content. This includes images or any other content that infringes on anyone else's legal rights, including copyright, as well as content that relates to terrorism, sexual abuse imagery or sexualisation of children, dangerous or illegal acts (such as rape, organ sale, or human trafficking), or illegal products and services (such as endangered animal products and illegal or diverted drugs). We also disallow potentially illegal online activity such as doxxing or content that contains a specific threat of harm or depicts illegal activity.

Despite elevated inherent risks related to illegal content, such as CSAM, terrorist content, and illegal activity, the results of the systemic risk assessment showed Maps' preparedness—such as taking action to stop fake review scams and tackling fake contributions—enabled it to achieve low levels of residual risk for illegal content and activity. As explained earlier in this report,[21] we have a well-developed process for responding to legal orders to remove content, and the efforts described below to enforce our policies ensure the removal of content that is illegal or that violates our policies.

[21] *See supra* at Handling Government Removal Requests.

Addressing Content that Violates our Policies

The systemic risk assessment reviewed several risks relating to a wide variety of potentially harmful content, such as hate speech, violence and gore, harassment and bullying, gender based violence, and content promoting practices harmful to health. However, the enforcement of Maps' user-generated content policies, which favour authoritative information and genuine experiences, lowers residual risks.

Developing Content Policy

We have created strict policies to make sure that user-generated content is based on real-world experiences and to keep irrelevant and offensive comments off Maps. Our policies against topics like fake engagement, misrepresentation, and misinformation continually evolve in response to changing threats.

Our user-generated content policy describes our overall approach while our prohibited and restricted content policies clearly set out what is not allowed on Maps, covering civic discourse, deceptive content, mature content, regulated goods and services, dangerous and illegal content, and low-quality information. These user-generated content policies are more restrictive than those for many other Google services, reflecting our increased emphasis on relevant, authoritative information and genuine experiences for Maps. Our "off-topic" and "fake engagement" policies are good examples of Maps' unique approach to content. These policies have evolved over time to guard places and businesses from violative and off-topic content when there's potential for this type of content to lead to harmful and targeted abuse. For example, when governments and businesses started requiring proof of COVID-19 vaccination before entering certain places, we put extra protections in place to remove Google reviews that criticised a business for its health and safety policies or for complying with a vaccine mandate.

Other policies relevant for content on Maps include our Local Guides Program Terms and Conditions, which set out who is qualified to be a Local Guide and appropriate conduct standards; Google-Contributed Street View Imagery Policy, explaining how we treat inappropriate content and the criteria we use for publishing Street View imagery to Maps; My Maps Content Policy, which sets out policies for creating and sharing custom maps; and Guidelines for representing your business on Google, setting out guidelines for Business Profiles.

Once a policy is written, it's turned into training material—both for our operators and classifiers—to help our teams catch policy-violating content and behaviour.

Enforcing Content Policy

Contributions to Maps should accurately represent the location in question. Where user-generated contributions distort truth we remove content, including reviews, photos, or videos not related to the location or business where they are tagged. If user-generated content is inaccurately placed on the map, or is associated with an incorrect listing, the contribution may be rejected. When a user submits a review, we automatically send it to our system to make sure the review doesn't violate any of our user-generated content policies before posting the review. Given the volume of reviews we regularly receive, we have found that we need both the nuanced understanding that humans offer and the scale that automated detection provides to help us moderate contributed content.

Undertaking Automated Detection and Removal

Automated detection is our first line of defence because automated systems are good at identifying patterns that help determine if the content is legitimate. This includes whether the review contains offensive or off-topic content, whether the Google account has a history of suspicious behaviour, such as a history of posting violative content, and whether there has been uncharacteristic activity, such as many reviews over a short period of time. The vast majority of fake and fraudulent reviews are removed before anyone sees them because all reviews are run against classifiers before being posted.

Our human operators regularly run quality tests and complete additional training to remove bias from the machine learning models. By thoroughly training our models on all the ways certain words or phrases are used, we improve our ability to catch policy-violating content and reduce the chance of inadvertently blocking legitimate reviews from going live. We review and update our classifiers, including review for quality and accuracy across language, gender, ethnicity, and religion, and our assessment identified this as a priority for continuous improvement over time.[22]

If our systems detect no policy violations, then the review can be posted within a matter of seconds. However, our automated systems continue to analyse the contributed content and watch for questionable patterns, such as a group of people leaving reviews on the same cluster of Business Profiles or a place receiving an unusually high number of 1- or 5-star reviews over a short period of time.

In addition, we make it easy for people using Maps to flag any policy-violating reviews, with businesses and consumers both able to report reviews and flag inappropriate user profiles.

Our EU DSA Biannual VLOSE/VLOP Transparency Report includes a Maps accuracy metric for each of the 26 official European Economic Area (EEA) Member State languages, and an overall accuracy metric across all automated content moderation decisions that is language-agnostic.

Undertaking Human Review

A team of human operators work alongside automated systems to remove reviews that violate our policies, and when appropriate suspend user accounts. We deploy thousands of trained operators and analysts globally who help with content evaluations that might be difficult for automated systems, such as understanding reviews that include local slang.

Our EU DSA Biannual VLOSE/VLOP Transparency Report discloses the human resources capacity evaluating Maps content across the official EU Member State languages.

Undertaking Enforcement Proactively

In 2022 we launched a significant update to our machine learning models that helped us identify novel abuse trends many times faster than in previous years. For example, our automated systems detected a sudden uptick in Business Profiles with websites that ended in ".design" or ".top", which our team of

[22] *See supra* at Evaluating Content Across Languages for further discussion of how Maps, and Google as a whole, are addressing this identified residual risk for Maps.

analysts quickly confirmed to be fake. They were therefore able to quickly remove the Business Profiles and disable the associated accounts.

Our EU DSA Biannual VLOSE/VLOP Transparency Report discloses the number of actions we took on Maps, including advertisements presented on Maps, segmented by type of illegal content or violation of terms and conditions. The most recent report (April 2024) demonstrates that scams and fraud and content that is inappropriate or unhelpful (e.g., doesn't reflect the real world) are the most common policy violations on Maps, accounting for over 95% of cases. Other violations (e.g., hate speech, data protection and privacy violations, unsafe or illegal products) are much less prevalent. Over 99% of these actions were taken following automated detection. The report also discloses the number of complaints received from users located in EU Member States about content moderation decisions and the outcome of these complaints.

Going forward, Maps will continue to invest in new technologies and processes to keep information on our services helpful and reliable.

Posting Restrictions

When we find that user contributions for certain types of places are consistently unhelpful, harmful, or off-topic, we may limit or suspend user-generated content for those places. Maps has developed a measured response regarding posting restrictions:

- Short-term restrictions, when posting may be turned off for a particular place for a short period of time to help protect the place or area from a spike in irrelevant or offensive content.
- Long-term restrictions, when posting on a particular place may be turned off for a longer period of time if its category or geographic area has experienced a continuous pattern of low value or off-topic posts.
- Partial or full restrictions, when, depending on the volume and pattern of policy violating content, a particular place may have posting restrictions on some or all of the types of user-generated content (including text reviews, ratings, photos and videos).

Posting Restrictions for Repeat Violators

We found that policies such as our posting restrictions greatly reduce the opportunity for repeat offenders to manipulate our systems through inauthentic use, reducing residual risks across the board.

Assessment Results for Specific Content Risks

Protecting Civic Discourse

We assessed the risk that mis- and disinformation relating to elections, civic discourse, democratic participation, or civil unrest may be available on Maps. While this risk may occur in the context of user-generated content, Maps is designed for a low likelihood of content going viral, and more severe outcomes (such as influencing the result of an election) are highly unlikely. Our preparedness for elections and prohibition of any information that may be deceptive or misleading about civic processes, newsworthy events, or civic discourse significantly reduce residual risk.

In addition to reviewing flagged content, our team proactively works to identify potential abuse risks, which reduces the likelihood of successful abuse attacks. For instance, when there is an upcoming event with a significant following—such as an election—we implement elevated protections for the places associated with the event and other nearby businesses that people might look for on Maps. We continue to monitor these places and businesses until the risk of abuse has subsided. To avoid the spread of election-related mis- and disinformation, we prevent people from editing the phone numbers, addresses and other information for places like voting sites.

This approach was especially significant during the assessment period given the large number of elections taking place globally and in the EU. You can read more about Google's overall approach in Delivering Reliable Information.

Protecting Consumers and the Freedom to Conduct a Business

Maps provides information to users that enable them to find and navigate to a business. For this reason, prominent inherent risks include the risk that disinformation, misinformation, or fraudulent content about a business is available or that unfair commercial practices take place on Maps. This is typically driven by intentional manipulation of the Maps service and might include positive or negative fake reviews, "review bombing" by competitors, fraudsters creating false business listings, or "predatory removals," which occur when a bad actor demands payment for the removal of fake reviews. These risks can disproportionately impact less technologically literate users and newly opened businesses, which are typically more vulnerable than established brands.

We work to stay ahead of scammers and protect small businesses by continuously monitoring for fraudulent content on our products, using a combination of people and technology. One of the best tools we have to fight back is our understanding of inauthentic use patterns on Maps, which informs our classifiers. These classifiers detect and remove policy-violating content across a variety of languages, and also scan for signals of abnormal user activity.

Our teams and protections are built to fight two main types of bad actors: content fraudsters and content vandals.

Fraudsters, who are ultimately motivated by money, try to trick people with scams like fake reviews to attract customers or fake listings to generate business leads. To deter them, we preemptively remove opportunities for them to profit from fake content, and have focused efforts on detecting content coming from click farms where fake reviews and ratings are being generated. Through better detection of click farm activity, we are making it harder to post fake content cheaply, which ultimately makes it harder for a click farm to sell reviews and make money.

Content vandals, who may be motivated by social and political events or simply want to leave their mark online, often post fake reviews or edit the names of places to send a message, or add off-topic photos as pranks. Content vandalism can be more difficult to tackle than fraud as it is often random. Impeding content vandals requires anticipation and quick reaction, and as certain places become more prone to vandalism, we adjust our defences—such as when we modified our algorithms to preemptively block racist reviews when we observed anti-Chinese xenophobia associated with COVID-19.

These risks are further mitigated by the implementation of our Ads policies, such as the Misrepresentation Ads Policy, which disallows ads that deceive people, and the Restricted Businesses Policy, which restricts certain kinds of businesses with products prone to abuse. These Ads policies are complemented by relevant Maps User Contributed Content policies, such as the Misrepresentation Policy, which doesn't allow users to mislead or deceive others, and the Impersonation Policy, which doesn't allow users to impersonate any person, group, or organisation.

We have long recognised these inherent risks as priorities, and our wide range of measures to remove policy-violating reviews, stop fake Business Profiles, and protect targeted businesses serve to reduce these to much lower residual risks. Our actions here also address equality and non-discrimination risks, since this activity can disproportionately be targeted at those under-represented as content contributors, such as minority businesses.

Respecting Freedom of Opinion, Expression, and Media Pluralism

The systemic risk assessment explored several risks relating to content removal, users reporting potentially violating content, and media pluralism (e.g., the plurality, polarisation, and diversity of perspectives available). Maps' focus on providing topical and authoritative information and genuine experiences rather than being a forum for dialogue lowered freedom of expression and media pluralism as an inherent risk. Maps does remove high volumes of content that violates our policies against off-topic or misleading information about locations and businesses, and for this reason the systemic risk assessment found that some residual risk remains for over-moderation of user-generated content.

However, during the assessment period, and consistent with Articles 20 and 35 of the DSA, we established and implemented new appeals channels for potentially erroneous user-generated content removals and feature restrictions, resulting in an improved preparedness evaluation and lower residual risk for risks to freedom of expression than the prior assessment. Maps' merchants have long been able to appeal potentially erroneous account and listing suspensions, but the new appeals channels offer users a path to redress and also gives Maps better information for improving in the first instance.

Overall, we believe that this approach is reasonable, appropriate, and proportionate given the nature, purpose, and intended use of Maps.

Service Design

Respecting Privacy

Some prominent inherent risks for Maps relate to privacy, reflecting the locational nature of the Maps service, the existence of user-generated content, and challenges associated with reversing privacy impacts once they have occurred. Our privacy risks and mitigations cover three dimensions: users of Maps; contributors to Maps; and images shown on Maps that may involve users, non-users, and contributors.

Protecting Users of Maps

Maps uses location data to make its service functional and useful for users. Real time location information plays a very important role for Maps, such as assisting in providing accurate driving directions, the latest transit status, and useful search results.

The Google Privacy Policy governs how user data is collected and used by Maps and other Google services and is designed to ensure that we collect data only where it is necessary for the user's intended purpose. In addition to the use of real time location data, users may turn on Location History in their Google account settings to opt into preserving precise historical location data. Location History is off by default. On Maps, real time location data is used even when Location History is off, and people who use our services can also choose to share (or not share) their real time location with others regardless of their choice of settings for Location History.

Our well-established policies, procedures, and options for users result in low residual risk for Maps collecting, processing, aggregating, or sharing more user information than is necessary for the stated purposes.

Protecting Contributors to Maps

The nature of user-generated content tied to locations on Maps opens up the potential for unintended or malicious disclosure of private or highly personal information about users attached to a specific location. While the complexity of data choices, and the link between content and location, means that the unintentional sharing of information will always remain a risk, the user choice whether to submit results in much lower levels of residual risk. Maps is primarily a local discovery and navigation service and not designed for sharing personal information, so we also concluded that malicious data sharing is unlikely to occur on the service and that, when considered alongside our existing mitigations, the residual risk of malicious data sharing is low.

Addressing Risks Relating to Images and Personal Information on Maps

We take several steps to protect the privacy of individuals when Street View imagery is published to Maps. We have developed cutting-edge face and licence plate blurring technology that is designed to blur identifiable faces and licence plates within Google- and user-contributed imagery in Street View. If we do not automatically or completely blur an image, users and non-users can request that Maps do so if their face or licence plate requires additional blurring, or if they would like us to blur an entire house, car, or body.

During the assessment period, and consistent with Article 35 of the DSA, we refined our personal information policy, which directs users to request content removal if they believe personal information has been posted without their consent.

Protecting Children's Rights

The nature and purpose of Maps (helping users navigate from A to B, making available accurate and reliable information about places, business, and experiences) results in lower levels of inherent risk for

children's rights. With policies designed to ensure accuracy and relevance of content we are able to place fewer restrictions on children when compared to other Google services.

The Maps experience is largely the same for children in the EU, except that those under 13 (or the minimum age in their country) cannot contribute content (including photos, ratings, and reviews), publish public place lists, add or edit places on the map, or turn on Location History. Children under 13 (or the minimum age in their country) whose accounts are managed with Family Link can only share their real-time location with their parents, and won't see where they went with their devices or get recommendations based on visited places. You can read more in Google Maps and your child's Google Account.

As described in more detail above, Google Ads policies, including the specific policies on Ads allowed on "made for kids" content and "ad-serving protections for teens," apply to ads shown on Maps. We prohibit personalised ads to any users determined to be under the age of 18, for whom ads may only be served based on non-personalised contextual information, such as the content being viewed.

Because of the underlying nature of the service design, the safety functionality built into Maps resulted in lower residual risk for children's activity.

Enhancing Accessibility

During the 2024 assessment period we introduced several new accessibility features and updates that make accomplishing daily tasks more accessible.

For example, we made Lens in Maps (which uses AI and augmented reality to help people orient themselves in an unfamiliar neighbourhood) more accessible and useful for people who are blind or low-vision by adding screen reader capabilities that provide auditory feedback of the places around the user, with helpful information like the name and category of a place and how far away it is.

We also made the option to request wheelchair-accessible walking routes available globally wherever we have data available. Not only is this helpful for people who use wheelchairs, but it's also useful for people travelling with things like luggage or strollers. This feature builds on our wheelchair-accessible transit navigation option that shows step-free transit routes.

For those in need of hearing assistance, business owners can now add the Auracast attribute to their business profile. Auracast broadcast audio allows venues—like theatres, gyms, places of worship, and auditoriums—to broadcast enhanced or assistive audio to visitors with Auracast-enabled Bluetooth hearing aids, earbuds, and headphones.

Google Play

Play

Description of Service and Associated Risk Profile

On Google Play, users find and download their favourite apps, games, books, and more. Play provides millions of apps and games to over 300 million users[23] in the EU. Play ranks and organises apps in order to help users discover the most relevant apps for them on Play through features such as categories, For You, and recommended for you. Ads and sponsored content are clearly marked.

Multiple factors are considered to decide which apps to show when users search, how many apps to show, and how they appear, including:

- **Relevance:** We show apps that are relevant to the page users are on or what they searched
- **Quality of the app experience:** We show apps that have good in-app user experiences based on several factors, including app design
- **Editorial value:** We curate recommendations based on what's noteworthy and interesting
- **Ads:** When developers advertise their apps, we make sure the ads are appropriately identified
- **User experience:** We show apps that perform well in the Play Store and that users continue to enjoy after installation

These factors are weighted differently depending on the user device, user preferences, and where they are looking in Play.

Users can manage how their experience is personalised on the Activity Controls of their Google accounts, where they can turn off personalisation by turning off Web and App Activity, or by deleting past activity.

Play connects millions of developers to billions of users worldwide and invests in the platform, tools, services, and marketing opportunities that support developers. This investment allows small or nascent developers to benefit from economic opportunity and contribute to a healthy, competitive ecosystem. In fact, 97% of developers pay no service fees to benefit from Play. We believe higher numbers of active developers, subject to compliance with our consumer-protection policies, results in wider choices for users.

[23] Average monthly counts based on distinct signed-in accounts of recipients.

This report will primarily focus on apps and games (collectively referred to as apps) as the main drivers of systemic risk relevant to Play. Our assessment also examined the other forms of content on Play, such as books, and found they posed less risk to users because the more standardised content lacks the dynamic data and user-generated challenges inherent in apps.

We take our responsibility to provide a safe and trusted experience for all users very seriously and provide a platform for developers to deliver apps safely to billions of people worldwide. To help achieve this, we establish and seek to enforce clear expectations via our Google Play Developer Program Policies, which cover topics such as restricted content, privacy, malware, and monetisation. We also help keep users safe by building protections into Play, requiring developers to follow high safety standards. You can read more in our description of How Google Play Works.

Play is a "platform of platforms." Many of the apps available through Play are also platforms themselves; in these instances, the app hosted on Play is the front door into a user experience controlled by the third-party app or game developer. This structure creates two dimensions of risk, which you will see reflected in our systemic risk assessment: risks related to the Play platform (e.g., hate speech in a review left on Play) and risks created by third-party apps on the Play platform (e.g., hate speech in a post within a social media app).

This separation reflects the appropriate allocation of systemic risk among Play and the apps that appear on the Play platform. While Play's risk assessment references both dimensions of risk, it is focused on our role in the mitigation of risks to the Play platform, with app-level mitigations most appropriately taken by the developers of those apps. As Recital 27 of the DSA notes, requests or orders related to the removal of illegal content should be "directed to the specific provider that has the technical and operational ability to act against specific items of illegal content, so as to prevent and minimise any possible negative effects on the availability and accessibility of information that is not illegal content."

Systemic Risk Assessment Results and Associated Observations

We assessed 42 different risk statements[24] for inherent risk (i.e., risk absent any action taken by Google) preparedness (i.e., the cumulative measures currently in place to mitigate the risk) and residual risk (i.e., risk after mitigation by Google). Residual risk serves as a guide for where further investment may be warranted. The full list of risk statements is found in the Annex A to this report.

This systemic risk assessment surfaced important themes relating to the inherent and residual risk.

In the first of the two sections that follow ("Content Moderation") we consider risks and mitigations relating to content moderation on Play, which primarily pertains to apps themselves as content. While there is inherent risk of illegal or harmful content appearing on apps, this section explains Play's app review and moderation program, which results in much lower levels of residual risk. We also discuss how we address other types of user-generated content, such as reviews on Play.

In the second of the two sections that follow ("Platform Design"), we consider risks and mitigations related to the way Play functions. Three important inherent risk themes that emerged during the assessment were related to privacy, security, and child rights. These risks reflect Play's role in the overall app ecosystem, and this section explains the actions we take that result in much lower levels of residual risk.

Taken together, these two sections address the four broad categories of systemic risks articulated in Article 34(1) of the DSA and the specific manifestations of those systemic risks that we evaluate. This report emphasizes those risks for which the assessment showed elevated inherent or residual risk, and describes Play's current risk mitigation practices as well as improvements consistent with Article 35 of the DSA. We highlight where changes in the external context or improvements to our mitigation measures cause significant modifications to inherent or residual risks between the 2023 and 2024 systemic risk assessments.

2024 Highlights

- Launched a new Play Reviews Policy Violation Report Form for users to flag Play Reviews that potentially violate Play policies and a new Appeals Form for users who have left reviews to appeal removals that they may disagree with
- Made changes/refinements to our Developer Program Policies, on topics such as health apps, photo/video permissions, manipulated media, generative AI, and child safety

[24] *See* Methodology Step One: Classification.

- Launched new developer validation ("Know Your Developer") and pre-publication app testing requirements, which reduced some residual risks, such as unintentional data sharing, scams, malware, and phishing
- Expanded the Play software developer kits (SDK) Index to enhance app quality and address risk

Content Moderation

Removing Illegal Content

Play has appropriate policies in place prohibiting illegal content on the platform through policies related to restricted content, intellectual property rights, and other policies preventing fraudulent or malicious apps. As discussed previously in this report, Google has a developed process for evaluating government requests to remove content.[25] Additionally, Play has reporting channels for users to report illegal content or content that violates Play policies. There is extensive overlap between content prohibited by Play's product policies and content that is illegal, meaning that our policy development and enforcement efforts work to mitigate the risks of both illegal and policy-violative content. These efforts are described in detail in the next section.

The Play risk assessment identified a range of relevant illegal content-related inherent risks, including risks such as CSAM, terrorist content, apps infringing intellectual property rights, and illegal activity like scams. However, the assessment found that robust policies that are binding on apps (as described below) and enforcement of these policies resulted in much lower levels of residual risk.

As noted above, illegal content appearing within apps available on Play is primarily the responsibility of developers, though (as described below) Play takes a variety of enforcement actions against developers with multiple or egregious policy violations.

Addressing Content that Violates our Policies

The systemic risk assessment reviewed numerous risks relating to a wide variety of harmful content, such as content impacting human dignity, promoting discriminatory beliefs, inciting, praising, or glorifying violence, promoting practices harmful to health, inciting gender-based violence, or constituting harassment and bullying.

There were several factors that caused us to conclude that these risks are of much lower residual risk for Play. There are high costs associated with developing an app as compared to a single piece of user-generated content, thus well-designed policies ensure developers are effectively disincentivised from spending time and resources developing apps that clearly violate Play policies. Additionally, when app

[25] For more information on how YouTube and Google respond to government requests to remove content, *see supra* at Handling Government Removal Requests.

developers submit their apps to Play, we use a combination of automated processes and human review to assess these apps before they can be published for distribution on the Play Store. The automated processes—which include static and dynamic components—scan an app's code, app images, the developer profile, and the app description.

We review millions of apps submitted to Play each year, including technical reviews of code for malware. If we identify policy violations at this pre-publication stage, we reject the developer's submission and give the developer an explanation of the policy issues along with instructions on how to correct them. Once the issues are addressed, the developer can resubmit their app or app update for further review. If we find no policy violations, we publish the app or app update to the Play store. As explained below, we also have robust developer education processes to keep apps in compliance with evolving policies and enforcement mechanisms when they fall out of compliance.

These estimates of residual risk also rest on the distinction raised earlier: between user-generated content moderation that Play can undertake and content moderation responsibilities held by app developers, who may themselves be providing a user-generated content platform.

For example, we establish clear requirements around robust, effective, and ongoing user-generated content (UGC) moderation in apps. But only an app developer operating a UGC platform can remove specific pieces of content (e.g., a specific post in a social media app; a video from a streaming app) from its platform. Google can only remove the UGC platform app in its entirety—including all legitimate content within it—from Play. This limitation raises questions of proportionality, fairness, freedom of expression, and user impact, all of which must be balanced against the risks that may be posed by the specific underlying content.

In March 2024 we launched a Policy Violations Form for Play Reviews for users to flag reviews that potentially violate Play policies and a new Appeals Form for reviewers to appeal removals that they may disagree with. We also created a new Help Center article covering all reports.

Maintaining Developer Policies

Our Play Developer Policies set out what developers can and cannot provide users on the Play platform and are the foundation upon which Play delivers apps and games safely to billions of people worldwide. These policies cover areas such as restricted content, impersonation, monetisation and ads, privacy, malware, and mobile unwanted software, and are relevant across most of the risk statements included in this systemic risk assessment.

Our User Generated Content (UGC) Policy requires providers of apps that contain UGC services to implement ongoing UGC moderation and sets out requirements in the areas of informed consent, defining objectionable content and behaviours on the app, and undertaking reasonable UGC moderation that is consistent with the type of UGC hosted by the app. It also includes requirements for reporting channels and an in-app system for blocking UGC and users. We believe that requiring providers of apps that contain UGC services to implement these policies—and requiring developers to play a role in enforcing them—provide an appropriate, proportionate, and effective approach for Play.

Developing Policy

We update the Developer Policies over time to reflect insights into new and emerging risks, and conduct regular reviews of our policies based on developer feedback, external media, expert and stakeholder feedback, and internal enforcement data. The practice of constantly updating our policies based on emerging threats also reduces many or most of the systemic risks considered in this risk assessment. We maintain a page detailing changes to our policies, including when these changes take effect and resources to help developers prepare for app updates.

Over the years, Play has taken strides in developing our policies such that the risks posed by apps and the content that appears on apps have been appreciably lowered.

For example, we created the UGC Policy mentioned above in response to the evolution of social media platforms, which included significant fleeting and/or real-time content, and new types of UGC, which had resulted in heightened societal concern about user safety. After market research, user studies, and collecting insights from developers in global markets, Play established a developer policy requiring in-app moderation for all UGC apps. Play's UGC Policy reflects the belief that users should have a direct means to contact social media platforms, which should be held accountable for consistently moderating content.

When specific user harm concerns arise, Play's policy development team goes through a rigorous process to understand the issue, develop guardrails, internally test those guardrails, and then introduce new policies into our ecosystem. After introduction, we monitor the impact of our policies to refine or expand protections, as needed.

For example, in 2021, we introduced our Personal Loans policy. This policy was developed in response to user feedback from India and several Southeast Asian markets that developers were charging high and often illegal interest rates to users, and that some developers were blackmailing users with the permissions they had obtained through apps. We mandated declaration and disclosure of financial agreements between users and financial apps or their developers, so that we can verify the legitimacy of the loan agreements and make sure the loan terms are clear to users. We later expanded the policy to prohibit these apps from accessing sensitive data, such as photos and contacts, which were being used to verify credit worthiness in markets without formal credit-scoring mechanisms.

In 2020, we added a Stalkerware policy to address code within apps that collects personal or sensitive user data from a device and transmits the data to a third party. Our policy requires prominent disclosure and consent for a narrow set of permissible uses and prohibits these apps for all other uses. Only apps designed and marketed for enterprise management or for parents to monitor their children's activities are allowed to have such functionality. Play prohibits apps used to track anyone else, even with their knowledge and permission.

Some of the main policy changes made during the assessment period included:

- Adding a new Health apps policy to introduce new requirements and disclosures for apps that meet the definition of Health apps and updating the policy to reflect current public health guidance.

- Introducing a new Photo and Video Permissions policy to reduce the number of apps permitted to request broad photo and video permissions.
- Updating our Manipulated Media policy to include audio as an additional example of a type of media covered under the policy.
- Introducing a Child Safety Standards policy requiring Social and Dating apps to follow specific child safety standards and self-certify compliance on Play Console before publishing.
- Emphasizing with developers that all generative AI apps must comply with our policies and have in-app user reporting or flagging features that allow users to report or flag offensive content to developers without needing to exit the app. Developers should utilise user reports to inform content filtering and moderation in their apps.

We recognise that smaller developers may have fewer resources to help them understand our policies or keep up with changes, so over the last three years we have expanded our education and support efforts. We now offer the Google Play Academy, where developers can take courses to better learn our platform, and PolicyBytes videos about policy updates. We stream global webinars throughout the year where we make major policy announcements, and we offer the Google Play Developer Help Community for developers to get advice from other expert developers.

Well-developed policies were a significant factor in lowering the content-specific residual risks on Play. We strive to ensure these policies are constantly reexamined and updated, as well as binding for developers.

Enforcing Policy

Play uses a combination of human and automated evaluation to review apps and app content to detect and assess content which violates our policies and is harmful to users and the overall Play ecosystem. Using automated models helps us detect more potential violations and evaluate potential issues faster, which helps us better protect our users and developers. The policy-violating content is either removed by Play's automated models or by trained operators and analysts. The results of these manual reviews are then used to help build training data to further improve our machine learning models.

Our EU DSA Biannual VLOSE/VLOP Transparency Report discloses the human resources capacity evaluating Play content across the official EU Member State languages.

If an app does violate any of our policies, we take appropriate, necessary, and proportionate action pursuant to our enforcement processes. These actions may include app rejection (for apps and app updates submitted for review prior to being made available on Play), app removal (for existing apps), app suspension, limited visibility, limited regions, and account termination (for multiple suspensions or an egregious policy violation). Additionally, Play users can report an app policy violation and flag individual app reviews as inappropriate through a link on the Play listing. We offer an appeals mechanism for developers who believe there has been an enforcement error.

In order to protect developer's rights, when we deploy new policies on Play, developers generally have at least 30 days from the announcement of the new policy to make changes to their apps, and longer if the

updates are likely to require significant technical changes. Because app removal can negatively impact users and developers, in addition to giving time for compliance, we invest heavily in efforts to educate developers about our policies and how to comply. Education lessens the need for enforcement and keeps well-intentioned developers and apps on Play. Enforcement of, and education about, our policies are key aspects of Play's moderation and user safety program that resulted in lower residual risks for much of our assessment.

Addressing Specific Content-Related Risks

Specific content-related risks feature less prominently as inherent or residual risks for Play, since individual apps (rather than Play) have a greater determining role in creating and managing these risks. For example, social media apps are available on Play, but those apps are primarily responsible for enforcing their own UGC policies. Play does face risks when it comes to UGC in the form of app reviews, such as efforts to influence the visibility of apps, either positively or negatively, with inauthentic reviews (known as "review bombing"). However, our efforts to address this risk, combined with the fact that users are often searching for a specific app, reduces the residual risks considerably.

Preventing Review Bombing and Ensuring Rating and Review Integrity

While Google cannot undertake content moderation for all in-app content of apps available on Play, we work to ensure the integrity of app, game, books, and movie reviews on Play.

There are several ways that we work to moderate ratings and reviews. Both qualitative comments and quantitative ratings (i.e., one to five stars) are monitored, especially to detect coordinated campaigns to either artificially boost or downgrade a listing's rating. We deploy specialised algorithms to identify signals that may indicate coordinated attacks (e.g., duplicate or repeat reviews), which are then reviewed by humans. And as discussed earlier in this report, we are improving our systems' ability to detect violative content across different languages.[26]

We also work to ensure app ratings present an accurate picture of the current user experience by calculating ratings using a percentage of the most recent reviews, not the average of all the ratings, to determine the overall rating for an app. This methodology protects against impact spikes that sham ratings can have on an app's rating and helps lead instead to more accurate ratings that reflect true user sentiment towards app experiences. We believe that the work we do to root out sham ratings leads to a more transparent app ecosystem, which ultimately supports the visibility and availability of a diversity of viewpoints and content on our platform.

During this assessment period, and due to changes in the geopolitical climate, the Play Reviews team enacted a process that temporarily pauses the ability for users to post reviews during sensitive events and times of conflict to allow more time for our systems and teams to detect, address, and enforce issues as they arise and help ensure correct enforcements at scale.

Our EU DSA Biannual VLOSE/VLOP Transparency Report discloses the number of actions we took on Play, including advertisements presented on Play, segmented by type of illegal content or violation of terms and

[26] *See supra* at Evaluating Content Across Languages for further discussion of how Play, and Google as a whole, are addressing this identified residual risk for Play.

conditions. The most recent report (April 2024) demonstrates that spam and content that is inappropriate or unhelpful are the most common policy violations, accounting for over 95% of cases. Other violations (e.g., privacy and data protection, scams and fraud, and illegal and harmful speech) are much less prevalent. Over 99% of these actions were taken following automated detection. The report also discloses the number of complaints received from users located in EU Member States about content moderation decisions and the outcome of these complaints.

Protecting Civic Discourse

We conducted a comprehensive evaluation of systemic risks associated with civic discourse. Some inherent risks relate to (1) the risk of apps engaging in mis- and disinformation relating to elections, civic discourse, or democratic participation and (2) digital threats such as targeted account hijacking, phishing, and targeted disinformation campaigns. While significant mitigations are in place to address these inherent risks (described below), room for improvement with respect to these dynamic threats remains.

We have made significant investments in addressing civic discourse mis- and disinformation risks through the introduction and enforcement of clear Play Policies.

For example, we introduced minimum requirements that apps must meet prior to being classified in the News category, including transparency requirements about the source and ownership of in-app news content, requirements applicable to news subscription services, and requirements regarding the use of affiliate marketing and ad revenue.

To protect integrity in elections, our Deceptive Behaviour policy prohibits apps from making misleading claims or providing false information about the app, including demonstrably deceptive or false content about an app's capabilities or functionality that may interfere with voting processes. For example, an app that misleads voters into believing they can cast their vote through the app would violate these policies. The policy additionally prohibits apps that promote or help create false or misleading images, video, and/or text, and requires apps that manipulate or alter media to prominently disclose or watermark the altered media.

We found that app development and usage were the most relevant for the systemic risk assessment because apps contain layers of their own hosted content, may collect user data, and are constantly being updated and evolving. However, the assessment also considered other types of content, such as books on Play. Our Publisher Content Policies for Google Play Books are specific to book publishers and set out what books publishers can and cannot distribute to users on the Play platform. These policies cover areas such as hate speech, child safety, misleading content, and copyright. Because of well-developed and enforced policies in this area, we did not identify a significant residual systemic risk with respect to offerings on Play other than apps, such as books.

Platform Design

Knowing Developers and Protecting Users

One of the many ways we keep Play a safe and trusted platform is by verifying the identity of developers and their payment information. This helps prevent the spread of malware, reduces fraud, and helps users

understand who's behind the apps they're installing. These mitigations help enable us to address the higher inherent risks arising from the wide availability of generative AI tools—such as the greater ease by which bad actors can create fake apps and scams—to achieve lower levels of residual risk.

In July 2023, we announced new verification requirements for anyone creating new Play developer accounts, and are now in the process of implementing these verification requirements with all existing Play developers. These verification requirements were introduced to boost trust and transparency on Play (such as preventing bad actors from distributing malware) and help ensure that developers comply with Play policies.

Our revised list of verification requirements now include the use of D-U-N-S numbers, a unique nine-digit identifier that is widely used to verify businesses, alongside requirements such as legal name and address, email, and Google payment profile. In addition, developers with unverified bank accounts will have their developer presence and apps removed from Play.

These and other investments enabled us to identify bad actors and fraud rings more effectively. During 2023 we banned 333,000 bad accounts from Play for violations like confirmed malware and repeated severe policy violations. We also prevented 2.28 million policy-violating apps from being published on Play, in part thanks to our investment in new and improved security features, policy updates, and app review processes.

Developers are able to appeal enforcement actions on Play apps, though our most recent EU DSA Biannual VLOSE/VLOP Transparency Report (April 2024) shows that less than 1% of all automated enforcement actions are reversed following a successful appeal submitted by EU developers.

The verification methods differ between individual accounts and organisation accounts. In November 2023 we also introduced new app testing requirements for individual developers publishing new apps on Play, as well as a review by Play before production. Called "Start on Play," this requires individual developer accounts to run a closed test with a minimum of 20 testers for 14 days before an app can be distributed on Play. These new requirements are intended to address the risk that new individual developer apps on Play may violate Play policies.

In early 2024 we also rolled out a new "Government" badge in the Play Store to help identify official government-made apps. Completing the government declaration requirements also allows government apps to become eligible for official endorsement signified by a clear visual treatment on the Play store, provided further eligibility criteria is also verified. At launch, the label appeared on official state and federal government apps in Australia, Brazil, Canada, France, Germany, India, Indonesia, Japan, Mexico, South Korea, United Kingdom, and the United States, with plans to grow over time.

While our strategy is focused on knowing developers and preventing policy-violating apps from being published in the first place, sometimes we have reason to take legal action against fraudsters with multiple egregious violations. For example, in 2024 we filed a lawsuit against two fraudsters who made multiple misrepresentations to upload fraudulent investment and crypto exchange apps on Play to scam users globally. This lawsuit is a critical step in holding these bad actors accountable and sending a clear message that we will aggressively pursue those who seek to take advantage of our users.

Protecting Privacy

Reflecting the fact that Play exists in the app ecosystem and offers apps in categories that are likely to involve the use of personal data (e.g., banking or government services) some of the highest inherent risks for users who access content through Play relate to privacy.

While we strive to maintain an open and accessible Play and maximise user choice, we also enforce safety standards for apps through our developer policies, ensuring we provide a more safe and secure environment for app users than would exist without Play. These measures are typically sufficient to lower residual risk considerably, such as risks relating to the collection and use of sensitive personal data without consent; however, the determined and constantly evolving nature of bad actors caused us to conclude that some elevated residual risk for phishing, malware, and malicious apps remains. An explanation of this elevated residual risk and related Article 35 mitigations are below.

Our developer policies create consistent safety standards for apps that appear on Play, and generally give users additional transparency and control over their personal data. These policies include heightened protections for personal and sensitive user data, which prohibit developers from selling personal and sensitive user data, and require developers to limit the access, collection, use, and sharing of personal and sensitive user data acquired through the app to purposes reasonably expected by the user. We maintain data deletion requirements that require developers to delete associated data when they receive an account deletion request unless the user indicates they want their data preserved or certain other exceptions apply.

We also offer a "Data safety section" for apps. The Data safety section provides developers with a transparent way to show users if and how they collect, share, and protect user data, before users install an app. Developers are required to tell us about their apps' privacy and security practices by completing a form in Google Play Console. After a developer completes and submits the Data safety form, Play runs automatic checks on key elements of the information provided as part of the app review process. This information is then shown on the app's store listing on Google Play. We cannot wholly know what data a developer collects and shares and so compliance remains the responsibility of the developer; however, if we become aware of a discrepancy between app behaviour and this declaration, we may take appropriate action, including enforcement action. With strengthened platform protections and policies, and developer outreach and education, we prevent submitted apps from unnecessarily accessing sensitive permissions.

Apps that are deceptive, malicious, or intended to abuse or misuse any network, device, or personal data are strictly prohibited. You can read more about our approach to topics such as user data, permissions, misrepresentation, and deceptive behaviour in Privacy, Deception and Device Abuse. However, motivated bad actors are constantly evolving their tactics to circumvent known protections on Play, so we assessed this as having some medium levels of residual risk remaining. In recognition of this challenge, during the assessment period Play established new verification requirements and pre-publication reviews (described above) that help address this risk.

Other privacy concerns relate to the use of software developer kits (SDKs). App developers often rely on third-party code, or SDKs, to integrate key functionality and services for their apps. We are clear with developers that our existing privacy and security requirements apply in the SDK context and are designed to help developers safely and securely integrate SDKs into their apps. In 2022, we launched the Google Play SDK Index to help developers evaluate an SDK's reliability and safety and make informed decisions

about whether an SDK is right for their business and their users. You can read more about our approach in SDK Requirements.

During the assessment period we partnered with certain SDK providers to limit sensitive data access and sharing, enhancing the privacy posture for over 31 SDKs impacting over 790,000 apps. We also significantly expanded the Google Play SDK Index, which now covers the SDKs used in almost 6 million apps. This valuable resource helps developers make better SDK choices, boost app quality, and minimise integration risks.

We believe these policies are investments to protect our users and help developers meet consistent standards. When it becomes apparent that a developer is not meeting our established requirements for privacy and user safety, we take action to remove offending apps or developers from Play. We have prevented policy-violating apps that were submitted for publishing from appearing on Play with improved security features and policy enhancements.

Protecting Children's Rights

The systemic risk assessment reviewed several risks relating to child rights, and found the highest inherent risks to include the risk that children under a defined minimum age may access services that they should not be able to, that children's data may be used for ads targeting, that apps may not function equitably for children of varied learning styles, and that apps primarily directed at children may not be of an adequate quality across languages, markets, and age groups.

However, the systemic risk assessment concluded that the combination of platform and service design measures and policies that Play has in place are reasonable, proportionate, and effective mitigation measures. The assessment reinforces our view that we provide a more safe and secure environment for app users than would exist without Play.

Several important features of Play address child safety risks on the platform:

Maintaining Additional Policies for Minors

We have additional requirements for apps that are targeted at children under the age of 13. Before an app is published on Play, the developer must certify whether children under the age of 13 are part of the target audience and, if so, the app must comply with the Google Play Families Policies (in addition to the standard Play Developer Program Policies). While developers generally are in the best position to identify the correct audience for their apps, in some instances, we may disagree with a developer's stated age designations and redesignate the app.

The Play Families Policies establish heightened obligations for developers regarding age-appropriate content, data practices (e.g., not making use of precise location data and no personalised ads for users known to be under 18), and social app features. Apps subject to these policies must also disclose in greater detail how they use the user data they collect. A dedicated enforcement workstream that uses both automated protections and human reviewers enforces the Google Play Families Policies.

Providing a Teacher-Approved Program

Play makes it easy for families to find quality content for children. Play's Teacher Approved program is a quality review program for apps that specifically target children under the age of 13. It collects ratings from teachers, children's education specialists, and media specialists, who rate and approve apps based on a range of quality criteria (i.e., whether apps are fun and inspiring, age-appropriate, and thoughtfully designed). Approved apps are included in Play's Kids Tab, along with a description of their quality attributes, to help families easily review the apps and make informed choices for their children.

The program provides an additional layer of review, insight, and quality control on top of the Google Play Families Policies.

Obtaining Age Assurance

Users can view Play on the web without being signed into a Google Account, but must sign in to download, purchase, or install content on Play, whether on the web or on the mobile store. Adult content is not available in a signed-out state and is blocked for signed-in users under the age of 18. As described further above, Google utilises age assurance technology, along with a neutral age-screen in the Google Account sign up process, to help determine which recipients are likely under the age of 18. Recipients identified as likely being under 18 are subject to heightened privacy, content, and safety protections. To reduce the burden on our recipients and in accordance with data minimisation principles, these processes are carried out at the Google Account level, so that the results can then be used in connection with all Google services, such as on Play.

As a part of age assurance during Google Account sign up, if a user is under 13 (or the minimum age in their country) then a parent, guardian, or caregiver's consent is needed to continue to sign up for or use the Google Account.

When a child reaches their country's minimum age to manage their own Google Account, the child can choose to continue their current parental supervision settings or manage their own account. Family Link facilitates a range of parental controls on Play for supervised Google Accounts, including purchase controls, approving or blocking apps, and filtering content based on content ratings.

Enforcing Content Ratings and Content Restrictions

We incorporate official content ratings from the International Age Rating Coalition (IARC) into Play ratings. The IARC is administered by a group of participating regional ratings agencies. IARC ratings are designed to help developers communicate locally relevant content ratings to recipients. Ratings are assigned by a regional authority based on a rating questionnaire completed by the developer and displayed in each app's listing page on Play. IARC ratings may be updated when developers make changes to their app's content or features that affect issues in the IARC questionnaire.

IARC ratings are used to aid parental controls and to restrict access by recipients under the age of 18 to mature-rated content where legally required. For supervised Google Accounts, parents can filter or block content based on IARC ratings (i.e., limit their child to seeing content rated PEGI 16 or below). Unrated apps are treated as high-maturity apps for the purpose of parental controls until they receive a rating.

Play blocks the purchase or download of mature-rated content in the EU, unless we have signals providing sufficient confidence that the recipient is an adult. In some circumstances, we require users to provide additional verification (e.g., by providing evidence of a government ID or credit card) of their age. We might require such verification if a user is trying to access mature-rated content or services, and we cannot otherwise establish with sufficient certainty that they are an adult, or if our model has classified the user as under 18 but the user wishes to verify eligibility to access such content.

Shopping

Description of Service and Associated Risk Profile

Google Shopping helps users discover and learn about the products they are interested in, whether from a big-box retailer, direct-to-consumer brands, or the local store. Users use Shopping to search for products and compare prices between different merchants. They then buy products directly from the merchant on the merchant's website or at their physical store, not on Google. Our mission is to democratise e-commerce by supporting an open network of retailers and shoppers, help businesses get discovered, and give users more options when they are looking to buy.

Shopping uses a variety of factors to determine which products are displayed in search results, including a product's price, availability, and relevance to the user's query. Users can filter Shopping results by price, brand, and other criteria.

Merchants use the Merchant Center to manage their product data for Shopping and either use free product listings or ads to promote their products. All ads are clearly marked as "Sponsored" or "Ad."

Our Shopping Graph is a dynamic, AI-enhanced, and real-time dataset of product listings, sellers, brands, reviews, product information, and inventory. Listings are updated constantly based on information retailers share directly via Google Merchant Center or from what retailers and brands post across the web. The Shopping Graph makes those sessions more helpful by sorting through a vast set of products to connect people with around 50 billion listings globally across the web. Shopping is used by over 30 million average monthly users in the EU.[27]

In addition to the content promoted by merchants, Shopping includes user-generated content in the form of product and merchant reviews and ratings. Google collects some reviews and ratings directly through Google Customer Reviews, a free program that merchants enable to allow Google to collect feedback on their behalf. Shopping also features reviews and ratings collected using a merchant's own UGC service or a third party service working in a software as a service model (e.g., Yotpo, Avis Vérifiés).

You can read more in How Shopping Works, How Merchant Center Works, and Shopping Graph.

[27] Average monthly counts based on distinct signed-in accounts of recipients.

Systemic Risk Assessment Results and Associated Observations

We assessed 39 different risk statements[28] for inherent risk (i.e., risk absent any action taken by Google), preparedness (i.e., the cumulative measures currently in place to mitigate the risk), and residual risk (i.e., risk after mitigation by Google). Residual risk serves as a guide for where further investment may be warranted. The full list of risk statements is found in Annex A to this report.

This systemic risk assessment surfaced important themes relating to the inherent and residual risk. Because Shopping operates on limited types of content that are directly related to products available on Shopping, many of the risks covered by the systemic risk assessment (such as CSAM, illegal hate speech, and election mis- and disinformation) have a lower likelihood of appearing.

However, risks relating to privacy, the freedom to conduct a business[29], consumer protection[30], and intellectual property[31] feature more prominently given the role of Shopping in presenting and raising the visibility of products sold by merchants. For these themes the systemic risk assessment identified several areas of important inherent risk that are being appropriately addressed, as seen through high preparedness evaluations, resulting in much lower levels of residual risk.

In the following two sections we consider the risks and mitigations relating to illegal and policy violating content ("Content Moderation") and the design and functioning of Shopping ("Service Design"), though in practice there are several interactions and relationships between the two.

Taken together, these two sections address the four broad categories of systemic risks articulated in Article 34(1) of the DSA and the specific manifestations of those systemic risks that we evaluate. This report emphasizes those risks for which the assessment showed elevated inherent or residual risk, and describes Shopping's current risk mitigation practices as well as improvements consistent with Article 35 of the DSA. We continue to invest in efforts to reduce residual risk, but did not identify any significant modifications to inherent or residual risk between the 2023 and 2024 systemic risk assessments.

2024 Highlights

- Enhanced methods to review and verify merchant identity-related signals
- Implemented infrastructural changes aimed at reducing the time required for detection and enforcement, enabling prompt identification and enforcement action

[28] See Methodology Step One: Classification.
[29] Article 16 of the EU Charter: Freedom to Conduct Business.
[30] Article 38 of the EU Charter: Consumer Protection.
[31] Article 17 of the EU Charter: Right to Property.

Content Moderation

Removing Illegal Content

Identifying and Blocking Illegal Products and Services

One risk associated with Shopping is that merchants may promote or attempt to sell illegal products and services through Shopping. Here, we assessed that determined bad actors seeking to use Shopping services for the sale of illegal products or services constitute higher levels of inherent risk, but our effective mitigations result in much lower levels of residual risk. For example, our EU DSA Biannual VLOSE/VLOP Transparency Report discloses that we take action on cases relating to animal welfare, healthcare and medicine, and other unsafe or illegal products millions of times per year.

Shopping has a robust set of policies that prohibit the sale of illegal products and services, including those relating to gambling, abuse of the network, local legal requirements and safety standards, dishonest behaviour, and healthcare and medicines.

Prohibiting and Detecting Violations of Intellectual Property Rights

Shopping maintains policies that address the sale of goods that infringe on the intellectual property rights of others, such as our counterfeit, trademark, and copyright policies.

Shopping prohibits the sale or promotion of counterfeit products. Malicious actors may attempt to leverage Shopping to disseminate counterfeit goods, but our robust reactive and proactive enforcement scheme means Shopping is well prepared to address this risk, resulting in lowest residual risk.

Shopping uses well-established proactive detection measures for counterfeit violations, which include techniques like keyword matching and detection of signals that may indicate merchants are promoting trending products with unrealistically low prices.

Additionally, trademark owners can report merchants offering counterfeit goods in a dedicated reporting channel. Where a merchant is identified as promoting counterfeit goods, its Merchant Center account is typically suspended.

Trademark owners can also report Shopping content that uses their trademarks in a way that is likely to cause confusion about the origin of a product. Our teams review each notice carefully, including confirming that the reporter has valid trademark rights. Where the notice is complete and we determine that the content violates our trademark policies, we remove the content from Shopping.

We provide a simple and efficient mechanism for copyright owners from countries/regions around the world. To initiate the takedown process, a copyright owner who believes content is infringing sends us a takedown notice for that allegedly infringing material. When we receive a valid takedown notice, our teams carefully review it for completeness and check for other problems. If the notice is complete and we find no other issues, we remove the content from our services.

Addressing Content that Violates our Policies

Maintaining Google Shopping Policies

We have two categories of policies—Free Listings Policies and Shopping Ads Policies—that outline what is and is not permitted on Shopping, including for product listings pulled from what retailers and brands post on their websites.

The Free Listings Policies and Shopping Ads Policies prohibit content that is harmful to customers or the overall shopping and advertising ecosystem.

Both sets of policies cover four broad areas:

1. Prohibited content, meaning content that is not allowed to be listed, such as counterfeit products, dangerous products, and inappropriate content;
2. Prohibited practices, meaning things merchants cannot do if they want to list products, such as misrepresentation of content;
3. Restricted content that can be listed with limitations or in certain locations only, such as adult-oriented content, alcoholic beverages, and healthcare-oriented content; and
4. Editorial and technical content, meaning website standards, such as irresponsible data collection and use.

In addition, Shopping enables users to report listings and ads that violate policies and/or contain illegal content, and enables brand and trademark owners to report merchants misusing their brand or trademark. A dedicated team reviews and actions these incoming complaints.

Maintaining Guardrails for User-Contributed Content

User-contributed product and seller reviews are intended to enhance the user experience by helping users discover and select products and online sellers on the basis of opinions and feedback from other customers. We have developed user-contributed content policies and product rating policies covering content such as hateful content, misrepresentation, and fake reviews to help ensure everyone who views user-generated content has a positive experience.

An automated system processes reviews before they show up on Google to remove spammy or inappropriate language. Spammy content includes reviews with the same content posted multiple times or from multiple accounts.

After a review is published, it cannot be modified or updated by Google and we are not able to contact reviewers or ask reviewers to update what they wrote.[32] However, we may take down reviews that are flagged to us, in order to comply with legal obligations.

[32] Google Shopping has collected some reviews from the EU via the Google Customer Reviews program, and in this case users are able to delete their own reviews.

Google also enables users to report user reviews that may violate the law and to provide feedback on user reviews to improve the user experience. We are in the process of introducing additional reporting functionality to allow users to report policy-violating content as well.

Preventing Unfair Commercial Practices

We strive to create a healthy digital shopping ecosystem that is trustworthy and transparent. Customers should feel confident about the offers they are browsing and the businesses they are purchasing from. Unfair commercial practices—such as scams or representing products inaccurately—pose an inherent risk to the overall ecosystem. Shopping's policy enforcement processes significantly address this risk, resulting in low levels of residual risk.

For example, our policy on misrepresentation requires merchants to be upfront, honest, and provide users with the information that they need to make informed decisions. We disallow promotions that represent products in ways that are not accurate, realistic, and truthful. In addition, merchants are encouraged to take part in user-generated content programs to help shoppers review "real world" feedback (from Google and external sources) about product and merchant quality.

Our abuse of the network policy bans malicious content, sites that offer little unique value to users and are focused primarily on traffic generation, retailers who attempt to gain an unfair advantage in Shopping campaigns, and retailers who attempt to bypass our review processes. We also maintain a list of certain kinds of businesses with products prone to abuse. This list informs prioritisation in risk management and is regularly updated based on Google reviews, feedback from users, regulators, and consumer protection authorities.

Preventing Fraudulent Business Information

We assessed the risk that disinformation, misinformation, or fraudulent content about a business, such as fake reviews, are discoverable on Shopping. For some time, we have been using signals such as IP location, social media presence, and third party consumer research sources to mitigate these risks. In addition, we have introduced several new mitigations to address the residual risk identified in our 2023 assessment, such as reviewing and verifying merchant identity-related signals, including VAT information and identity verification.

Regarding fake reviews, Shopping has automated content checks that focus on content quality, and we employ intermittent analyses aimed at identifying anomalous review contributions. We run these checks both on an individual level (a specific merchant) and on the review source level (a review aggregator). Examples of what we might investigate further include elevated levels of 1-star or 5-star reviews or an unusual number of reviews provided by a single user or for a specific entity. We also deploy teams of trained operators and analysts who audit reviews and ratings.

Service Design

Respecting Privacy

The use of sensitive data in eCommerce (such as credit card numbers, user names, and passwords) results in critical inherent risks relating to data collection and use, and data sharing. There are two dimensions to privacy risk and mitigation on Shopping: the privacy practices of merchants, and our own privacy practices.

We do not process payments and are not involved in shipping products, so we do not collect sensitive payment data (e.g., credit cards) or pass it onto merchants, nor do we control the actions of merchants and retailers. However, we do set high expectations for merchant and retailer data collection and use on their websites, and prohibit unsafe collection or use of personal information, and misuse of personal information. Under this policy, merchants may not collect data for unclear purposes, use personal information in ways customers have not consented to (e.g., re-selling users' contact information), or without appropriate security measures in place (e.g., not obtaining certain data over non-secure SSL server connections). We have also established checkout requirements covering aspects such as accurate pricing, user information, and language use.

Where our privacy practices are concerned, storage of signed-in user data by Google is controlled by Web & App Activity and the collection and use of data is controlled by the Google Privacy Policy. To block specific advertisers or opt into personalised ads, users can visit My Ad Center. By default, Shopping ranks product listings based on relevance to a user's current search terms.

Vetting Merchants

All products and merchants go through in-depth reviews before they can list on Shopping.[33] These reviews use a combination of automated and/or human evaluation to ensure compliance with our policies, with the more complex, nuanced, or severe cases often reviewed by specially trained experts. Thanks to the Shopping Graph, our dataset of the world's products and sellers, our automated systems can quickly review whether a business is legitimate, whether the products shoppers see are accurate, and whether merchant content follows our policies. This automated vetting process has helped us more efficiently and accurately review a massive amount of merchants and products.

Sometimes we make mistakes in our decisions when vetting merchants and enforcing our policies, which may result in the unwarranted removal of products or accounts from our services. For this reason, we have enhanced our appeals process by (1) creating an appeals path for content removed based on counterfeit complaints, (2) creating appeals paths for all content removals, and (3) enabling merchants suspected of fraudulent activities to submit their EU VAT ID as an additional data option during appeal, which increases likelihood of successful account reevaluation.

[33] Google Shopping does not physically inspect products. Product review is limited to a review of virtual signals that may indicate violations of our policies.

Monitoring Merchants and Listings

Our safety efforts do not stop once a product listing goes live. Our automated systems are always monitoring for violating activity, and our team of human reviewers is on standby to review issues that might need a more nuanced perspective, such as a sudden drop in prices, a significant shift in product mix, or a change in business information. After they are onboarded, we review merchants and their listings, making sure nothing has suspiciously changed since they first came to Google. We take different types of actions when we see odd behaviour, such as removing listings that violate our policies, or suspending a merchant's Shopping account. In most cases (all except sanctioned accounts), these actions can be appealed by the merchant.

Our EU DSA Biannual VLOSE/VLOP Transparency Report discloses the number of actions we took on Shopping (including unpaid content and advertisements), by type of illegal content or violation of terms and conditions. The most recent report (April 2024) demonstrates that aside from technical violations (such as data defects, which account for more than 85% of cases), the most common violations are in the areas of sexualised content, unsafe/illegal products and services, healthcare and medicine, animal welfare, and scams/fraud. Other violations (e.g., hate speech, public security, protection of minors, negative effects on civic discourse) are much less prevalent. Over 99.9% of these actions were taken following automated detection.

The report also discloses the number of complaints received from users located in EU Member States about content moderation decisions and the outcome of these complaints. During the reporting period, <0.01% of all automated content moderation actions on Shopping were appealed by content or account owners based in the EU, and 80% of these appeals were successful.

Protecting Children's Rights

Shopping leverages measures applied to all Google services for age assurance for signed-in (including centralised Google Account solutions) and signed-out recipients.

In addition, Google ensures that adult and non-family safe listings and ads on Shopping are restricted from minors and recipients for whom we do not have an inferred or declared age. When we have insufficient signals to indicate that a user is an adult, we err on the side of turning on children's protections by default because of the critical importance of protecting minors. In this Shopping context, this makes sure that by default they cannot access products (e.g., adult products) which may not be safe for their age.

Lastly, Google has a suite of automated and manual processes aimed at scalably identifying and preventing content that depicts harm to children, such as CSAM. For product images that we get from merchants directly we use automated tools and human reviews to identify and block instances of CSAM. Due to the way in which we source user reviews (directly from merchants and from third-party aggregators), Google expects those third parties hosting the content to moderate it before it reaches our service; however, we still run our own protections for images and remove and report any CSAM we find on our service.

YouTube

Description of Service and Associated Risk Profile

YouTube's mission is to give everyone a voice and show them the world. We believe that everyone deserves to have a voice, and that the world is a better place when we listen, share, and build community through our stories. From music to education, from comedy to news, YouTube touches every corner of society, offering access to information in the video format to anyone with an internet connection. The internet is a force for creativity, learning, and access to information, and supporting the free flow of ideas has always been and remains at the heart of YouTube's video-first mission.

YouTube allows users to watch, upload, and share videos. YouTube is available to all EU users free of charge, and users can opt to pay for a premium subscription that removes paid ads and offers other features. Both are covered in this systemic risk assessment.

YouTube's focus on voice, stories, and community means that the service potentially impacts a wide range of rights afforded by the EU Charter, such as freedom of expression and information, media pluralism, freedom of the arts and sciences, freedom to conduct a business, as well as broad civic participation rights.

We strive to make YouTube as open as possible and empower users to easily access, create, and share information. In addition to providing a service for users to express their creativity and ideas, we are an important source of economic opportunity for creators, with whom we share revenue from ads that are served on their video content. Yet, as with all open internet services, there are inherent challenges and risks that arise which we must also address, including those from users that upload violent or dangerous content, sensitive and graphic content, and misinformation. Bad actors actively seek to exploit open services like YouTube for their own nefarious purposes, even as we continue to invest in robust systems designed to stop and deter them.

Over the years, we have worked tirelessly to develop policies and products that protect the YouTube community. As reflected in our Community Guidelines (policies broadly covering spam and deceptive practices, violent or dangerous content, misinformation, sensitive content, and regulated goods) and Legal Removals processes (procedures to ensure we comply with legitimate user and government requests to remove illegal content), YouTube is committed to keeping the service safe, for our users, advertisers, and society at large, while balancing open and free creative expression across the service. Beyond removing harmful content, we also leverage our recommendations systems and monetisation tools to promote a healthier ecosystem.

YouTube's prominent role as an online video-sharing service means that we naturally have a responsibility to protect the service from harmful content that may be uploaded, as well as other abuses of the service. YouTube's business model only works when our viewers, creators, and advertisers have confidence that we are living up to our responsibility as a business. In other words, responsibility is a business imperative: viewers do not want to see harmful content, advertisers do not want to be associated with it, and creators and YouTube depend on each other to attract users and advertisers alike.

Systemic Risk Assessment Results and Associated Observations

We assessed 39 different risk statements for inherent risks (i.e., risk absent any action taken by YouTube), preparedness (i.e., the cumulative measures currently in place to mitigate the risk), and residual risks (i.e., risk after mitigation by YouTube). Residual risk serves as a guide for where further investment may be warranted. The full list of risk statements is found in Annex A to this report.

At a high level, the sorts of risks addressed in this report can be divided into two sets: risks posed by the presence of a particular type of illegal or policy-violating content, and risks posed to users based on the design and functioning of a service. Thought of another way, some risks are mitigated by preventing, removing, or raising visibility of certain types of content (i.e., content moderation), and others are mitigated by changing the design or functioning of the service or the way users interact with the service (i.e., service design).

Important inherent risks identified in this assessment include risks associated with the presence of illegal or potentially harmful content, which we address via content moderation. Our investments enable us to achieve much lower levels of residual risk; however, given the complexity of balancing YouTube's mission of giving everyone a voice, while addressing harmful content, elevated levels of residual risk remain in relation to misinformation, disinformation, civic discourse, harassment and bullying, and public health.

Other notable inherent risks are associated with the design and functioning of a service, such as privacy, security, and child rights. Below we explain the service design choices which significantly lower residual risks related to the way YouTube functions, such as privacy and security measures and protections for minors on YouTube. However, while we have made significant investments in the safety of our younger users, such as preventing access to age-inappropriate content, we assessed elevated levels of residual risk related to problematic internet use given limitations of existing research into the existence or nature of a link between service use, the types of content being viewed, and potential overuse.

The structure of the below follows this division. We first address content moderation on YouTube, explaining YouTube's content policy development, enforcement, and the measures, like the Violative View Rate (VVR), which we use to gauge the efficacy of our moderation practices. The second section explains service design choices, which address risks related to the way YouTube functions, such as privacy risks or protections for children using YouTube.

Most of the systemic risks addressed in Article 34(1) of the DSA are related to fundamental rights, which are indivisible and interdependent. Because these rights (and associated risks) are interrelated, the practices YouTube employs to ensure users' rights frequently address more than one, or many, rights and risks articulated in Article 34(1) of the DSA. With this in mind, we have gathered together specific manifestations of systemic risks into groups that allow for efficient explanation of YouTube's existing mitigating practices, as well as improvements consistent with Article 35 of the DSA. We highlight where changes in the external context or improvements to our mitigation measures change inherent or residual risk estimations between the 2023 and 2024 systemic risk assessments.

2024 Highlights

- Introduced optional training modules for the Three-Strike System for Repeat Violators
- New measures to manage risks related to generative AI content including a new labelling tool for creators
- New privacy complaint tool to enable removal of content that looks or sounds like an individual altered or made using AI

Content Moderation

Removing Illegal Content

YouTube is one of the world's largest open video-sharing services. It is not surprising that bad actors work to upload illegal content on YouTube in violation of our express prohibitions (such as child sexual abuse material, terrorist and violent extremist content, hate speech, and non-consensual intimate images). This is why illegal content is one of our most critical inherent risks, and the reason we invest significantly to address the same—both alone and in collaboration with others, as described below. Similarly to our 2023 assessment, these investments and partnerships have resulted in low estimates for illegal content residual risks.

As discussed previously in this report, YouTube, and Google more broadly, have a robust process for evaluating government requests to remove content.[34] But in the absence of an order to remove content or a valid complaint from a rightsholder (as in the case of content infringing on intellectual property rights), YouTube enforces its Community Guidelines. There is extensive overlap between content prohibited by our Community Guidelines and content that is illegal, meaning that our enforcement efforts work to mitigate the risks of both illegal and policy-violative content.

Two Examples: Terrorist or Violent Extremist Content and Child Sexual Abuse Material (CSAM)

Terrorist or violent extremist content, and CSAM are examples of the overlap between illegal and policy-violative content. Enforcement efforts in both areas make use of signal sharing and hash matching (i.e., digital fingerprinting) to identify potentially violative content. Although CSAM is always illegal, the legal status of violent and extremist content varies widely according to context (based on the jurisdiction and the way the content is presented, as in the case of a documentary).

[34] For more information on how YouTube and Google respond to government requests to remove content, *see supra* at Handling Government Removal Requests.

Identifying and Removing Violent Extremist Content

Content that violates our policies against terrorist and violent extremist content includes material produced by designated terrorist organisations, content glorifying violent acts, and recruiting or fundraising on behalf of extremist groups, even if the content is not affiliated with a designated terrorist organisation. YouTube also prohibits violent or gory content intended to shock or disgust viewers, or content encouraging others to commit violent acts.

YouTube is committed to identifying and removing content that promotes terrorism or violent extremism on our service. Over the years, we have heavily invested in human review and machine learning technology that helps us quickly detect, review, and remove this content. Content that is removed is also used to improve our automated detection tools for better coverage in the future. In the rare cases users do see a video they believe is violative of our policies, we provide users with the option to flag, including for videos that "promote terrorism."

We're also a founding member of the Global Internet Forum to Counter Terrorism (GIFCT), where we work with other tech companies to keep terrorist and violent extremist content off the web and train and provide resources to smaller companies. In 2016 we created a hash-sharing database with industry partners where we share hashes (a type of "digital fingerprint") of terrorist content to inhibit its further spread. Today, this shared database is formally operated by GIFCT, which consists of 28 member companies (and growing), and the hash-sharing database contains hashes corresponding to more than 400,000 distinct images, videos, and textual items. This industry-wide collaboration helps address the systemic risk that illegal terrorist and violent extremist content spreads across services and supports smaller companies facing similar challenges. YouTube also uses these hashes for its own detection purposes and to test pertinent policies.

Whether violent extremist content is first detected by our own automated detection systems, by a GIFCT hash, or by a user flag, these moderation decisions are fed back into our machine learning technology to improve future detection.

Detecting, Removing, and Reporting CSAM

Similarly, we have heavily invested in engineering resources to detect CSAM in ways that are precise and effective, and have long used this technology to prevent the distribution of known CSAM videos on YouTube. This is an area where Google as a whole has been an industry leader, and this report previously addressed other company-wide efforts to combat the distribution of child sexual abuse material.[35] We have always had clear policies prohibiting content on YouTube that sexualises or exploits children. We use machine learning systems to proactively detect violations of these policies and have human reviewers around the world who quickly remove violations detected by our systems or flagged by users and our priority flaggers.

While some content featuring minors may not violate our policies, we recognise that the minors could be at risk of online or offline exploitation. This is why we take an aggressive approach when enforcing these policies, including for a feature like comments (i.e., a minor and ancillary feature pursuant to Recital 13 of the DSA). Our automated systems help to proactively identify videos that may put minors at risk and apply

[35] *See supra* Detecting, Removing, and Reporting CSAM.

our protections at scale, such as restricting live features, disabling comments, and limiting video recommendations for videos featuring minors.

Our proprietary CSAI Match technology,[36] which we licence to several other technology companies free of charge, allows us to detect known CSAM images and videos. In cases where a video contains CSAM or a user solicits CSAM through comments or other communications, our team reports it to the National Center for Missing and Exploited Children (NCMEC), who then liaise with global law enforcement agencies such as Interpol and Europol.

Once we have identified a video as illegal and reported it to NCMEC, the content is hashed (given a "digital fingerprint") and used to detect matching content. This hashing and scanning technology is highly precise at detecting known CSAM and enables us to detect illegal content more quickly. We maintain a database of known CSAM hashes and any content that is matched against this list is removed and reported to NCMEC.

Prohibiting and Detecting Infringement of Intellectual Property Rights

While losses due to copyright infringement can be serious, the risk of intellectual property (IP) infringement produced only elevated levels of inherent risk because of the relatively small number of users that are primarily affected, and because the related harms are not as difficult to remediate as, for example, serious physical harm or harms to vulnerable populations. Additionally, considering the suite of options protecting both institutional and individual rightsholders—such as the copyright webform, Copyright Match Tool, and Content ID, which are explained in more detail below–our assessment produced high preparedness ratings for our industry-leading protection tools.

All rightsholders have access to the YouTube copyright removal request webform, which is a streamlined and efficient way to submit copyright removal requests, and is available in 80 languages. It is designed for infrequent use by creators who hold few copyrights and rarely find their content on YouTube. For the vast majority of rightsholders, the webform is the only tool they need. Nevertheless, creators who have used the webform to remove videos from YouTube have access to powerful features, including the ability to ask YouTube to automatically prevent copies of the removed videos from being reuploaded.

For creators who experienced a higher amount of reposting of their copyrighted content and needed to submit more frequent copyright removal requests, we built the Copyright Match Tool to facilitate those creators' attempts to protect their intellectual property rights. The Copyright Match Tool is available to any YouTube user who has submitted a valid copyright removal request through the webform. Once a takedown request is approved, the Copyright Match Tool starts scanning YouTube uploads for potential matches to the videos reported in the removal request. The tool surfaces these potential matches to the claimant so they can decide what action to take next. For creators in the YouTube Partner Program, the tool automatically scans for potential matches on other channels, maintains a log of those matches for review, and through an easy-to-use interface allows the creator to archive the match, submit a takedown request, or contact the user. As of December 2023, over 3 million channels on YouTube have access to the Copyright Match Tool.

Efforts such as the Copyright Match Tool equip creators with the resources they need to protect their content, and are evidence of YouTube's high level of preparedness to prevent the non-authorised use of

[36] CSAI is child sexual abuse imagery and is a subset of content that can be considered CSAM.

copyright-protected materials. You can read YouTube's bi-annual Copyright Transparency Report for a description of the other means by which YouTube protects rightsholders.

Addressing Content that Violates our Policies

Our open service embraces a wide diversity of voices to entertain, teach, showcase talents, advocate, and build businesses. YouTube's commitment to free expression enables this diversity. But free expression on an open service can create tension with other fundamental rights, such as the right to security or the right to privacy. YouTube is available in over 80 languages, with billions of monthly active users worldwide (over 400 million in the EU). Because of YouTube's scale, even a relatively small number of bad actors can have systemic impacts on the service. YouTube acknowledges that its scale and extensive reach necessitate that we carefully balance the fundamental rights of users with any potential harms that may arise from misuses of our service.

Content that may be lawful but still creates harm (such as content impacting human dignity, promoting discriminatory beliefs, inciting, praising, or glorifying violence, promoting practices harmful to health, inciting gender-violence, or that constitutes harassment and bullying) can be uploaded to YouTube's open service, rendering it one of our most critical inherent risks. We have continued to develop thoughtful and comprehensive approaches for addressing this type of content (described below), but several factors make these risks more challenging to address than illegal content, including the need to take proportionate and reasonable measures that respect the right to freedom of expression and information, the need to consider the likelihood of real world harm, and the need to consider context to determine whether content actually violates policy. This results in some remaining elevated residual risk, such as in relation to content promoting practices harmful to health (including health misinformation), misinformation and disinformation related to civic discourse, and harassing and bullying content.

Below we describe some of YouTube's new and continuing efforts to identify and respond to all policy violative content. We then address some of the specific types of content violations we examined in this risk assessment, and discuss the mitigations for these specific risks.

Developing Policy

Preventing systemic risks related to content starts with YouTube's Community Guidelines. These "rules of the road" allow creative expression while prioritising the protection of the YouTube community from harmful content. As explained previously in this report,[37] YouTube's policy development process is robust, involving extensive internal analysis before implementation, regular reviews and updates, and engagement with internal and third-party experts to address issues before they reach, or become widespread, on our service. You can read more in How YouTube Works and in our blog post on policy development at YouTube.

It is not a coincidence that our Community Guidelines closely mirror many of the potential systemic risks addressed by Article 34 of the DSA. For years, YouTube has been attuned to these same risks. Our policies cover areas (such as hate speech, harassment, child safety, and violent extremism) across five broad categories: spam and deceptive practices, sensitive content, violent or dangerous content, regulated goods, and misinformation. Dozens of individual policies fall under these five categories, and our

[37] *See supra* at Designing Appropriate Content Policies.

assessment concluded that YouTube's policies provide excellent coverage of the risks identified both in Article 34(1) of the DSA and the Recitals elaborating on those systemic risks. Some of the most relevant Community Guidelines are explained below, such as those related to election misinformation or harassment and bullying. But one can find Community Guidelines that correlate to any of the DSA systemic risks. For example, YouTube's Community Guidelines related to sensitive content, violent and dangerous content, and regulated goods provide complete coverage of the illegal content related concerns detailed in Rectial 12 of the DSA.

The world moves quickly and our policies need to keep up. That's why we regularly review our policies to make sure that — similar to the laws that govern civil society — they reflect the changes that occur both on and off our platform.

Providing EDSA Exceptions

We recognise that some content that may otherwise violate our Community Guidelines but nonetheless provides compelling educational, documentary, scientific, or artistic value should remain available for viewers. We call this the "EDSA" (Educational, Documentary, Scientific, or Artistic) exception, and it is a critical way to make sure that important speech stays on YouTube, while protecting the wider YouTube ecosystem from harmful content. To educate creators, we include information about EDSA in our Help Center. To help determine whether a video might qualify for an EDSA exception, we look at multiple factors, including the video title, descriptions and the context provided in the video's audio or imagery, as well as the public interest of the content. These decisions are nuanced and context is important. Examples include hate speech that is condemned in a documentary about war, content targeting minors with insults that might appear as part of an educational anti-bullying campaign, or nudity that has scientific value or constitutes artistic expression.

Enforcing Policy

In addition to developing robust policies, we use a wide range of tools to enforce these policies. By combining multiple methods and approaches, YouTube continually improves the service for our viewers and creators.

Undertaking Automated Detection and Removal

Automated detection of problematic content enables YouTube to enforce our Community Guidelines at scale. Sophisticated automated systems are our primary tool. In Q2 2024, over 95% of videos (compared to 93% in Q1 2023) and over 99% of comments that were removed were first detected by automated means. And automated detection isn't just identifying huge amounts of problematic content; it's doing so quickly, before the impact is widespread. For example, in Q2 2024, over 80% of the videos we removed had ten views or fewer (compared to 72% in Q1 2023).

Once models are trained to identify potentially violative content, YouTube either automatically removes violative content (when the precision of our enforcement is very high) or the model nominates the content for human review against our Community Guidelines. Content moderators then confirm or deny whether the content should be removed.

This collaborative approach helps improve the accuracy of our models over time, as models continuously learn and adapt based on content moderator feedback. And, by maintaining complete coverage of all EEA languages deployed by YouTube, our content moderators can also bring native understanding to both the review process and our model training.

Maintaining a Priority Flagger Program

While we facilitate and encourage flags by users, generic user flags typically have low actionability rates. Our Priority Flagger program (formerly called Trusted Flaggers) complements our automated systems and helps spot potentially problematic content. We developed the YouTube Priority Flagger program to streamline the reporting processes for government agencies and non-governmental organisations (NGOs) that are particularly effective at notifying YouTube of content that likely violates our Community Guidelines–though each flag is reviewed by a human to assess whether it is a violation or not. The program provides these partners with dedicated reporting processes and a channel for ongoing discussion and feedback about YouTube's approach to various content areas. The program is part of a network of more than 300 government partners and NGOs that bring valuable expertise to our enforcement systems. Participants in the Priority Flagger program receive training in enforcing YouTube's Community Guidelines, and because their flags have a significantly higher action rate than the average user, we prioritise them for review. However the scale of YouTube and the effectiveness of automated systems meant that in Q2 2024 Priority Flaggers accounted for only 0.7% of videos removed from the service.

Enforcing a Three-Strike System for Repeat Violators

YouTube recognises creators' significant investments in their video content. For that reason, we already provide creators with proportionate due process when we think it is necessary to take enforcement action against a creator or their content. We have consistent penalties for violating our policies, exemplified by our three-strike system. Generally, after one Community Guidelines violation, the user gets a warning, but with subsequent violations the user begins to accrue strikes. Strikes carry increasing penalties when a channel receives them within a 90-day period:

- 1st strike - 1 week suspension;
- 2nd strike - 2 week suspension; and
- 3rd strike - channel termination

Consistent with our commitment to championing users' fundamental rights, including freedom of expression and transparency, we recently updated our strikes policy to incorporate optional trainings. The optional trainings are short in-product educational experiences based on the specific Community Guidelines policy that has been violated. If a user receives a Community Guidelines warning, the user can access the policy training from their Studio account. If the user completes an optional policy training, their warning will expire after 90 days. If a user violates a different policy after completing the training, the user will receive another warning. Repeated violations of our policies–or a single case of severe abuse–may still result in the termination of a user's account. This improvement is directed at avoiding the need to restrict user content by empowering users to ensure they comply with YouTube's Community Guidelines in the future.

We developed our three-strikes policy to balance terminating bad actors who repeatedly violate our Community Guidelines with the need to make sure people have an opportunity to learn our policies and appeal decisions. At the same time, we work hard to make these policies as understandable and transparent as possible, and we enforce them consistently across YouTube. We do not hesitate to issue strikes and terminate channels whose content repeatedly violate our policies, irrespective of whether the channel has a large audience.

While legitimate users get three strikes, we directly terminate egregious offenders such as uploaders of CSAM or channels dedicated to posting spam. In Q2 2024, we terminated more than 2.5M channels for spam.

Strikes, terminations, and content removals are only a few pieces of a larger puzzle. These complex problems necessitate multifaceted solutions, and dealing with material such as misinformation or potentially sensitive content on YouTube is no exception. While our Violative View Rate (described below) shows that YouTube has made strides in removing clearly violative material, more nuanced harms are not solely addressed by removals under our Community Guidelines. These areas of harmful content, which often brush up against our policy lines, require a comprehensive approach that includes raising authoritative content and rewarding creators who meet the higher bar required for our partner program.

Dealing with Generative AI content

The recent increase in the accessibility of generative AI tools has reduced the barriers to creating synthetic content, which may be shared on YouTube. Generative AI has the potential to unlock creativity on YouTube and transform the experience for viewers and creators on our platform. But just as important, these opportunities must be balanced with our responsibility to protect the YouTube community. All content uploaded to YouTube is subject to our Community Guidelines—regardless of how it's generated—but we also know that AI will introduce new risks and will require new approaches.

We believe it's in everyone's interest to maintain a healthy ecosystem of information on YouTube. We require creators to disclose synthetic content when it's realistic, meaning that a viewer could easily mistake what's being shown with a real person, place or event and the alteration is not inconsequential.

We provide creators with a tool in Creator Studio to label their videos as synthetic prior to sharing them. When creators fail to apply the label and YouTube is able to confirm that the content is synthetic or manipulated and seems realistic, YouTube may apply a label that creators will not have the option to remove. We have continued to iterate on our policies relating to disclosing the use of synthetic or altered content to help users better understand when they should disclose their use of content editing and generation tools, and how to do so. We're still in the early stages of this work, and will continue to evolve our approach as we learn more.

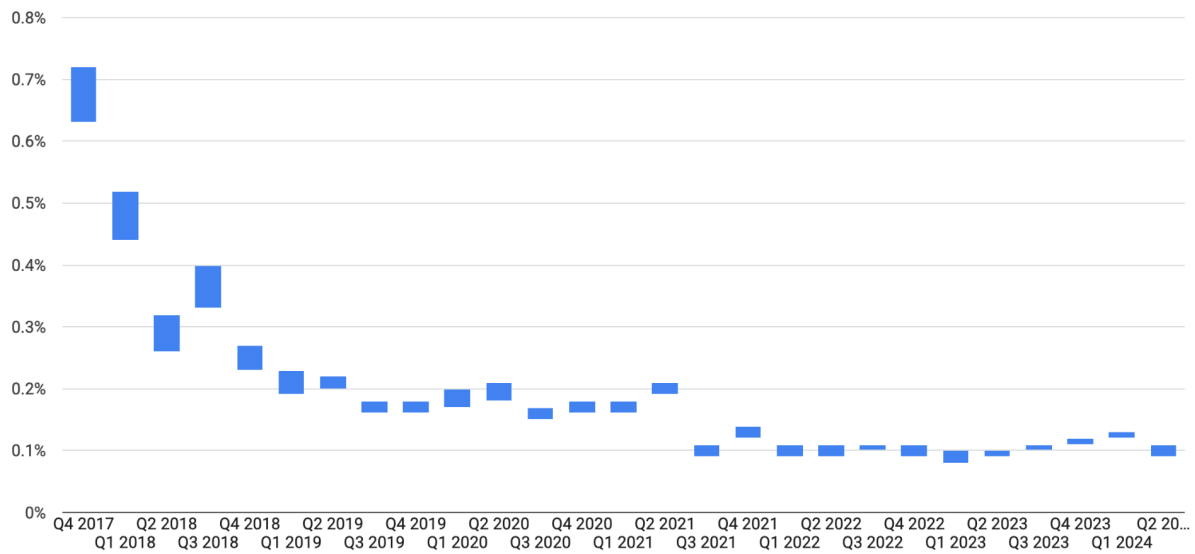
Recently, we developed a Privacy Complaint Process in relation to AI-generated or other synthetic content that looks or sounds like an individual. If someone has used AI to alter or create content that looks or sounds like an individual, the individual can ask for it to be removed. In order to qualify for removal, the content must depict a realistic altered or synthetic version of the person's likeness. This process was available in Europe prior to the 2024 European Parliamentary Elections in June 2024, and is now available in all regions.

Measuring Success: Violative View Rate

As described above, automated detection tools allow for the quick detection of problematic content. YouTube strives to remove content that violates our Community Guidelines before users are exposed to it. In Q2 2024, we removed 59.9%% of violative videos before they had a single view, and 24.3% of videos when they had one to ten views.

To measure our progress on removing violative videos before they are viewed, we developed a metric called Violative View Rate (VVR), which has been publicly reported since 2021. This metric, updated and made publicly available quarterly, estimates the percentage of total views on YouTube that are of violative videos (i.e., videos that are inconsistent with our Community Guidelines).

VVR data gives critical insight into how well we are protecting our community. Although metrics like the turnaround time to remove a violative video or the number of takedowns are important, those statistics do not fully capture the actual impact of violative content on viewers. The VVR is a better measure because it tells us how widely violative videos have been viewed before they are taken down. Two videos could be removed from YouTube within 24 hours, but one may have 100 views while the other has 1 million views. This is a 100% takedown rate within 24 hours, but that metric obscures the most important information. Because we care most about the potential for harm to users, and potential harm can arise by actual exposure to violative content, we have chosen to focus attention on a metric that specifically measures user exposure. We believe the VVR is the best way for us to understand the extent to which harmful content may reach viewers, and to identify where we need to make improvements. We are committed to being transparent about this metric and working to continue to reduce it over time, as we have since 2017.



Graphical depiction of the YouTube Violative View Rate (VVR)

Calculating VVR serves a second purpose: it helps us gain insight into the type of content we should remove but sometimes miss. Our methodology for calculating the metric allows us to do this. We calculate

VVR by taking a sample of videos on YouTube and having content reviewers gauge which videos violate our policies and which do not. By sampling, we gain a more comprehensive view of the violative content that evades our detection and enforcement systems. With that understanding we can improve those systems and, over time, further decrease the VVR.

Over the years, we have seen the VVR fluctuate—both up and down. For example, immediately after we update a policy, this number may temporarily rise as our systems ramp up to catch content that is newly classified as violative. Our methodology for this reporting mechanism has been validated by MIT Sloan professor of statistics Dr. Arnold Barnett as "thoroughly sensible and statistically sound."[38]

Our VVR reports indicate that violative views today are around 0.1% of all videos viewed (i.e., out of every 1,000 views on YouTube, just one is of violative content). We recognise that even if the prevalence of violative content is low, it might still represent a large volume of content in absolute terms, and significant investments are required to maintain these low levels. This report describes the efforts we undertake to prevent users from seeing violative content, and identifies potential areas for improvement.

Elevating Authoritative Sources

Removal of violative content is not the only way that YouTube makes adjustments to balance the freedom of expression with other rights such as safety and security. Over the past several years we have invested significantly in the systems that take authoritativeness of the channel into account when making recommendations where accuracy and authoritativeness are key, including news, politics, and medical information. Our systems are trained to elevate authoritative sources in search results, particularly in sensitive contexts. We raise high-quality information from authoritative sources in search results and information panels, in turn helping people find accurate and useful information. Whether you are searching for something evergreen or a current event, YouTube aims to surface videos from sources like public health authorities, research institutions, and news outlets, within the top search results. These efforts are particularly important when it comes to connecting people with information from high-quality sources at the moments that matter most—for example when learning about a breaking news event, or searching for health information. So when a user in France searches for, "arreter de fumer," the top listed videos are from top hospitals, cancer centres, and Sante Publique France, the national public health authority.

Providing Information Panels with Topical Context

For YouTube search queries or videos related to topics prone to misinformation, such as COVID-19 and climate change, we surface information panels which provide content sourced from independent third parties. These panels give viewers additional context and help them make more informed decisions about what they are watching. Depending on the topic, the panels will point to information from sources like health authorities, Wikipedia, Encyclopedia Britannica, and the United Nations. These information panels will show regardless of what opinions or perspectives are expressed in a video. YouTube also uses information panels to inform users when content has been uploaded by a news outlet funded in whole or in part by a government.

[38] Arnold Barnett (2021) YouTube's Violative View Rate Methodology, Massachusetts Institute of Technology.

Addressing Specific Content Risks

Addressing Misinformation and Disinformation

Risks related to intentional manipulation of the service (e.g., dis- or misinformation impacting civic discourse or content promoting practices harmful to health) are complex, constantly evolving, and societal-wide issues. They necessitate a multifaceted approach, combining policy enforcement and content evaluation with real-time context and information for users. YouTube has many measures in place (described below) but must still contend with determined and highly-motivated bad actors constantly evolving their techniques, resulting in some elevated levels of residual risk.

YouTube has developed measures to respond to situations where the risk of misinformation is at its greatest. Where misinformation violates our policies we are quick to remove such content, and we will terminate a channel for egregious or repeated offences. But we employ other techniques to combat misinformation in addition to simply taking down content or channels. We have a higher bar for monetised content, so that creators are not incentivised to create untrustworthy clickbait or other low-quality, misleading videos.

YouTube's measures directed to mitigating the risk of polarisation and segmentation of perspectives in news media similarly help combat these risks. For example, YouTube provides a News Watch Page, which helps viewers explore different sides of a news story, including mechanisms to help ensure diversity of news sources. In the News Watch Page, users can see: updates, with the most recent video coverage of the news story; explanations and commentary, with additional context on the news topic; live News, with live streams showing what's happening in the moment; and shorts, to quickly catch up on the news stories' latest updates. Users can open the watch page for a specific news story by clicking on a video with the newspaper icon on the YouTube homepage, news destination page, or in search results.

Addressing Public Health Related Violative Content

In our assessment, we identified critical inherent risks in relation to content promoting practices harmful to health (e.g., self-harm, anorexia, health misinformation) and assessed our preparedness at "effective" rather than "optimal," in part due to the viral nature of this content.

Our methods for addressing public health-related misinformation exemplify the multifaceted approach described above. We raise information from authoritative health sources, as determined by external experts and externally published principles, and provide context on the sources of health information for users via information panels below each of those videos.

Crisis resource panels are an important part of the suite of health products on YouTube. These are information panels that help easily connect users with authoritative and helpful information in times of crises. YouTube's crisis resource panels allow users to connect with live support from recognised crisis service partners. The panels may surface on the Watch page, or in YouTube search results. Currently, topics covered include suicide, self-harm, eating disorders, and topics related to certain health crises or emotional distress.

Another recent policy change with inherent challenges and trade-offs concerns medical misinformation, with scientific understanding evolving all the time and important topics (such as vaccines) being a source of fierce debate, notwithstanding consistent guidance from health authorities about their effectiveness. Our Community Guidelines already prohibited certain types of medical misinformation, but we worked with experts to expand them. We've also taken what we've learned so far about the most effective ways to tackle medical misinformation to simplify our approach for creators, viewers, and partners. In particular, we have streamlined dozens of our existing medical misinformation guidelines to fall under three categories – Prevention, Treatment, and Denial. These policies apply to specific health conditions, treatments, and substances where content contradicts local health authorities or the World Health Organization (WHO). To determine if a condition, treatment, or substance is in scope of our medical misinformation policies, we evaluate whether it's associated with a high public health risk, publicly available guidance from health authorities around the world, and whether it's generally prone to misinformation.

Addressing Civic-Discourse-Related Violative Content

Article 34(1) of the DSA directs YouTube to conduct an assessment of systemic risks to civic discourse and electoral processes. Applying our risk assessment methodology, YouTube evaluated the inherent risk of civics misinformation to be higher, leaving elevated levels of residual risk despite our preparedness, largely due to external factors including the dynamic and viral nature of misinformation in politics, during elections, and at times of crisis and civic unrest.

With users around the world coming to YouTube to learn about politics and develop informed opinions about current events, we have a responsibility to support an informed citizenry and foster healthy political discourse. We provide a range of resources for civics partners such as government officials, candidates, civics organisations, and political creators to ensure a broad range of voices are heard.

Among other items, our Community Guidelines prohibit content that has been technically manipulated or doctored in a way that misleads users and may pose a serious risk of egregious harm, content that aims to mislead people about voting processes, and content encouraging others to interfere with democratic processes, such as obstructing or interrupting voting procedures. Other policies that are relevant during elections include:

- **Voter suppression:** Content aiming to mislead voters about the time, place, means, or eligibility requirements for voting, or false claims that could materially discourage voting.
- **Candidate eligibility:** Content that advances false claims related to the technical eligibility requirements for current political candidates and sitting elected government officials to serve in office. Eligibility requirements considered are based on applicable national law, and include age, citizenship, or vital status.
- **Incitement to interfere with democratic processes:** Content encouraging others to interfere with democratic processes. This includes obstructing or interrupting voting procedures.
- **Hate speech:** Content that promotes violence or hatred against individuals or groups based on certain attributes. This includes, for example, content that shows a political rally attendee dehumanising a group based on a protected attribute, such as race, religion, or sexual orientation.

- **Impersonation:** Content intended to impersonate a person or channel, such as a political candidate or their political party.
- **Misinformation:** Certain kinds of misleading or deceptive content with serious risk of egregious harm, including content that may pose a serious risk of egregious harm by falsely claiming that old footage from a past event is from a current event.
- **Harassment & cyberbullying:** Content that threatens individuals, including content that threatens individuals such as election workers, candidates, or voters.

In addition to our robust policies about what is not allowed on YouTube, we also devote significant resources to systems that raise the visibility of authoritative content, as described above.[39] These techniques are designed to ensure that users find the trustworthy content they are looking for on topics that can be targets for manipulation by bad actors.

Detecting and Removing Harassment and Bullying in YouTube Comments

Experience also shows that comments on YouTube are sometimes misused to directly and indirectly threaten the wellbeing of creators and other users particularly at risk of being the targets of harassment and abuse. In 2024, we recognised increased inherent risk in our assessment of risks related to bullying and harassment. This was due to a change in risk statements aimed at achieving a more precise assessment of human dignity-related risks. YouTube's automated moderation systems are specifically and proportionately designed to mitigate these risks to user and creator safety. With these protections, in Q2 2024 we removed about 94 million comments for violating our Community Guidelines prohibiting harassment and cyberbullying. YouTube's automated detection systems are removing this content at scale, but the overall number of bullying and harassing comments produced one of the higher inherent risk ratings in our assessment, so we believe there is more work to be done to protect our users.

For a minor and ancillary feature like YouTube comments, which accounts for a very small percentage of the time users spend on the YouTube service (less than 1% of time as of our Year 1 Report), much of the violative content is directed at creators (i.e., users that upload videos), who are the heart of YouTube. Viewed another way and as referenced in our Year 1 Report, users both globally and in the EU spent over 120x more time watching videos than they did engaging with comments in Q4 2022. Consistent with Recital 40, YouTube has developed appropriate and proportionate strategies for detecting and removing offensive content in comments aimed at creators, including those creators particularly at risk of being subject to hate speech, sexual harassment, or other discriminatory actions. YouTube deliberately casts a wide net, using automated technologies optimised to identify and remove any comments appearing under videos directed at creators at particular risk of being subject to harassment, discriminatory actions, or bullying.

In Q2 2024, over 80% of actioned comments were removed because they were spam (i.e., deceptive, high-volume commercial content that harms the user experience). Of the remaining 19%, comments were removed for other important user safety reasons, such as harassment and cyberbullying (6%), child safety (5%), and violent or graphic content (4%).

[39] *See supra* at Elevating Authoritative Sources.

While our size influences our risk profile, so does the format of the content with which users engage across our service. Users come to YouTube to create, share, and view audiovisual content, i.e., videos. Comments are a secondary feature, which creators can opt to enable and permit users to contribute additional textual feedback under creator videos. Our Community Guidelines apply to all content on the service, regardless of its format. But when it comes to how those policies are enforced and the corresponding consequences for users who post secondary text content in comments, there are critical differences as compared to video content. In other words, comments occupy a fundamentally different place in YouTube's video-first ecosystem.

Moreover, a user's investment in commenting on a video does not compare to a creator's investment in terms of time, effort, and resources in creating the video content available on YouTube. Creators often take many steps to create a YouTube video: research, scripting, filming, editing, audio-mixing, thumbnail creation, and search engine optimisation. They use multiple devices and software applications and can spend many hours of production work creating a single video, costing time and money. By contrast, commenting requires very little effort: a few keystrokes amounting to much less time and effort than video creation.

There are other aspects of YouTube comments that make them unlike videos:

- Creators have control over whether comments on their videos are enabled or not, can remove any comments under their videos for any reason, can edit comments under their videos, can create block lists for words or phrases permissible in comments under their videos, and can block specific users from commenting on their videos.
- Comments are not searchable, recommended,[40] nor accessible via the YouTube Homepage.
- Comments are not a factor in a creator's ability to monetise their video content.
- Comments are intrinsically tied to the video to which they relate, and not independent pieces of hosted content. If a video is removed or taken down, the comments associated with it are automatically taken down. The same is not true when a comment is moderated, which has no impact on the availability of the video on the service.
- Comments are not enabled on all versions or interfaces of YouTube. Additionally, comments are not available on certain types of videos featuring minors, on YouTube Kids, or for embedded videos.

Given the above considerations, the moderation of comments on YouTube does not pose the same risks to freedom of expression or information present in video moderation.

Prohibiting and Removing Hate Speech

Our assessment found lower levels of residual risk of hate speech. YouTube's hate speech policy outlines clear guidelines prohibiting content that promotes violence or hatred against individuals or groups based on certain attributes. We enforce this policy rigorously and regularly report on the removal of hateful

[40] Comments are not recommended by an algorithm to increase engagement, but they can be sorted chronologically or by most engagement depending on settings.

content from our service. For example, in Q2 2024, we removed over 163,000 videos for violating our hate speech policies.

We have made significant progress in our work to quickly remove hateful content from our service. In 2019, we updated our hate speech policy, resulting in an increase of the number of daily hate speech comment removals. In Q2 2024, we removed over 28 million comments for violating our hate speech policies.

A 2020 report by the Institute of Strategic Dialogue showcased the efficacy of our hate speech policy update: "Following YouTube's change of hate speech policies we found a significant reduction of such content on the platform... an analysis of the volume of these mentions over time reveals a dramatic drop in content around spring 2019, demonstrating the effectiveness of YouTube's ban on Holocaust denial content."[41]

Additionally, all our policies, including our hate and harassment policies, include penalties for creators who repeatedly brush up against the line, including removal from the YouTube Partner Program.

As outlined on our Help Center page, under YouTube's hate speech policy, we may remove content or issue other penalties–such as terminating an account–when a creator repeatedly targets, insults and abuses a group based on attributes such as race, ethnicity, sexual orientation, or gender identity and expression, across multiple uploads.

Additionally, YouTube is a founding signatory to the EU Code of Conduct on Countering Illegal Hate Speech. Each year, YouTube takes part in the annual monitoring exercise, responding to flags from NGOs specialising in hate speech.

Service Design

Respecting Privacy

YouTube's main source of revenue is advertising—a portion of which is shared with creators participating in the YouTube Partner Program—and we use the information we collect for the purposes described in our Privacy Policy, including to provide the service, customise services, provide recommendations, personalise search results, and serve relevant ads. We also take our responsibility to protect user information seriously, and while advertising makes YouTube free of charge for everyone, we do not sell personal information to anyone. YouTube's data practices turn a significant inherent risk into a much lower residual risk.

Our Privacy Policy and YouTube's Help Center page on privacy provide transparency over what information we collect, why we collect it, how we process it, and how users can manage their information. Your data in YouTube is a powerful, easy-to-use tool designed to give users control over the privacy settings that are right for them, and provides further information on the data we collect and use across our services.

If a user turns YouTube watch history off and has no significant prior watch history, features that require watch history to provide video recommendations will be disabled.

[41] Jakob Guhl, Jacob Davey (2020), Hosting the 'Holohoax', Institute for Strategic Dialogue.

We are likewise taking steps to mitigate the risk that personal information is used or disclosed inappropriately. Recently we have implemented two specific mitigations under Art. 35. First, we enhanced our access controls relating to personal information. Second, we updated the automated systems we use to flag phishing ads and accounts at their source, learning from the latest methods adversarial actors use to circumvent systems. While systems are constantly improving, attackers swiftly shift tactics in an attempt to game the systems.

Protecting Children's Rights

In this assessment and consistent with our obligations under DSA Article 28 regarding "Online Protection of Minors", YouTube considered numerous risks particular to children. As described below, these include the risk that children access or are exposed to content they should not see, or conversely that their access to content is overly restricted; the risk that YouTube stimulates behavioural addictions in children; and the risk that children's data are used to target ads.

YouTube is heavily invested in the safety of its younger users. As described above, we have well-developed and advanced tools to quickly detect and remove illegal content. We collaborate with industry partners, and make available first-in-class tools to allow other services to remove illegal content at scale as we do. Our policies provide additional protection, under which we remove harmful but legal content. Below we describe the policies and protections that go beyond content removal, and ensure that our service is designed in a way that is aimed at keeping children safe.

YouTube pursues many policies and programs to protect children on the service, and we seek the input of experts to shape those efforts. Our Youth and Families Advisory Committee is made up of experts in children's media, child development, digital learning, and citizenship from a range of academic, non-profit and clinical backgrounds, and provides advice when we update our family product experiences and policies. Other components include rules and guidelines for when children appear in content, restricting access to mature content, and protecting minors at risk. You can read more in Fostering Child Safety.

Maintaining Guardrails for Children's Access to Content

We assessed the risk that children under a defined minimum age access YouTube services that they should not be able to or are exposed to harmful, hateful, or age-inappropriate content. We implement a wide range of measures (such as minimum age requirements, signals for estimating the age of users, "made for kids" content, parental controls, and granular age categories in YouTube Kids) to address the risk of children accessing age-inappropriate content. Although this risk will never be eliminated, our measures result in a significantly lower residual risk profile. The mitigations described below also resulted in a much lower residual risk of children's access to content being over or under restricted.

We are always looking at ways to create an appropriate environment for family content on YouTube, so we invest heavily in the policies, technology, and teams that help provide families with the best protection possible. Our holistic child rights approach has several important components.

We age-restrict content that does not violate our policies, but is nonetheless inappropriate for viewers under 18. This includes videos containing adults participating in dangerous activities that children may

imitate or videos related to regulated substances, sexually suggestive content, or violent and vulgar content. Videos that are age-restricted are not viewable by signed-out users either.

YouTube Kids is a separate app built from the ground up to be a safer and simpler experience for kids to explore, with tools for parents and caregivers to guide their journey. The app is a filtered version of YouTube and has a much smaller set of content available than YouTube's main app and website. This is because we work to identify content that is age-appropriate, adheres to our quality principles, and is diverse enough to meet the varied interests of kids globally.

Supervised experience on YouTube is for parents who decide their tween or teen is ready to access YouTube through a supervised Google Account. Videos a child can watch depend on the content setting their parent selects when setting up a supervised experience. We have disabled a number of standard features normally available in YouTube, like comments, uploads, purchases, and live chat. To reinforce healthy screen time habits, reminders for breaks and bedtime are set to "on" by default.

When we find that a channel is owned by a user under 13 and that user is unsupervised, we terminate that account. YouTube employs automated detection systems to find signals on YouTube channels that indicate that the channel may be owned by a user under the age of 13. These systems rely on content signals to find such channels, which are then flagged for a team to review more closely when they appear owned by an underage user. Channels identified as potentially owned by underage users are sent through the account recovery process and are given two weeks to provide evidence that the owners are 13 years or older or to obtain parental consent and establish parental supervision. Channels flagged in this way are disabled and thereafter deleted if the owner does not prove they are 13 years of age or older or establish parental supervision within two weeks of the initial notice.

YouTube also employs automated detection to determine whether young minors are livestreaming themselves without supervision. These accounts are further reviewed by a team to determine whether to disable the account.

Addressing Potentially Addictive Behaviour in Children

We also assessed the risk that the interface, design, or features of YouTube stimulate behavioural addictions in children using the service. Google has many measures in place to address this risk (such as parental controls, the unique experiences designed for kids described above, and surfacing high-quality content), but the general lack and limitations of existing research into the existence or nature of a link between service use (e.g., screen time), the types of content being viewed, and addiction results in some elevated levels of residual risk.

Despite inconclusive research, we take steps to manage the potential risk of excessive use of our service by children. For users that declare themselves to be under 18 when they create their Google Account, Take A Break and Bedtime reminders are turned "on" by default. These are aimed at reinforcing healthy screen time habits. Our built-in timer also lets parents limit screen time by telling kids when it's time to stop watching.

Additional protections apply for YouTube Kids and YouTube Supervised Experience. YouTube Kids allows parents to set the amount of time their child can spend on the service. Additionally, Family Link accounts

allow parents to control the time children spend with their device or with specific apps, including YouTube Kids and YouTube Supervised Experience.

Protecting Children's Data

Personalised ads are prohibited on YouTube Kids, as well as for users in a supervised experience on YouTube, consistent with our obligations under Article 28(2) of the DSA. For videos "made for mids", meaning children are the primary audience or the video is still directed to children based on factors such as the subject matter of the video, whether the video has an emphasis on kids characters, themes, toys or games, we limit data collection and use, and as a result, we restrict or disable some service features. For example, we do not serve personalised ads on content "made for kids", and some features are not available on these videos, like comments and notifications. All creators are required to indicate whether or not their content is "made for kids". Accordingly, our assessment resulted in lower residual risk of children's data being used to target ads.

Protecting Children's Safety in YouTube Comments

For years, YouTube has been attuned to the pernicious threat of predatory conduct towards children in comments. To combat this threat, we have continually refined our automated detection systems to remove potentially predatory comments. It is critical that automated detection measures related to child safety be designed to cast a wide net because they must detect comments that are often facially innocuous. In egregious cases, we terminate the account and report the content to the National Center for Missing and Exploited Children (NCMEC), an organisation that works with global law enforcement agencies to protect children. We have developed and launched increasingly effective automated detection measures in order to ensure that YouTube remains a safe space. For a minor and ancillary feature like comments specifically, these automated systems are deliberately designed to cast a wide net so as to identify and remove as much material in comments as possible that may potentially be harmful to children (e.g., sexualisation of minors, information regarding minors, CSAM). We also use machine learning in developing our systems to identify hundreds of millions of non-violative videos depicting children and automatically turn off comments to avoid any chance of the child being the subject of harassing or predatory comments. We choose to err on the side of safety to protect this vulnerable population from exploitation, and continuously work to refine our automated approaches for identifying and removing any content in comments that potentially threaten the safety of children.

YouTube commits resources and has developed reasonable and proportionate enforcement strategies optimised to detect and remove as many comments threatening the safety of children as possible, consistent with the express obligations under DSA Article 28. As noted above, this content is the most challenging for any service hosting user-generated content. Much of this content appears innocuous to many viewers but may still be used in ways that YouTube wants to prevent (for example, by individuals seeking sexual gratification). Because the content may be posted innocently and omit objectively problematic content, the challenges of addressing potential misuse are significant. YouTube has developed machine-learning tools and content policies to identify this and similar types of content that may appear innocuous or humorous, but may put minors in potentially risky situations.

As an example, in early 2019, YouTube learned that some innocuous videos (such as a home video of a young girl jumping into a pool in a swimsuit) could potentially appeal to bad actors. These videos do not sexualise or endanger minors, and thus do not violate YouTube's content policies. However, bad actors could present some risk of engagement by viewing or commenting on the video. Accordingly, YouTube developed a comprehensive approach to address these issues: combining machine-learning tools and content policies to remove violative comments and apply restrictive measures to the discoverability of this type of content.

Where an egregious violation occurs, we terminate the commenter's account. The poster of an egregious comment removed for child safety reasons receives a notice of the termination and is given the right to appeal. In the unlikely event termination was the result of a false positive, the account is reinstated and the poster's fundamental rights are protected. The vast majority of removed comments, however, do not result in the suspension or termination of a user's account, and the uploader of an actioned comment remains free to use the service, similarly protecting the users' fundamental expression rights.

We believe our tailored approaches to moderating comments and videos strike the appropriate and proportionate balance. Data, comparatively minor user engagement, and the structure of our service place comments in an ancillary position to videos. But comments pose an outsized risk of harm to creators and young users. We choose to err on the side of safety above other considerations in this narrow context because the weight of the competing interests clearly demands it.

Promoting Equity

In 2020, we established a dedicated Racial Justice, Equity, and Product inclusion team to explore practices, policies, and norms that could reproduce bias and inequity on YouTube.

One of the core programs under the Racial Justice, Equity, and Product inclusion team is YouTube's Inclusion Working Group (IWG). This group works to institutionalise inclusion and equity across YouTube's products, content policies, and business–prioritising equity considerations prior to product launch. Members include executive sponsors, a dedicated product inclusion lead, and representatives from employee resource groups across YouTube. Since its inception in 2020, the IWG has partnered in over 1,400 projects to understand and consider equity and inclusion early in the development process. The IWG's work has improved how we detect racially hateful comments and prepared teams to identify and respond to new forms of online hate.

YouTube and Google more broadly also prioritise equality of access to information, including for those with disabilities. Google and YouTube have a well established process for measuring accessibility. Google Accessibility Rating (GAR) is a scale measuring a software product's level of technical accessibility. It's used to evaluate products and features during the product development lifecycle. GAR is a rating system that ranges from zero (least accessibility support) to four (high accessibility support), allowing teams to monitor the level of technical accessibility over time. To that end, Google also provides resources for partners such as creators and developers to use Google products and leverage research to increase accessibility.

5. Conclusions

Our mission is to organise the world's information and make it universally accessible and useful. But for information to be helpful, it must also be reliable. That's why we take our responsibility seriously to protect users from harm, deliver reliable information, and partner to create a safer internet.

We have long developed and implemented methodologies to assess our services prior to launch and throughout their use. These methodologies are based upon well-established global approaches to the identification, prioritisation, and mitigation of risk, and have been tailored to the specific challenges of the technology industry.

Our risk assessment and mitigation priorities for the coming year are those we set out in this report. These include continuing to address the evolving risks and opportunities associated with the wide availability of generative AI tools, maintaining our participation in collaborative approaches that pursue "whole of society" strategies for addressing systemic risk, and implementing strategies that are responsive to the changing context and the evolving tactics of bad actors. We will continue to take an approach that is informed by engagement with external stakeholders and responsive to insights arising from the latest research.

We are committed to continuous improvement in our approach and submit our second EU DSA systemic risk assessment report in this spirit. We welcome the opportunity to receive input and feedback on our assessments from the European Commission and other stakeholders, and build upon our work to date in subsequent reports.

Annex A: Full List of Risk Statements

While the 2024 risk statements are substantially the same as 2023, some improvements were made to reduce duplication, ensure completeness, and respond to feedback provided by the auditor and European Commission.

Illegal Content, Behaviour, and Products and Services

- Risk that content or behaviour that constitutes or facilitates Child Sexual Exploitation and Abuse is available or takes place on a service
- Risk that illegal content or behaviour representing, praising, glorifying, assisting, facilitating, or supporting (including through fundraising) terrorist organisations or acts of terrorism or violent extremism is available or takes place on a service
- Risk that illegal hate speech is available on a service
- Risk that content or behaviour constituting illegal harassment or bullying (including doxxing, stalking, and threats of violence) is available or takes place on a service
- Risk that intellectual property is available on a service in ways that violate or facilitate the violation of legal protections
- Risk that a service is used for or facilitates illegal online behaviour
- Risk that the promotion or sale of illegal products and services takes place or is facilitated on a service

Freedom of Expression and Media Pluralism

- Risk that a service removes content that does not constitute a necessary or proportionate removal of content with a legitimate purpose
- Risk that users are not able to report potentially violating content on a service
- Risk that users are not able to appeal content removals on a service
- Risk that the users' ability to make autonomous and informed decisions about what they view on a service is impaired by limited transparency or options
- Risk that the visibility of content on a service adversely impacts media pluralism

Privacy and Data Protection

- Risk that a service collects, processes, aggregates, and / or shares more user information than is necessary for the stated purpose or without the informed consent of users
- Risk that private or highly personal information of users is unintentionally made available on a service
- Risk that private or highly personal information about users or others is maliciously made available on a service
- Risk that sensitive personal data is used to target paid speech at users of a service without the informed consent of the user
- Risk that content or applications enabling or facilitating phishing, malware, data breaches or other digital threats is available on a service

Human Dignity

- Risk that content praising, supporting, promoting, inciting, or that constitutes vulgarity or profanity is available on a service
- Risk that content praising, supporting, promoting, inciting, or that constitutes violence or gore is available on a service
- Risk that content praising, supporting, promoting, inciting, or that constitutes gender-based violence, including sexual, physical, mental and economic harm, or threats of violence, coercion, and manipulation, is available on a service
- Risk that non-illegal content or behaviour representing, praising, glorifying, assisting, facilitating, or supporting (including through fundraising) terrorist organisations or acts of terrorism or violent extremism is available or takes place on a service
- Risk that non-illegal hate speech is available on a service
- Risk that non-illegal content or behaviour constituting harassment or bullying is available or takes place on a service

Consumer and Business

- Risk that unfair commercial practices take place on a service
- Risk that misrepresentation or misinformation about a business is available on a service

Child Rights

- Risk that children under a defined minimum age access services that they should not be able to for reasons of age-based restrictions
- Risk that children under a defined minimum age are exposed to harmful, hateful, or age-inappropriate content or conduct on a service
- Risk that children's access to and / or use of a service is limited more than is necessary or proportionate for a legitimate purpose
- Risk that children's data are used by a service for ads targeting in ways that have adverse impacts on children's rights, including their right to be protected from economic exploitation
- Risk that an application or service does not perform equitably for children with varied learning styles, learning challenges, or disabilities
- Risk that an application or service primarily directed at or predominantly used by children is not of adequate quality across languages, markets, and age groups and has adverse impacts on children

Equality and Non-Discrimination

- Risk that a service selects organic content or paid speech based on factors that result in discrimination
- Risk that an application or service is not of adequate quality across languages, markets, and age groups
- Risk that some populations are under-represented as content contributors on a service, with adverse impacts on minority businesses
- Risk that algorithms on a service are less well trained in some languages, dialects, and vernaculars than others
- Risk that an application or service does not function equitably for users with disabilities

Civic Discourse

- Risk that misinformation and disinformation relating to elections, civic discourse, democratic participation, or civil unrest are available on a service
- Risk that digital threats such as account hijackings, phishing attempts, or disinformation campaigns are targeted at users of a service during election times and other important civic discourse milestones

- Risk that content with value as evidence in legal process and access to remedy is removed and/or deleted by a service

Public Health

- Risk that content or behaviour which enables, assists, or promotes practices harmful to health is available on a service
- Risk that content or behaviour targeting individuals based on intrinsic attributes (such as protected group status or physical traits) is available or takes place on a service
- Risk that the interface, design, or features of a service stimulate compulsive use of the service for users, including children

Annex B: List of Mitigations

Background

This annex contains additional mitigation measures being put in place consistent with Article 35(1) of the DSA. As part of the systemic risk assessment for each VLOP and VLOSE, we evaluated our existing mitigation measures for each risk statement. As explained in this report, we have long invested in efforts to address user trust and safety and have thus already put in place an extensive array of mitigations. These existing mitigations are discussed in the report, where relevant to the reporting of the results. Please see below for a list of new or enhanced mitigations being put in place consistent with Article 35(1) to address the salient residual systemic risks identified in the Article 34 assessment.

Article 35 Mitigation Types

Mitigation Type	Full Article 35 Mitigation Description
Adapting the design, features or functioning of services	Adapting the design, features or functioning of their services, including their online interfaces
Adapting terms and conditions and its enforcement	Adapting their terms and conditions and their enforcement
Adapting content moderation processes	Adapting content moderation processes, including the speed and quality of processing notices related to specific types of illegal content and, where appropriate, the expeditious removal of, or the disabling of access to, the content notified, in particular in respect of illegal hate speech or cyber violence, as well as adapting any relevant decision making processes and dedicated resources for content moderation
Testing and adapting algorithmic systems	Testing and adapting their algorithmic systems, including their recommender systems
Adapting advertising systems and adopting targeted measures	Adapting their advertising systems and adopting targeted measures aimed at limiting or adjusting the presentation of advertisements in association with the service they provide

Reinforcing internal processes, resources, testing, documentation and supervision	Reinforcing the internal processes, resources, testing, documentation, or supervision of any of their activities in particular as regards detection of systemic risk
Initiating or adjusting cooperation with trusted flaggers	Initiating or adjusting cooperation with trusted flaggers in accordance with Article 22 and the implementation of the decisions of out-of-court dispute settlement bodies pursuant to Article 21
Initiating or adjusting cooperation with other online platform providers	Initiating or adjusting cooperation with other providers of online platforms or of online search engines through the codes of conduct and the crisis protocols referred to in Articles 45 and 48 respectively
Taking awareness-raising measures	Taking awareness-raising measures and adapting their online interface in order to give recipients of the service more information
Taking targeted measures to protect the rights of the child	Taking targeted measures to protect the rights of the child, including age verification and parental control tools, tools aimed at helping minors signal abuse or obtain support, as appropriate
Ensuring that information is distinguishable through prominent markings	Ensuring that an item of information, whether it constitutes a generated or manipulated image, audio or video that appreciably resembles existing persons, objects, places or other entities or events and falsely appears to a person to be authentic or truthful is distinguishable through prominent markings when presented on their online interfaces, and, in addition, providing an easy to use functionality which enables recipients of the service to indicate such information

Mitigations Applicable to Multiple Services

DSA Article 35 Mitigation Type	Mitigation	Description
Reinforcing internal processes, resources, testing, documentation, or supervision.	Improving cross-Google signal sharing to improve scam detection	We will incorporate additional signals regarding incidents and bad actors across services to improve detection of bad actors.
Adapting terms and conditions and its enforcement	Improve safeguards to curb misuse by malicious advertisers	We are investing in updates to refine the process for provisioning tools available to advertisers to encourage responsible use of our platforms and mitigate speed and scale of abuse.
Reinforcing internal processes, resources, testing, documentation, or supervision.	Hardening upstream protections and leveraging responsible feature access to mitigate bad actors	We are further bolstering Google account security and anti-fraud/scams protections upstream to minimise the risk of threat actor access to Google accounts.

Google Maps

DSA Article 35 Mitigation Type	Mitigation	Description
Adapting the design, features or functioning of services	Language expansion for content moderation	We will continue to focus on expanding the breadth of languages supported in content moderation with the EU official languages as priority.

Google Play

DSA Article 35 Mitigation Type	Mitigation	Description
Reinforcing internal processes, resources, testing, documentation and supervision	Malicious app removals	We are expanding the application of existing capabilities to more quickly identify policy violations, malicious apps, and abuse.
Adapting content moderation processes	Reducing profane slang in Play Store reviews	We will continue to reduce the presence of profane slang terms in app reviews across languages on the Play Store.

Google Search

DSA Article 35 Mitigation Type	Mitigation	Description
Testing and adapting algorithmic systems	Personal hardship coping support improvements	We are working on improvements to features like Related Questions in Search that intend to help users to cope with personal hardship and distress by promoting helpful resources (e.g., grieving, financial troubles, severe health conditions etc.).
Testing and adapting algorithmic systems	Enhance mitigations to address explicit fake content on Search	We are continuing to develop and implement measures that address the spread of explicit fake content (e.g., non-consensual explicit fake content) on Search.
Reinforcing internal processes, resources, testing, documentation and supervision	Expanding Search asset coverage for Proactive Data Governance	We are continuing to increase the Search asset coverage for our Proactive Data Governance platform, to increase our ability to find and remediate sensitive data access risks.
Testing and adapting algorithmic systems	Improve blurring functionality for graphic violence in Search	We are improving the SafeSearch blur functionality to better detect content that should be blurred.
Testing and adapting algorithmic systems	Enhance automated detection for violence	We are developing and implementing improvements in automated detection systems for content depicting graphic violence.
Testing and adapting algorithmic systems	Evaluate presence of gender-based violence abuse types on Search	We are evaluating the prevalence and nature of gender-based violence (GBV) abuse types on Search and, based on that evaluation, developing and implementing strategies to mitigate the impact of identified GBV abuse types on users.

Google Shopping

Please see “Mitigations Applicable to Multiple Services”

YouTube

DSA Article 35 Mitigation Type	Mitigation	Description
Adapting terms and conditions and its enforcement	Updating Misinformation policies	We are continually evaluating the need to update our Misinformation policies to address emerging threats.
	Improvements targeting adversarial abuse	We will make continued improvements targeting adversarial abuse.
	Updating hate speech, harassment, and cyberbullying policies	We will continually evaluate the need to update our Hate Speech and Harassment and Cyberbullying policies to address emerging threats.
Adapting the design, features or functioning of services	Enhanced supervision tools	We will introduce supervision tools for parents of teens.
Testing and adapting algorithmic systems	Enhance automated detection for business misinformation	We continually update and improve our automated detection systems based on user flags and other information.

Annex C: List of Consultations

Safety expert engagement during 2024 assessment period

As discussed within the Consulting with Experts section of the report, we regularly engage with academics, independent researchers, and civil society groups. During the assessment period (July 1st, 2023 - June 30th, 2024) we engaged with thousands of safety experts on systemic risk topics, including academics, independent researchers, and civil society groups.

Selected engagements in support of systemic risk assessments

Below is a non-exhaustive list of engagements undertaken by Google Trust and Safety teams during the assessment period that included discussion of the risks addressed by the systemic risk assessments. These engagements demonstrate the breadth of our coverage and commitment to dialogue with external partners on these critical matters.

- Google Growing Up in the Digital Age Child Safety Summit
- European Advertising Standards Alliance (EASA) 2024 Ads Safety Event
- Advancing Explainability Through AI Literacy and Design Resources
- Consumer Leadership Academy with Consumer Empowerment Project (CEP)
- Content Safety in Generative AI at Applied Machine Learning Days (AMLD)
- DSA Trusted Flaggers Panel at FIC Trust and Safety Forum
- European Virtual Forum (EVF) 2024: A Network of Trusted Flaggers
- Google Fighting Misinformation Online: Election Integrity in the Age of AI (Warsaw)
- Google Fighting Misinformation Online: Elections 2024 (Brussels)
- Global Anti Scam Summit Europe 2024
- International Conference on the Promotion of Digital Wellbeing
- Digital Trust and Safety Partnership (DTSP) and Global Network Initiative (GNI) European Rights and Risks Summit

- Trust and Safety Professionals Association (TSPA) EMEA Summit: Panel on Alternative Dispute Resolution
- Trust and Safety Summit UK