



# Systemic Risk Assessment

*August 2024*

Assessment Report

Introduction.....	3
Executive Summary .....	3
Background.....	5
Products at a Glance.....	6
Existing products .....	6
New since last year.....	11
Methodology .....	19
Year 2 Design Enhancements .....	19
Risk Assessment Process .....	20
Risk Assessment Tools .....	22
Summary of Results.....	22
Key Findings.....	22
Scoring Rationale.....	47
Mitigations.....	49
Best Practice Highlights .....	49
Monitoring the Effectiveness of Mitigations.....	60
Progress Made on Enhanced Mitigations.....	61
Planned Enhanced Mitigation Efforts.....	62
Conclusion .....	62
Appendices .....	63
Appendix A: Methodology Expanded.....	63
Appendix B: List of Article 35 mitigations.....	71

## Introduction

LinkedIn Ireland Unlimited Company—the provider of LinkedIn’s services in the European Union (EU)—has been designated by the European Commission as a Very Large Online Platform (VLOP) and is therefore subject to the European Commission’s Digital Services Act (DSA) Article 34 requirement to conduct a Systemic Risk Assessment. This Systemic Risk Assessment Report (the “Report”) is responsive to the DSA Article 42(4)(a) requirement to provide a report setting out: the results of the Article 34 risk assessment, the specific mitigation measures put in place pursuant to Article 35(1), and information about the consultations conducted by the provider in support of the risk assessments and design of the risk mitigation measures.

The Report reflects the Year Two assessment, through August 2024, of systemic risks indicated by DSA Article 34 stemming from the design or functioning of LinkedIn or from the use of LinkedIn services in the EU. The assessment considered the specific risk categories and influencing factors outlined in Article 34 with a focus on their impact in the EU; however, LinkedIn also evaluated these risks and influencing factors at a global level. Accordingly, the findings and mitigations discussed in this Report are largely applicable and relevant to the other regions and countries in which LinkedIn offers its services.

The Report describes assessed risk related to the LinkedIn platform and services holistically. The Report also addresses specific risks or mitigations as they may apply uniquely to specific LinkedIn products. The Report includes an executive summary, background on LinkedIn as it relates to the Systemic Risk Assessment, a brief overview of LinkedIn products to inform their risk profiles, LinkedIn’s approach to and the methodology used for the Systemic Risk Assessment, a summary of the Systemic Risk Assessment results, and an overview of LinkedIn’s risk mitigation efforts, both in place and planned.

## Executive Summary

LinkedIn is committed to keeping its platform and services safe, trusted, and professional and to providing transparency to its members, the public, and to regulators. LinkedIn’s vision is to create economic opportunity for every member of the global workforce. Its mission is to connect the world’s professionals to make them more productive and successful. Members come to LinkedIn to find a job, stay informed, connect with other professionals, and learn new skills. As a real-identity online networking service for professionals to connect and interact with other professionals, LinkedIn has a unique risk profile when compared with many social media platforms. With this in mind, LinkedIn invests heavily in numerous Trust and Safety domains to proactively enhance the safety, security, privacy, and quality of the LinkedIn user experience.

New and sustaining LinkedIn investments include:

- The launch of various free verifications, which aim to prevent fraudulent activity by empowering members to make more informed choices about the authenticity of the people, jobs, and Pages they engage with on LinkedIn;
- Enhanced user choice and optionality for what information users would like to share across LinkedIn services and features and the ability to opt-out of recommended jobs;
- The continued development of advanced models for detecting and preventing the creation of fake or inauthentic accounts, fraudulent activities, or sharing of harmful content, including Child Sexual Abuse Material (CSAM), hate speech, and harassment;
- Enhanced internal tooling for content moderation to drive accountability, quality assurance, cross-functional collaboration, and reporting;
- Clear, actionable, and informed training for LinkedIn Trust and Safety team members with a robust quality assurance program that is proactively and continuously improved to yield more meaningful results; and

- Investment in human reviewer subject matter expertise as well as testing and trials of additional content review methodologies to enhance decision quality and minimize subjectivity in the content moderation lifecycle.

The systemic risks most relevant to users of LinkedIn’s platform and features are (1) Consumer Protection and Fraud, (2) Mental and Physical Well-being, and (3) Discrimination and Hate. Accordingly, LinkedIn has robust policies, detection capabilities, and enforcement processes to mitigate risk in these areas. Some examples include:

- LinkedIn prohibits fake accounts and fraudulent activity on the platform to help ensure a high level of Consumer Protection. LinkedIn has developed sophisticated models to detect and prevent the creation of fake accounts; to remove phishing links, malware, known or suspected scam content, and fraudulent content; and to permanently restrict the accounts of known fraudsters or scammers. All platform content is scanned for malware and phishing upon origination. Since August of 2023 LinkedIn has published at least three Transparency Reports specifically related to LinkedIn’s operations. In the last published Transparency Reporting period, automated defenses blocked 90.5% of the fake accounts that were caught during the July - December 2023 period, with the remaining 9.5% stopped by manual investigations and restrictions. 99.6% of the fake accounts were stopped proactively, before a member report.
- LinkedIn has robust policies and processes in place to proactively address threats of violence, gender-based violence, controlling behavior, self-harm, and other concerns related to Mental and Physical Wellbeing. LinkedIn has deployed models and hash matching algorithms to detect policy violations across user-generated text, video, and images. To strike the right balance between keeping the LinkedIn platform safe and honoring professional speech, depending on the precision levels of these algorithms, content flagged by them will either be automatically moderated or sent to a human reviewer.
- LinkedIn is innovating to reduce hate speech, harassment, and other harmful content or activities linked to Discrimination and Hate on the platform. These initiatives include the optional advanced safety feature for EU members that allows LinkedIn’s automated machine learning models to detect sexual harassment and certain other abusive content<sup>1</sup> in messages received by the member, giving the recipient the ability to view and report the message to LinkedIn for review if desired. LinkedIn also places a high priority on pre-launch product reviews and testing to help reduce discrimination and bias in algorithmic systems.

Artificial intelligence (AI) is not new to LinkedIn. LinkedIn has long used AI to enhance its members’ professional experiences. In 2024, an area of priority for LinkedIn was to enhance its AI and Generative AI (GenAI) capabilities, both in the responsible application of GenAI to directly enhance the LinkedIn user experience as well as in increasing investments in the use of AI algorithms to further enhance content moderation processes. Further leveraging the power of AI and expanding the use of GenAI features in the platform, supports LinkedIn’s objectives to help its members connect, increase productivity, and achieve success in their careers.

While AI has enormous potential to expand access to opportunity and ultimately transform the world of work in positive ways, LinkedIn acknowledges that the use of AI comes with unique risks. That’s why, consistent with its commitment to build a trustworthy platform, LinkedIn is committed to using AI responsibly. Inspired by, and aligned with, Microsoft’s leadership in Responsible AI, LinkedIn uses the following Responsible AI Principles to guide its work:

- **Advance Economic Opportunity:** People are at the center of what LinkedIn does. AI is a tool to further LinkedIn’s vision, empowering LinkedIn members and augmenting their success and productivity.

---

<sup>1</sup> LinkedIn automatically scans all content for malware, phishing, and CSAM.

- **Uphold Trust:** LinkedIn’s commitments to privacy, security and safety guide its use of AI. LinkedIn takes meaningful steps to reduce the potential risks of AI.
- **Promote Fairness and Inclusion:** LinkedIn works to ensure that its use of AI benefits all members fairly, without causing or amplifying unfair bias.
- **Provide Transparency:** Understanding of AI starts with transparency. LinkedIn seeks to explain in clear and simple ways how its use of AI impacts people.
- **Embrace Accountability:** LinkedIn deploys robust AI governance, including assessing and addressing potential harms and fitness for purpose, and ensuring human oversight and accountability. LinkedIn is committed to learning from, and helping, others as AI best practices, norms and laws evolve.

LinkedIn continues to advance its Trust and Safety efforts aligned with these principles and Microsoft’s commitments to advancing responsible innovation.

## Background

LinkedIn is a real-identity, specific-purpose platform, designed for professionals. It is a networking tool that enables members to establish their professional identities online, connect with other professionals, and build relationships for the purpose of collaborating, learning, and staying informed about industry information and trends. As such, the design and function of the platform are central to its overall risk profile related to systemic risks, and shape that risk profile in a few key ways:

- LinkedIn is a real-identity platform, where members must use their real or preferred *professional* names, and the content they post is visible, for example, to their colleagues, employers, potential future employers, and business partners. Given this audience, members by and large tend to limit their activity to professional areas of interest and expect the content they see to be professional in nature.
- LinkedIn operates under standards of professionalism, which are reflected both in content policies and enforcement, as well as in content prioritization and amplification. LinkedIn’s policies bolster a safe, trusted, and professional platform, and LinkedIn strictly enforces them. LinkedIn strives to broadly distribute high-quality content that advances professional conversations on the platform.
- LinkedIn services are tailored toward professionals and businesses. During the account creation process, new members who are listed as attending a secondary education institution must attest to being over the minimum age (16 years) to open an account.

These factors significantly reduce the likelihood of certain systemic risks stemming from LinkedIn’s platform and services. Nonetheless, LinkedIn proactively invests in establishing robust policies, advanced detection capabilities, and effective enforcement processes in line with the Digital Trust and Safety Partnership’s (DTSP) Best Practices Framework. Additionally, as bad actors may resort to creating fake personas to attempt abusive or malicious activities on the platform, LinkedIn invests resources in developing techniques to prevent the creation of fake Profiles and provides identity verification and other authenticity signals to members.

One of LinkedIn’s core values is “**members first.**” This means that every day, when making decisions large and small, LinkedIn asks “Is this the right thing to do for our members?” This value drives LinkedIn, unites the company, and pushes it to remain worthy of the trust of LinkedIn members worldwide.

As LinkedIn’s impact on the world of work has grown, the company recognizes that its responsibility has grown as well. LinkedIn understands that in the critical areas of privacy, security, safety, and fairness, it needs to continue to deliver for members and the world beyond. So, in addition to this foundational value of members first, LinkedIn uses a set of Guiding Principles for Trust to ensure its employees make the best possible decisions to protect LinkedIn members and maintain their trust.

- LinkedIn provides its members with **clarity, consistency, and control** over their data. LinkedIn tells members what it will do with their data. LinkedIn then does what it says it will do with this data. And LinkedIn strives to provide its members with simple, easy to understand settings and controls so that they are in control of their data.
- LinkedIn is focused on keeping its platform **safe, trusted, and professional**. It uses systems, technology, and reports from its members to detect and quickly remove any content that violates LinkedIn's Professional Community Policies. Members also rightfully expect content they encounter on LinkedIn to be legitimate. Removing fake Profiles, jobs, and companies is paramount to this objective.
- LinkedIn believes two members with equal talent should have **equal access** to opportunity. To achieve this goal, LinkedIn is committed to building a product with no unfair bias that provides access to opportunity to all its members. There is a lot of work still to do, but LinkedIn is focused on working across the company, with members and customers, and across the industry to close the network gap.
- LinkedIn is a global platform with an obligation to **respect the laws** that apply to it. It also contributes to the dialogue that shapes these laws so that LinkedIn can fulfill its vision of creating economic opportunity for every member of the global workforce.

Members come to LinkedIn to find a job, stay informed, connect with other professionals, and learn new skills. It's LinkedIn's responsibility to give them a safe and trusted platform to do just that and be open about how LinkedIn makes decisions to build the professional community they expect.

## Products at a Glance

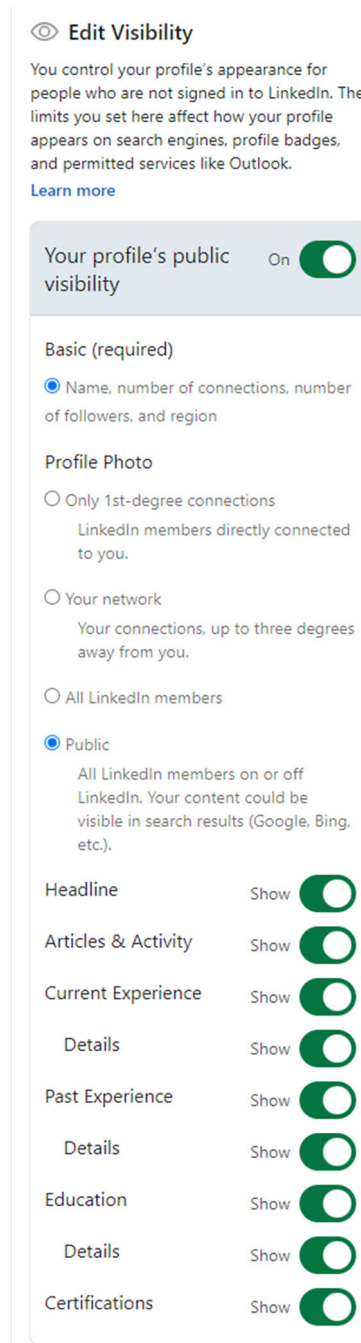
LinkedIn has members in over 200 countries and territories worldwide and operates a diversified business with revenue coming from its Talent Solutions (LTS), Marketing Solutions (LMS), Sales Solutions, and Premium Subscriptions products. While the Systemic Risk Assessment evaluates risk stemming from use or misuse of the LinkedIn online platform holistically, key products and features are described here to facilitate understanding of the platform's functioning and use.

### Existing products

**Profile:** Each member account has a LinkedIn "Profile" associated with it. Functioning as an online resumé or curriculum vitae, the LinkedIn Profile serves as the member's professional landing page on which they can manage their own personal, professional brand. Members must use their real or preferred professional names and include accurate information within their Profiles. Additionally, members must not misrepresent their identity, create a member Profile for anyone other than themselves (a real person), or use or attempt to use another's account.

The Profile includes space for a personal headshot photo, as well as a "background" photo. There is an "about" section, as well as space for education history, work history, location, industry, contact information, skills, endorsements, and licenses and certifications, among other things. There is a "recent activity" section that shows snippets and thumbnails of recent posts or other actions by the member. Using their Profile settings, members can control the extent to which certain aspects of their Profile information is accessible via LinkedIn's in-product search feature (e.g., searching for a person by name or category), as well as by general browsing of the platform. For example, settings control who can see the member's Profile photo and last name. Additionally, although portions of Profiles may be accessible by default via search engines to non-members or members in a logged-out state, members generally can control this via settings like those displayed in Figure 1.

Figure 1



**Feed:** LinkedIn focuses on providing its members professionally relevant content through their Feed, which enables them to engage with people and topics they value. The LinkedIn Feed is the place on the platform where members go to view content, most of which is created by other members. Content can appear in a member's Feed because (1) the member follows or is otherwise connected through LinkedIn with the person or Page that posted it; (2) one of the member's LinkedIn connections reacted to, commented, or shared the content; or (3) it bears relevance to that member's job or industry or interests. Members may also see content from Groups

they've joined, hashtags they follow, or Events they're attending. Members can customize and manage the content that appears in their Feed by following, unfollowing, blocking, or muting people, companies, and hashtags; hiding content in their Feed that they don't want to see; and reporting content that may violate LinkedIn's Professional Community Policies.

The following types of content can appear in the Feed:

- *Text*: This can be either short snippets / messages or long form posts that read more like an article or essay.
- *Images*: These can accompany text posts or be standalone images.
- *Video*: The Feed supports recorded videos, as well as links / embeddings of videos posted elsewhere. It is also possible to conduct a "livestream" event that appears in members' Feeds.
- *Ads*: Advertisements can appear in, or adjacent to, the Feed.
- *Notifications and Suggestions*: Suggestions for other people the member may want to connect with or follow, Pages and Groups the member may want to follow, LinkedIn Learning courses the member may want to take, and jobs the member may want to apply to may all appear in the Feed, as well as notices about connection updates or work anniversaries (subject to user settings).

A member's Feed experience is dynamic and specific to that logged-in member, based on, among other things, the member's LinkedIn Profile and network. Members can choose to view their Feed in a personalized or chronological manner. Logged-out users or guests do not have access to a LinkedIn Feed.

**Pages:** Individual companies, organizations, and institutions can create their entity Profiles on LinkedIn (called LinkedIn Pages) to allow LinkedIn members to discover, follow, and engage with those organizations and learn more about them. To create LinkedIn Pages, LinkedIn members must verify that they are authorized representatives of their organization and agree to the LinkedIn Pages Terms. Organizations with LinkedIn Pages can have one or more LinkedIn member Page administrators who manage the Page and post content (such as industry updates, information about the organization, and job posts) to build and promote the organization's brand to the professional audience on LinkedIn.

**Groups:** LinkedIn Groups are forums members join to discuss shared interests or experiences. Groups function similarly to Pages, but unlike Pages, Groups are created and maintained by individual members rather than organizations or institutions. A Group can be public or private. Group members can send message requests to other Group members and start discussions or conversations in the Group by posting in the Group's Feed text, images, a video, or a link. Group members may see each other's Group posts, likes, and comments within the Group, or in their Feed.

**Newsletters:** LinkedIn's Newsletter feature allows logged-in members to author a set of regularly published articles on LinkedIn, typically about a specific professional topic that interests the author. Each LinkedIn Newsletter has its own Page where members can learn about the newsletter and its author, see past editions, and subscribe to or share the Newsletter. The Page includes an overview of the Newsletter, the author's Profile information, and a list of the articles in the newsletter.

**Messaging:** LinkedIn's messaging feature allows members to contact and communicate with other LinkedIn members through direct messages. LinkedIn members can send (1) a message to a connection or multiple connections in a Group chat; (2) a message request to certain non-connections based on a shared context (i.e., a message request to a fellow Group member or coworker at your current company); or (3) a connection request

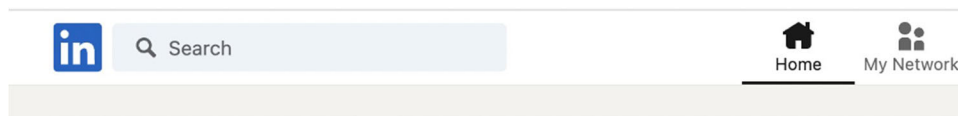


with a personal message to other members. In addition, “InMail” messages are a premium (for a fee) feature that requires the use of credits to message non-connections.

**Jobs:** The LinkedIn Jobs page contains job listings posted by employers or compiled from related sites and services. There are various options to apply for Jobs via LinkedIn, though some listings only allow applicants to apply for a Job directly from the employer’s own application website. Recommended Jobs are curated for the individual member based on their skills, job titles, education, interests, and various other attributes, including Jobs that the member has specifically indicated to LinkedIn they are interested in. Members are also able to search for Jobs by title, company, and other attributes.

**Search:** Using LinkedIn’s in-product search feature located at the top of most LinkedIn Pages, LinkedIn members can conduct a search across most of the platform, including posts and Profiles of other LinkedIn members. A searching member has the option to “See all results” for a keyword search or to narrow a search by categories such as Jobs or people.

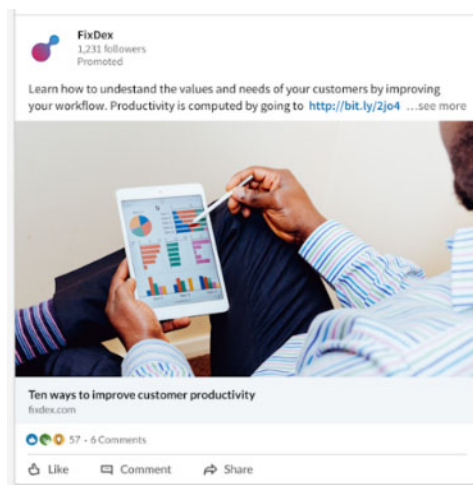
Figure 2



**LinkedIn Marketing Solutions (LMS):** LinkedIn’s advertising platform primarily consists of fully integrated first-party advertising on LinkedIn with a focus on “business-to-business” marketing (i.e., one business advertising to another business, as opposed to a consumer). For members in Europe, there are three types of ads on LinkedIn: Sponsored Content, Text Ads, and Dynamic Ads:<sup>2</sup>

- *Sponsored Content:* These are native posts in the Feed from the advertising customer. They can feature text, images, a video, or a job posting and links to the advertiser’s website of choice. These posts are identified as advertisements by terms such as “Promoted” that appear in the upper-left corner of the post.

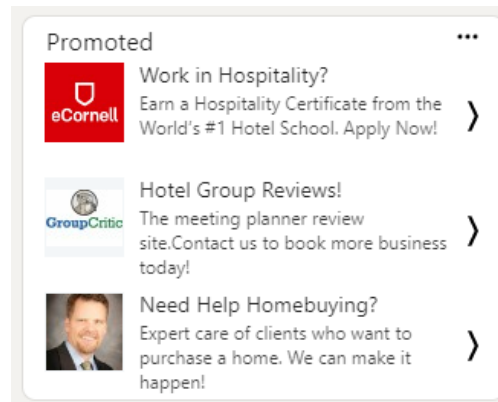
Figure 3



<sup>2</sup> There is a fourth type currently shown to members outside of the EU known as Sponsored Messaging. These appear as messages in the member’s LinkedIn inbox that are clearly labelled as “Sponsored.”

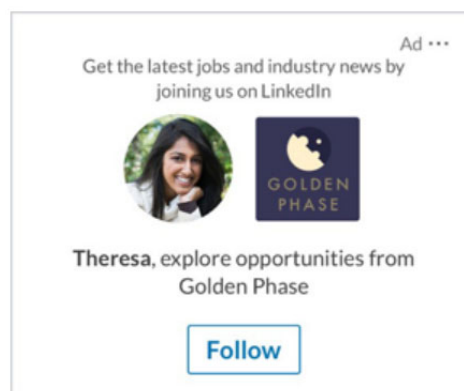
- **Text Ads:** These include a headline, brief text, and an image. They are shown either above or to the right of – and not in – the Feed. They are only shown on the LinkedIn website, not the mobile application. These are also identified as advertisements by a “Promoted” label.

Figure 4



- **Dynamic Ads:** These are personalized ads using the member’s Profile image, name, or job function to create customized ads directed toward that member. Dynamic ads appear to the right of – and not in – the Feed and are not shown on the mobile application. They are marked with an “Ad” label in their top-right corner.

Figure 5



LinkedIn members may receive targeted ads based on information provided and inferred from their LinkedIn Profile and usage information. For example, years of experience can be inferred from the sum of the years of work history indicated on a member’s Profile and professional interests can be inferred from LinkedIn Feed posts. Members can view details about why specific ads are displayed to them by clicking the three-dot icon in the upper-right corner of the ad and selecting “Why Am I Seeing This Ad?”

Using their [ad settings](#), a member can manage general advertising preferences, how data collected on LinkedIn is used to show the member ads, and how third-party data is used to show the member ads. If a member limits the information used to improve ad relevance, that member will still see ads on LinkedIn, but they may not be related to the member’s interests and/or experience.

## New since last year

**LinkedIn's approach to product updates:** At LinkedIn, new products and initiatives undergo a pre-launch review [process](#) to identify and address potential associated risks and abuse vectors that are likely to impact systemic risk. In cases where that impact is potentially critical, LinkedIn may initiate an out-of-cycle assessment of systemic risk to ensure adequate mitigations are in place prior to the launch. While no product updates this year met the criteria for an out-of-cycle update to the Systemic Risk Assessment (no updates were determined to be likely to have a critical impact on the risks identified pursuant to DSA Article 34), highlighted below are several changes to the service that were reviewed and implemented since the last Systemic Risk Assessment report.

**Verification:** LinkedIn's voluntary verification features allow members to verify certain information about the Pages they administer, the Jobs they post, or their Profile (like the member's association with a particular company or educational institution, or their identity, using a government-issued ID through one of LinkedIn's identity verification partners). Verified information, which is marked by a badge symbol in various locations on the platform, provides members with authenticity signals about the people, Jobs and Pages they encounter on LinkedIn. These signals are designed to help members make more informed decisions about connecting or engaging with other professionals, organizations, or Jobs.

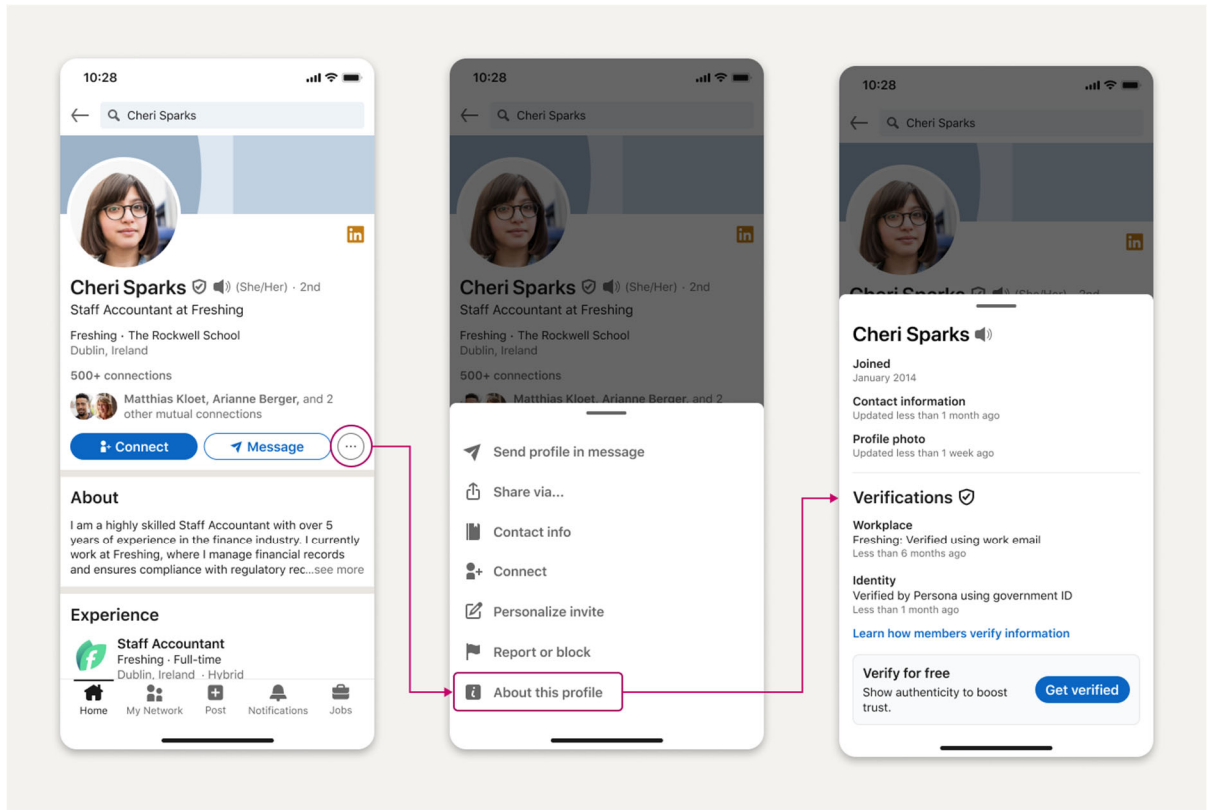
Currently, LinkedIn supports several forms of verification. Each form of verification has its own eligibility and availability criteria. Verifications are free, aimed at increasing trust and authenticity on the platform, and can be removed at any time by the member or entity whose information is verified.

Figure 6



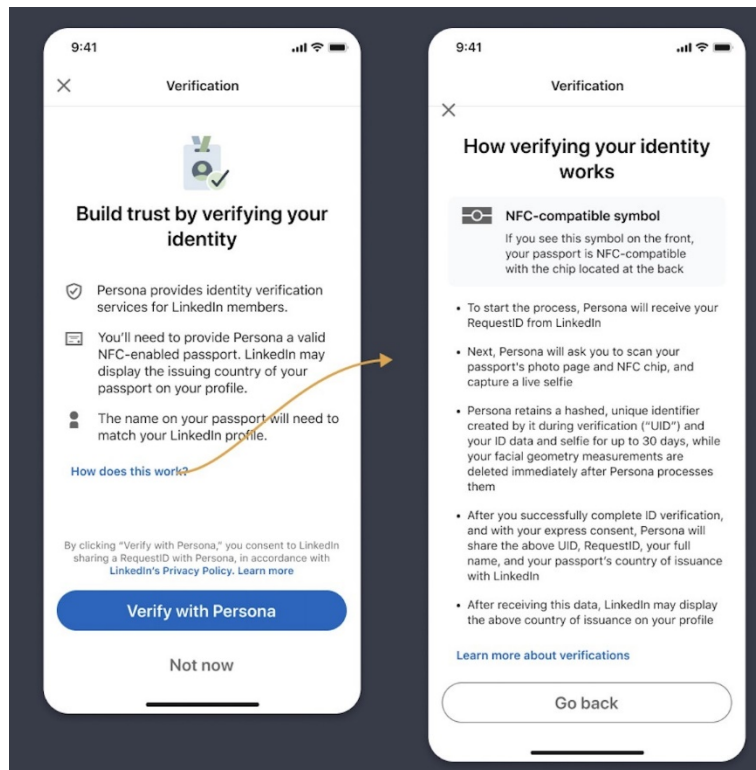
- **Verifications on member Profiles:** As described in greater detail below, members can add an identity, workplace, or educational institution verification to their Profile. Once the member has successfully done so, a verification badge is displayed next to the member's name on their Profile and information about the verification is included in their "About this Profile" section. The badge may also appear next to the member's name elsewhere within LinkedIn products (e.g., next to Jobs the member posts). This information will be visible to logged-in members.

Figure 7



- **Identity Verification:** EU Members may verify their identity through LinkedIn’s verification partner, Persona. LinkedIn uses different identity verification partners for other countries and regions, including CLEAR in the United States, Canada and Mexico and DigiLocker in India. As shown below, LinkedIn provides basic information about how the verification process works for members with Near-Field Communication or NFC-enabled government IDs (e.g., E-passports) issued in the EU. If a member opts to “Verify with Persona,” they are redirected to Persona, which will collect an image of the member’s government issued ID and a live selfie.

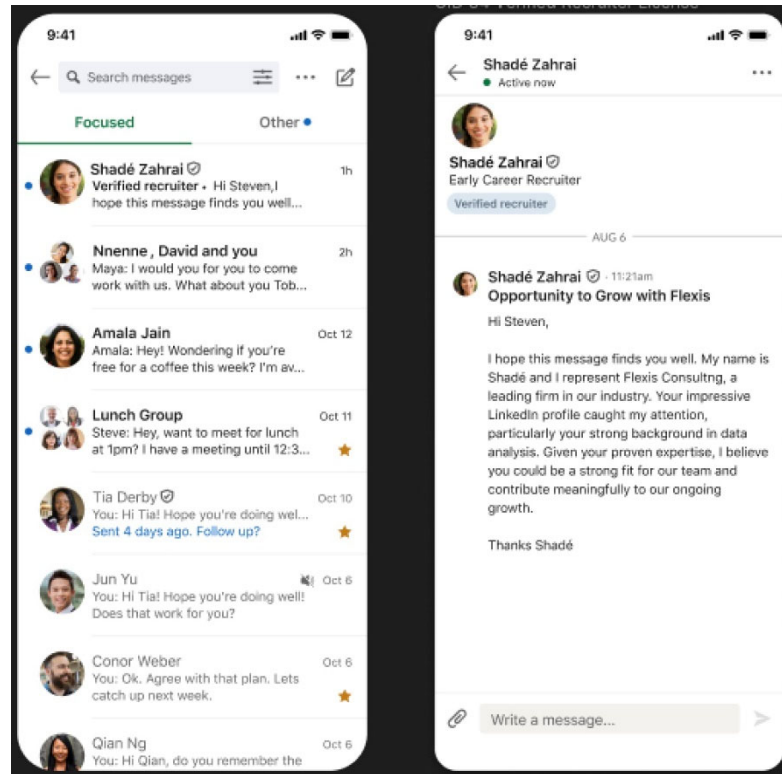
Figure 8



If the member successfully completes Persona’s verification process, LinkedIn does not receive a copy of any government IDs from Persona, but -with the member’s consent- Persona shares certain limited data about the member with LinkedIn (such as the member’s name as it appears on the government ID). LinkedIn uses this data to add a verification to the member’s Profile. LinkedIn may also use this data for security purposes (e.g., to prevent identity theft and fraudulent use of LinkedIn’s verification processes), in accordance with LinkedIn’s Privacy Policy.

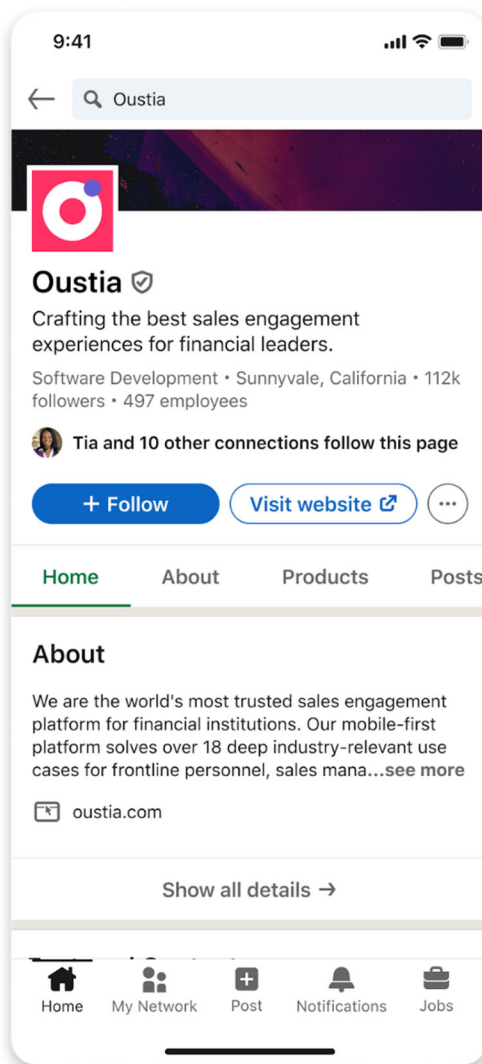
- **Workplace Verification:** Members may confirm their association with a specific company (whose LinkedIn Page is verified) using their work email, Microsoft Entra Verified ID, or a LinkedIn Learning license or active LinkedIn Recruiter license provided by the company.
- **Verified Recruiter Label:** When a member (1) verifies their workplace using a LinkedIn Recruiter license and (2) currently lists the company that issued that license as their current employer in the “Work experience” section of their Profile, a “Verified Recruiter” label will appear near the member’s name in InMails sent via LinkedIn’s Recruiter product.

Figure 9



- **Educational Institution Verification:** An educational institution verification indicates that the member has confirmed their association with an educational institution by using an email address containing a domain allowed by the institution, as indicated by the super admins of a verified institution's LinkedIn Page. Currently, the feature is only available to select educational institutions as LinkedIn continues to expand support for this feature.
- **Page Verification:** LinkedIn will display a verification badge next to the organization name on LinkedIn Pages that meet certain criteria (e.g., active Page admin presence, compliance with LinkedIn's policies, and accuracy of certain data points on the Page, like company location and URL). This verification badge helps signal that the Page is an official Page of the organization, which helps organizations build credibility with potential customers and helps members find the actual organizations they intend to engage with.

Figure 10



Once a Page is verified: (1) LinkedIn members can verify their employment with the Page’s organization using a Microsoft Entra Verified ID (if any), or a LinkedIn Learning or LinkedIn Recruiter license provided by the organization; and (2) the Page’s super admins can enable members to verify their association with the organization using a work or educational institution email address containing domains controlled by the super admins.

- **Verifications on Jobs:** To provide job seekers additional trust and authenticity signals as they apply for Jobs posted on LinkedIn, LinkedIn introduced a new badge on Job posts. When members see a verification badge on a Job post, it means that certain information about the company has been confirmed by LinkedIn, or information about the job poster has been verified by the job poster through LinkedIn or through a trusted third-party partner.

**GenAI Features:** Over the last year, LinkedIn has integrated multiple GenAI features into existing products to enhance LinkedIn member experiences, providing members with options to more easily consume and share

information. LinkedIn's GenAI features fall into two categories (1) the generation of new content and (2) the generation of insights based on existing content. These features include:

Generation of new content:

- **Feed:** LinkedIn's [Writing Assistant](#) can create draft LinkedIn posts (i.e., draft textual content which a member may choose to post on the platform) based on Premium subscriber prompts.
- **Messaging:** LinkedIn's Writing Assistant can help members create a first draft message (text only). Premium members can create a draft message using pre-set prompt suggestions or by typing in their own prompt. Recruiters can also create draft messages (for potential job candidates).
- **LinkedIn Talent Solutions (LTS):** For job posters, LinkedIn's AI-Powered job descriptions help recruiters create job postings with initial information provided by the job poster, such as job title, company name, workplace type, job type, and location, or from the job poster's profile, such as company and location. Recruiters will review and can edit all elements of the draft job posting (including the AI-suggested text) prior to publication.
- **LinkedIn Marketing Solutions (LMS):** LMS GenAI features are integrated into LinkedIn's existing Campaign Manager advertising platform and designed to simplify advertisement development and campaign initiation. Campaign Manager incorporates Microsoft Designer, a third-party design tool that leverages GenAI to empower marketers to create unique and high-quality graphics via prompts, in less time, and without any design background. Advertisers can also leverage AI to create a draft subject line for their conversational advertisement, which is an advertisement within the LinkedIn messaging UI.
- **Collaborative Articles:** Collaborative articles are textual articles focused on a specific professional topic that are published by LinkedIn with insights and perspectives contributed by LinkedIn members. Articles begin as AI-powered conversation starters developed with oversight from a LinkedIn team. Using a skills graph, LinkedIn matches articles with relevant member experts who can contribute their lessons, anecdotes, and advice based on their professional experience.
- **Ancillary Features:** Additional ancillary GenAI features have been launched throughout LinkedIn to enhance the value of the platform to its members. For example, Virtual Chat Assistant (VCA) is a text-based interface that offers AI-generated suggested responses to customer questions, drawing primarily on the help center content already available to those customers. For unresolved queries, the customer is connected with a live agent.

Generation of insights based on existing content:

- **Profile:** LinkedIn's AI-powered Writing Assistant shares personalized suggestions upon request to enhance the Headline and About section of a Premium subscriber's Profile. The Resume to Profile Suggestions feature also allows members to easily transfer information from an uploaded resume to their LinkedIn Profile. Members can also use Post Ideas to suggest questions or topics a member could write about in their next post based on the experience included in their Profile.
- **Feed:** AI-powered Insights on LinkedIn use GenAI to provide members with personalized insights and key takeaways on posted content.
- **LinkedIn Talent Solutions (LTS):** For job seekers, AI-powered job insights will help job seekers learn more about a job or company, understand their fit for a role, and how to best position themselves if they apply. Premium subscribers can also use GenAI to tailor their resumes and cover letters for a particular



role. LinkedIn leverages Microsoft technology to power some parts of this feature. Therefore, the Microsoft Privacy Statement will apply to any personal data members may input as a prompt for the AI-powered takeaways feature, as well as any personal data included in any Microsoft-provided results.

For enterprise seat holders of LinkedIn Recruiter, AI-Assisted Search and Projects enable recruiters to quickly search for candidates, create projects, and get recommendations by typing out needs into prompt box and chat experience.

- LinkedIn Marketing Solutions (LMS):** Accelerate is LinkedIn’s AI-powered advertising campaign type, which can help advertisers to more easily create campaigns. Accelerate aims to improve return on ad spend by continuously finding the right combination of targeting, creative, bidding, and placement for ad campaigns. When an advertiser creates an Accelerate campaign, they will provide information about the product, service, or brand, which is then analyzed to automatically generate suggested budget, schedule, audience, creative, and ad placements, each of which is reviewed by the advertiser before launching.

**Games:** LinkedIn launched a number of daily, thinking-oriented games designed to allow members additional opportunities to engage within their networks and spark conversations. The following games are available to EU members:

Figure 11

- Pinpoint—a word association game;

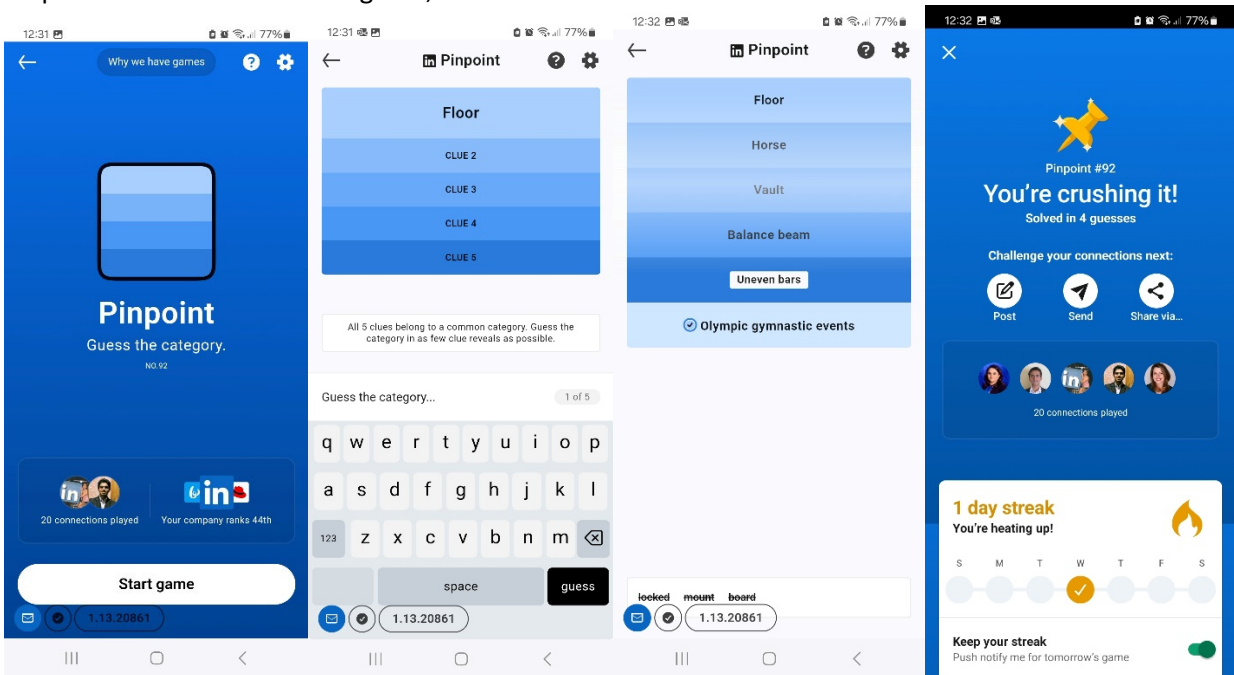


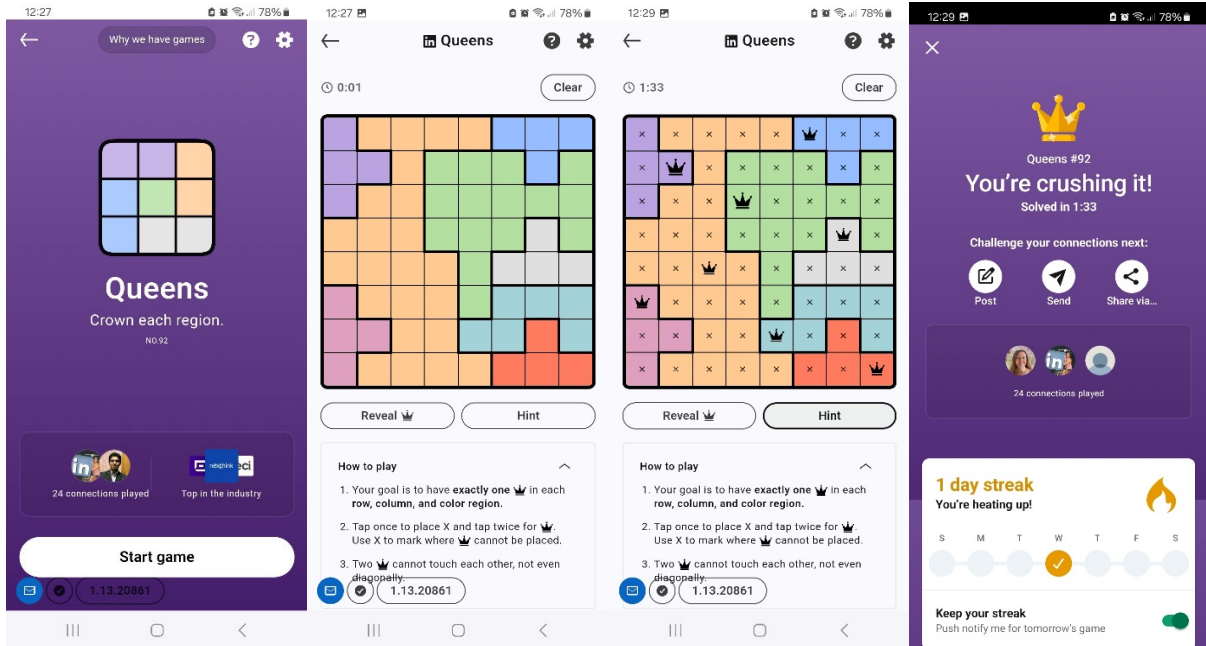
Figure 12

- Crossclimb—a timed trivia game; and



Figure 13

- Queens—a logic game.



After members play one of LinkedIn's games, they have the option to share scores with their network, see their streaks, and find the leaderboards. Games are available to all logged-in members. There are no prizes for winning games.

**Video Tab:** During the assessment period, LinkedIn also launched the Video Tab, a new entry point at the bottom of certain LinkedIn mobile experiences. When a member clicks on the Video Tab icon, it leads them to an immersive video player that allows members to swipe between editorially curated short form videos. The Video Tab only features video content stemming from the Feed, and each video in the Video Tab features the same underlying post, comments, and reactions as the original video in the Feed.

The reporting mechanisms that exist for user generated content across LinkedIn also are available in the Video Tab. Additionally, videos accessible through Video Tab are run through the set of classifiers that exist for videos available through the main Feed.

## Methodology

As in Year 1, LinkedIn aligned its risk assessment process and methodology with the DTSP's Safe Framework, the industry standard for the evaluation of online safety best practices. LinkedIn also engaged a global consulting firm with experience and expertise in the conduct of risk assessments across industries, including technology, energy, climate, finance, and human rights, to support the development of the process and methodology. During the assessment, probability of risk manifestation and severity were considered as inputs to assess the Inherent Risk (risk absent mitigations). Reasonableness, proportionality, and effectiveness of mitigations (informed by the best available information and scientific insights) were also considered as inputs for scoring and factored into calculating the Residual Risk (risk after mitigations).

LinkedIn first evaluated its organizational size and scale, product or digital service impact, and business landscape considerations to determine that a Level 3 Assessment – the most comprehensive level of assessment a company may undertake to examine Trust and Safety practices within the framework – would be appropriate for LinkedIn. It then followed the Safe Framework process to conduct a full assessment including **Discovery, Identification, Assessment, Testing, and Reporting**, which it believes to be proportionate not only to the platform but to the risks identified as well.

Throughout the process, LinkedIn examined risk across the twelve risk areas identified in Article 34. The methodology is summarized here and further expanded in [Appendix A: Methodology Expanded](#).

## Year 2 Design Enhancements

This year, LinkedIn followed the same methodology used in the Year 1 Systemic Risk Assessment, with some enhancements to incorporate lessons learned and enable a more tailored consideration of specific systemic risks to relevant mitigations. These enhancements include the following:

- Separating the Civic Discourse and Electoral Processes Risk Area from Public Security to enable a more direct focus on elections integrity (creating 12 instead of 11 Risk Areas for this review period);
- Replacing Gender-Based Violence with a broader Risk Area of Mental and Physical Wellbeing to consider risks related to individual wellbeing, such as off-platform safety, mental health, and gender-based violence, within the same category;
- Shifting specific risks, which may or may not be illegal or have legal components in various EU countries, out of the Illegal Content risk area into the more appropriate adjusted Risk Areas. These adjustments include shifting terrorist content into Public Security, stalking and threats of violence into Mental and Physical Wellbeing, and hate speech into Discrimination and Hate;
- Creating discrete Risk Scenarios within each Risk Area to more clearly and explicitly identify which risks are considered under each risk area, incorporating both risks specific to DSA Article 34 as well as LinkedIn's more holistic understanding of potential risks to users; and
- Documenting more systematically the Risk Manifestations – or how each risk scenario may or may not present on a product or feature – for each of the 12 Risk Areas across each in-scope LinkedIn product to

more clearly document the appropriateness of the implemented mitigations to the potential risks on the across each product.

## Risk Assessment Process

### *Discover*

As part of discovery, LinkedIn collected responses to the previously modified DTSP Safe Assessment questionnaire; examined external authoritative sources and guidance; and collected and reviewed relevant public policies and publications to identify any adjustments needed to the Systemic Risk Assessment process this year.

Inputs to Risk Assessment:

- DTSP questionnaire responses
- Authoritative sources
- Policies and publications

### *Identify*

LinkedIn built upon last year's risk identification process and focused this year on identifying updates and changes in risk trends, the product itself, and implemented mitigations. In support of this process, LinkedIn collected written updates from stakeholders on the manifestation and mitigation of risks across products and Risk Areas; reviewed relevant internal controls; conducted stakeholder engagements to dive deeper into products with relevant feature updates this year; conducted a tailored engagement with a civil society organization; and analyzed key internal and external metrics related to Trust and Safety policy enforcement as well trends in potential systemic risk related to LinkedIn.

Inputs to Risk Assessment:

- Product and feature changes
- Risk and Mitigation summaries
- Internal consultations
- External consultations
- Internal content moderation metrics
- Transparency reporting metrics
- Open-source data on public discourse related to LinkedIn and systemic risk areas
- DSA Risk and Control Matrix

### *Assess*

LinkedIn assessed the probability and severity of each of the twelve systemic risks stemming from use or misuse of its platform and services, as well as the maturity of risk mitigation measures implemented to arrive at a prioritization of systemic risk areas with potentially higher levels of Residual Risk. The assessment incorporated the referenced eleven inputs from the Discovery and Identification phases to inform scores and implemented a traditional risk assessment equation to aid in prioritization of risk for action.

LinkedIn employed a **data-driven probability assessment**, combining insights from internal, externalized, and open-source data to assess the likelihood of specific risks stemming from use or misuse of LinkedIn considering the vulnerability of the platform *absent* mitigation measures and user demand or likely frequency of attempts to perpetrate each harm on the platform. This methodology follows models used in climate and energy impact assessments.

LinkedIn leveraged an **objective, systems-based assessment of severity**, examining the systems impacted<sup>3</sup> by the systemic risk area (environmental, geographic, political, security, societal, and wellbeing), including scale and gravity of impact (for example, fraud poses a high economic risk at the individual level, but a lower environmental risk at the global level). This methodology follows models used in environmental impact assessments.

LinkedIn aligned existing mitigations and controls to the 35 DTSP Best Practices and evaluated their implementation according to the **DTSP Maturity Rating**, while also considering their proportionality, reasonableness, and effectiveness relative to the potential manifestation of risks across the LinkedIn product. This methodology follows the industry standard for evaluation of Trust and Safety best practices.

LinkedIn also considered for each applicable Risk Area and mitigation whether and how the following factors influence any of the systemic risks: the design of recommender systems and other relevant algorithmic systems; content moderation systems; applicable terms and conditions and their enforcement; systems for selecting and presenting advertisements; data-related practices; intentional manipulation including inauthentic use or automated exploitation; amplification and potentially rapid and wide dissemination of illegal or violative content; and linguistic and regional considerations.

LinkedIn considered the Inherent Risk for each risk area resulting from combined ratings of probability and severity and reduced the Inherent Risk rating by a percentage proportional to the assessed strength of relevant mitigations. This methodology produced a view of Residual Risk across the risk areas to enable prioritization of specific risk areas for further mitigation, in alignment with Article 35 of the DSA.

### *Test*

Quality Assurance is the process of taking systematic efforts to ensure that the product delivered to LinkedIn's members meets with their performance, design, reliability, and maintainability expectations. LinkedIn routinely conducts quality assurance activities and tests the effectiveness of its algorithmic systems and content moderation processes. The DSA Compliance Program will establish a testing strategy to test the current and future formal controls mapped to the Systemic Risk Assessment. For the period between the 2023 and 2024 Systemic Risk Assessments, LinkedIn is relying on the results of the third-party performed audit to inform on the effectiveness of those controls.

### *Report*

This Report fulfills step 5 of the DTSP Safe Assessment Framework process in addition to Article 42(4) of the DSA. LinkedIn reserves the right, as indicated in DSA Article 42(5) to remove such information from the publicly available version of this Report that may result in the disclosure of confidential information, cause significant vulnerabilities for the security of the service, or undermine public security or harm recipients.

---

<sup>3</sup> **Environmental System** is the interplay of natural processes and ecosystems that sustain life on Earth, including the atmosphere, land, water, and biodiversity; **Geographic System** is the physical and spatial environments that shape natural landscapes and human societies including the distribution of climates and resources across the Earth; **Political System** is the structures, institutions, and processes through which decisions are made and power is allocated within a society, shaping governance and public policy; **Security System** is the measures and institutions dedicated to protecting a society from internal and external threats, ensuring the safety and stability of the state and its citizens; **Societal System** is the social relationships, cultural norms, and institutions that bind individuals together, influencing values, behaviors, and community life; **Wellbeing System** is the overall health, happiness, and prosperity of individuals, encompassing physical and mental health, economic security, and quality of life.

## Risk Assessment Tools

LinkedIn employed a Systemic Risk Assessment workbook to facilitate the standardized completion of Systemic Risk Assessment and to store relevant ratings, scores, and rationale. The Risk Assessment team used this workbook to collect and assess the relevant risk factors and risk scenarios, to identify potential manifestation of risks across LinkedIn products and services absent safeguards, to capture and align implemented mitigations, to map mitigations to identified risks, and to calculate and display all risk assessment scores.

## Summary of Results

LinkedIn assessed systemic risk stemming from the functioning, use, or misuse of the platform and services within twelve key risk areas. To identify the residual systemic risk rating across each Risk Area, LinkedIn identified respective mitigations that actively work to address the Risk Scenarios in one or more Risk Area. Where a mitigation applies to more than one Risk Area, it may be discussed in more than one location throughout this Report. The full Risk Areas, definitions, and Risk Scenarios are included in [Appendix A: Methodology Expanded](#).

## Key Findings

### Risk Scores

Using the above methodology, LinkedIn calculated the risk scores listed below for each of the twelve Risk Areas covered by this Report. Definitions for each potential Inherent Probability, Inherent Severity, Inherent Risk, Mitigation Maturity, and Residual Risk rating can be found in [Appendix A: Methodology Expanded](#).

Figure 14

Risk Area	Inherent Probability	Inherent Severity	Inherent Risk Rating	Mitigation Maturity	Residual Risk Rating
<i>Consumer Protection &amp; Fraud</i>	Expected	High	Critical	Managed	Moderate
<i>Human Dignity</i>	Likely	Critical	High	Defined	Moderate
<i>Mental &amp; Physical Well Being</i>	Likely	Critical	High	Managed	Low
<i>Discrimination and Hate</i>	Likely	High	High	Managed	Low
<i>Protection of Personal Data</i>	Likely	High	High	Managed	Low
<i>Civic Discourse &amp; Electoral Processes</i>	Likely	Critical	High	Managed	Low
<i>Freedom of Expression and Information</i>	Likely	High	High	Managed	Low
<i>Public Health</i>	Not Likely	Critical	Moderate	Defined	Low
<i>Public Security</i>	Not Likely	Critical	Moderate	Defined	Low
<i>“Illegal” Content &amp; Activities</i>	Not Likely	Critical	Moderate	Managed	Low
<i>Private and Family Life</i>	Remote	High	Low	Managed	Minimal
<i>Rights &amp; Protection of Minors</i>	Remote	Critical	Low	Managed	Minimal



### *Risk Areas of Moderate Residual Risk*

Considering the probability, severity, and maturity of mitigations, LinkedIn assessed that two of the assessed risk areas fell into a Residual Risk category of Moderate: Consumer Protection and Fraud and Human Dignity.

#### **Consumer Protection and Fraud**

Risks related to Consumer Protection and Fraud include scamming, defrauding, spamming, or falsifying information, including the creation of fake accounts and use of bots or other deceptive commercial practices on LinkedIn. Absent sufficient mitigations, risks related to Consumer Protection and Fraud may manifest on the platform in ways such as users creating fake profiles and defrauding members through private messages or posting scams on Feed, advertisers posting links to fraudulent content or malware, or users posting fraudulent job postings.

The **Inherent Probability** of systemic risks concerning Consumer Protection and Fraud stemming from the function, use, or misuse of LinkedIn's services is assessed as "Expected." The LinkedIn platform favors knowledge sharing rather than virality, and therefore, tends to not facilitate the rapid and wide dissemination of content that does not meet a minimum quality bar. In turn, this limits the systemic risk posed by the spread of harmful content. However, while there is a lower likelihood of members sharing fraudulent content on LinkedIn due to its professional nature, there is a higher risk of bad actors creating fraudulent accounts and using them to defraud members due to the financial incentives of fraudulent activities. Out of the twelve risk areas, Consumer Protection and Fraud received the highest probability score due to the relative frequency of content or activities related to Consumer Protection and Fraud appearing in LinkedIn's content moderation systems as well as public discourse related to LinkedIn. Therefore, the probability is assessed as "Expected" that users could share or be exposed to fraudulent content or activities negatively impacting Consumer Protection on LinkedIn's platform absent sufficient mitigations.

The **Inherent Severity** of impact for fraudulent content or activities on the LinkedIn platform is rated as "High" due to the gravity of risks within this category and the potential for significant economic impact at the individual level. While economic, and even security, societal, and wellbeing, impact at the individual level can be severe, the risk is generally limited to the individual level and does not necessarily scale to the country, regional, or global impact scale. Not all risks within this category would be considered "High," as the impact of spam and deceptive advertisements are less likely to cause significant harm, compared to malware links or deceptive Job postings.

With the rating for Inherent Probability as "Expected" and the Inherent Severity as "High," the **Inherent Risk** associated with negative impacts to Consumer Protection and Fraud across LinkedIn products and features is therefore assessed as "Critical."

In addition to the DTSP Best Practices LinkedIn has effectively implemented across all Risk Areas, which are further described under "[Mitigations](#)," LinkedIn has implemented robust mitigations specific to Consumer Protection and Fraud following industry best practices with defined, documented, and managed processes. As such, the **maturity of mitigation** efforts LinkedIn has applied to the Consumer Protection and Fraud risk area is assessed as "Managed," which brings the **Residual Risk** rating down to "Moderate."

Considering the rapidly evolving nature of threats related to Consumer Protection and Fraud and the need to continually assess and enhance detection techniques, LinkedIn continues to prioritize the development and enhancement of strategies to further manage and mitigate risks related to Consumer Protection and Fraud, aiming to maintain a high level of consumer safety on the platform. The key mitigations currently implemented specific to Consumer Protection and Fraud are described in the subsequent paragraphs of this section.

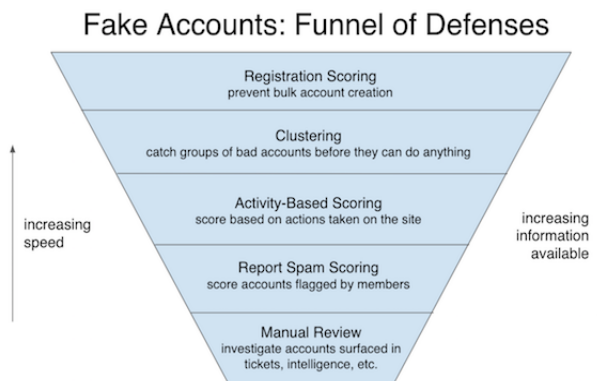
LinkedIn prohibits Profiles with false information including fake identities, misleading or deceptive information regarding business, qualifications, work experience, affiliations, or achievements. Members must use a Profile

photo that accurately reflects their likeness, and not associate with businesses that are not relevant to their professional associations. Creating or using another member’s account or Profile is not allowed, and sharing LinkedIn accounts is prohibited. Scamming, defrauding, or deceiving others on the platform are strictly prohibited, as is promoting pyramid schemes or romance scams, sharing malicious software, or engaging in phishing attempts.

LinkedIn has a dedicated Account Abuse team that works to action LinkedIn’s impersonation defense strategy, which encompasses verifying Profiles, performing a name to title to company match, tracking of member IP addresses, and leveraging member feedback to identify fake accounts at scale. LinkedIn applies a multi-layer approach to detecting fake accounts on platform.

- First, to proactively stop fake accounts at scale during registration, LinkedIn uses machine learning models to detect groups of accounts that look or act similarly. Every new user registration attempt is evaluated by a machine-learned model that gives an abuse risk score. Signup attempts with a low abuse risk score are allowed to register right away, while attempts with a high abuse risk score are prevented from creating an account. Attempts with medium risk scores are challenged by LinkedIn’s security measures to verify that they are real people.
- Second, LinkedIn creates clusters of accounts by grouping them based on shared attributes. LinkedIn then finds account clusters that show a statistically abnormal distribution, scores the clusters with supervised machine learning models, and then actions accounts in those clusters.
- Third, LinkedIn looks for activity-based models that either look for specific types of bad behavior typical of abusive accounts (e.g., excessive connection or invitations sent) or behavior that is anomalous.
- Fourth, LinkedIn gets signals from its members. Members give valuable information by reporting previously undetected accounts that can then go through additional model scoring and review.
- Finally, LinkedIn manually reviews accounts based on internal findings, peer leads, and member flags.

Figure 15



LinkedIn has also implemented a variety of verification features, as described in [New Since Last Year](#). To help these in-product authenticity signals serve their purpose (i.e., as user empowerment tools to inform user engagement on the platform), LinkedIn clearly marks verifications associated with Profiles, Pages, or Jobs with a verification badge icon and corresponding transparency about what such verifications means. This design and transparency aims to reduce the risk that users overly rely on or otherwise misinterpret verifications. LinkedIn also takes various steps to help prevent abuse of LinkedIn’s verification features. For example, in the EU,



members who verify their identity through LinkedIn's partner Persona must use certain hard-to-falsify documents like E-passports. In jurisdictions where it is legally allowed, LinkedIn's identity verification partners independently use liveness checks to confirm that the picture in the ID document submitted to the partner matches the member's face.

LinkedIn also offers several other features to protect its members from inauthentic accounts. Members can choose to only receive communications from people they know, thus allowing them to control who can contact them. The "About this Profile" feature shows members when a Profile was created, when the member's contact information and Profile photo were last updated, and if the member has verifications associated with their profile (like a workplace or identity verification).

For messaging, members can further protect themselves by opting into settings that detect harmful content in messaging (e.g., crypto scams), as well as limit the people who can attempt to connect with or message them. Once enabled through enhanced safety features in member account settings, LinkedIn also displays safety reminders to members if attempts to move a private conversation away from the platform are detected in their incoming messages. The member can click through to view or report the message. This flow is designed to empower and educate members to make informed safety choices before they move certain conversations off platform.

For Jobs, LinkedIn has a range of mechanisms in place to help protect job seekers from fraud and scams.

- LinkedIn has implemented anti-abuse models and processes that utilize both machines and humans to identify, review, and restrict fraudulent job posts.
- LinkedIn has integrated verified information into job posts, which helps job seekers assess the authenticity and origin of a job they are considering applying to and helps make attempted fraud less effective. By clicking on a verified information badge on a job post, members can view verified information about the job (e.g., whether the job poster has verified a work email address at the hiring company, how long ago the company Page for the job post was created, how long ago the job poster's account was created). Today, more than 20 percent of all job postings on LinkedIn have a verified information badge, and LinkedIn is committed to continue growing that number.
- LinkedIn has implemented tools to help companies protect their LinkedIn Pages from fraudulent and unauthorized jobs posted by bad actors. Job posts on LinkedIn must be associated with a LinkedIn Page. LinkedIn allows companies to limit those that can post jobs to their Pages to members at the company's email domain.
- LinkedIn has repeat offender systems to limit the ability of bad actors to repeatedly post jobs that violate LinkedIn's job policies. LinkedIn has implemented these systems at both the job poster level and at the Page level.
- In addition, to help protect job seekers on LinkedIn, LinkedIn has implemented in-product safety tips to help educate job seekers during their job search regarding how to identify signs of a job scam and protect themselves.

## **Human Dignity**

Risks related to Human Dignity include exposure to content pertaining to sexual exploitation, pornography, human trafficking, vulgarity, and gore. Absent sufficient mitigations, risks related to Human Dignity may manifest on the platform in ways such as members attempting to share content in Feed or messages that is violent, vulgar, or sexually explicit, or job posters attempting to post a job that facilitates or perpetrates human trafficking.

The **Inherent Probability** of systemic risks concerning Human Dignity stemming from the function, use, or misuse of LinkedIn's services is assessed as "Likely." While the overall market demand for mature content and the presence of violence and gore in popular media in Europe is high, there is a lower likelihood of members sharing such content on LinkedIn due to its professional nature. The LinkedIn platform also does not allow for rapid and wide dissemination of content that has not been selected for quality, which limits systemic spread of harmful content. Content shared by members is typically only available to members within their network, which, for any inauthentic accounts that may slip past LinkedIn's proactive defenses, is likely to be small. Based on the relative frequency of content or activities related to Human Dignity appearing in LinkedIn's content moderation systems as well as public discourse related to LinkedIn, the probability is assessed as "Likely" that users could share or access information or perpetrate activities negatively impacting Human Dignity on LinkedIn's platform absent sufficient mitigations.

The **Inherent Severity** of impact for content and activities that negatively impact Human Dignity is rated as "Critical" primarily due to the gravity of risks within this category, considering the potential for significant harm to wellbeing, societal, economic, and security systems at the individual and potentially broader level. Some risks within this category, such as human trafficking and prostitution, are potentially irremediable. Other risks, such as gore, vulgarity, profanity or pornographic content, may be remediable but can be significant.

With the rating for Inherent Probability as "Likely" and Inherent Severity as "Critical," the **Inherent Risk** associated with negative impacts to Human Dignity across LinkedIn products and features is therefore assessed as "High."

In addition to the DTSP Best Practices LinkedIn has effectively implemented across all Risk Areas, which are further described under "[Mitigations](#)," LinkedIn has implemented mitigations specific to Human Dignity following industry best practices with defined and documented processes. As such, the **maturity of mitigation** efforts LinkedIn has applied to the Human Dignity risk area is assessed as "Defined," which brings the **Residual Risk** rating down to "Moderate."

Recognizing that the Residual Risk is not "Low," LinkedIn continues to prioritize the development and enhancement of strategies to further manage and mitigate risks related to Human Dignity, aiming for an even safer user experience. The key mitigations currently implemented specific to Human Dignity are described in the subsequent paragraphs of this section.

LinkedIn's policies are set forth in its User Agreement, Privacy Policy, Professional Community Policies, Jobs Terms & Conditions, Jobs Policies and Ads Policies. Importantly, the Professional Community Policies clearly specify what members can and cannot do on LinkedIn. LinkedIn's Ads and Jobs Policies provide heightened requirements for advertising and job posts on LinkedIn. LinkedIn encourages civil conversation about all topics, including those touching on or relevant to Human Dignity; however, LinkedIn does not allow that conversation to be expressed in an abusive manner. LinkedIn prohibits engagement in sexual innuendos or unwanted advances and the sharing of harmful or shocking material. Content that facilitates prostitution and escort services is not allowed on the platform. LinkedIn's Ads policies prohibit ads from being offensive to good taste. This means ads must not be, for example, hateful, vulgar, sexually suggestive, or violent. All ads on LinkedIn must comply with LinkedIn's Professional Community Policies. LinkedIn also does not allow ads that are inappropriate during the time of or in relation to a tragedy, disaster, or similarly sensitive event. LinkedIn removes content that violates the policies and only makes content that meets a higher standard of quality and fits the professional focus of the platform eligible for broader distribution. LinkedIn removes depictions of extreme violence, along with image or video content that is excessively gory, gruesome, or disturbingly shocking. LinkedIn will also warn members when photos or videos that may contain sensitive or graphic content are detected. The content will initially be blurred with a warning, but members can decide if they would like to view it.

LinkedIn has internally developed classifiers with high levels of precision and recall that are trained on a set of confirmed policy-violating content identified by human reviewers used to proactively enqueue potentially policy-violating content for human review. In addition to proactive review, it is straightforward for members to flag potentially policy-violating content for human review. Content review is conducted by a global team of language and policy enforcement experts that adhere to strict limits with respect to the time in which the review must be completed. In addition, this team is subject to ongoing, mandatory training and is measured against established quality assurance standards.

LinkedIn's cross-functional Trust team, comprising engineers, product, data science, AI specialists, investigators, and legal teams, work together not only to detect and remove abusive content from LinkedIn, but also to educate LinkedIn members that unwanted sexual advances and other sexual harassment do not belong on LinkedIn. For example, LinkedIn provides guidance through Help Center articles with clear information and examples about the prohibition of sexual innuendos and unwanted advances on the platform, as well as guidance on how to report harassment or safety concerns. LinkedIn also warns members when photos or videos in Feed or Profiles are detected that may contain sensitive or graphic content. The content will initially be blurred with a warning, giving members control over whether they view this content, which can serve a newsworthy or educational purpose while complying with LinkedIn's policies.

European members can also opt in (through their message settings) to an optional advanced safety feature that allows LinkedIn's automated machine learning models to detect sexual harassment and certain other abusive content (beyond CSAM, malware, and phishing) in messages received by the member. When the setting is enabled and LinkedIn's automated systems detect potentially harmful content, depending on whether the sender and recipient have exchanged prior communications, the message will be directed to the recipient's spam folder or be hidden from the recipient by a warning. The warning can be dismissed, giving the recipient the ability to view and report the message to LinkedIn for review if desired. LinkedIn removes content that expresses physical or sexual attraction towards others, requests a romantic relationship, or subjects others to sexually suggestive or explicit content.

### *Risk Areas of Low Residual Risk*

Considering the probability, severity, and maturity of mitigations, LinkedIn assessed that eight of the assessed risk areas fall into a Residual Risk category of "Low:" Mental and Physical Well-being, Discrimination and Hate, Protection of Personal Data, Civic Discourse and Electoral Processes, Freedom of Expression and Information, Public Health, Public Security, and "Illegal" Content and Activities.

#### **Mental and Physical Wellbeing**

Risks to Mental and Physical Wellbeing include gender-based hate speech and violence, encouraging or assisting suicide, sexual harassment, stalking, encouraging coercive behavior, and addictive behaviors that may impact mental health. Absent sufficient mitigations, risks related to Mental and Physical Wellbeing may manifest on the platform in ways such as members attempting to share content in Feed or Messages that includes threats of violence, topics related to gender-based violence, or language encouraging or assisting suicide, or job postings that lead to harm to user physical or mental wellbeing.

The **Inherent Probability** of systemic risks concerning Mental and Physical Wellbeing stemming from the function, use, or misuse of LinkedIn is assessed as "Likely." There is a lower likelihood of members sharing content publicly on LinkedIn that would negatively impact Mental and Physical Wellbeing, considering the professional nature of the platform. Additionally, the LinkedIn platform favors knowledge sharing rather than virality, and therefore, tends to not facilitate the rapid and wide dissemination of content that does not meet a minimum quality bar, nor does it contain features that would promote addiction in members within its games or videos. However, considering the relative frequency of content or activities that may negatively impact Mental

and Physical Wellbeing appearing in LinkedIn's content moderation systems as well as public discourse related to LinkedIn, the probability is assessed as "Likely."

The **Inherent Severity** for content and activities that negatively impact Mental and Physical Wellbeing is rated as "Critical" primarily due to the gravity of risks within this category, considering the potential for significant harm to wellbeing, economic, and societal systems at the individual and potentially broader level. The impact of some risks within this category, such as violence against women, are irremediable. Other risks, such as behavioral addictions, may be considered remediable but still have the potential to cause significant damage to an individual's mental health or threaten their physical wellbeing.

With the rating for Inherent Probability as "Likely" and Inherent Severity assessed as "Critical," the **Inherent Risk** associated with negative impacts to Mental and Physical Wellbeing across LinkedIn products and features is therefore assessed as "High."

In addition to the DTSP Best Practices LinkedIn has effectively implemented across all Risk Areas, which are further described under "[Mitigations](#)," LinkedIn has implemented robust mitigations following industry best practices with defined, documented, and managed processes. As such, the **maturity of mitigation** efforts LinkedIn has applied to the Mental and Physical Wellbeing risk area is assessed as "Managed," which brings the **Residual Risk** rating down to "Low."

Considering the Critical Inherent Risk related to Mental and Physical Wellbeing, LinkedIn continues to prioritize the development and enhancement of strategies to further manage and mitigate risks related to Mental and Physical Wellbeing, aiming to maintain a high level of user safety on the platform. The key mitigations currently implemented specific to Mental and Physical Wellbeing are described in the subsequent paragraphs of this section.

As a professional network, LinkedIn maintains a high standard for safe and civil conversations between members. LinkedIn policies prohibit a wide range of actions and behaviors associated with harassment, bullying, or other abusive conduct. LinkedIn prohibits gender-based hate speech, symbols, and groups on the platform, and removes any content that attacks, denigrates, intimidates, dehumanizes, incites or threatens hatred, violence, prejudicial or discriminatory action against individuals or groups because of gender, gender identity, or sexual orientation. LinkedIn also removes content that personally attacks, intimidates, shames, bullies, disparages other LinkedIn members, or is otherwise abusive or uncivil. LinkedIn does not allow the sharing of harmful or shocking material. LinkedIn may permit content that evokes hateful rhetoric (including slurs) in the context of counter speech, reclamation, or members' personal experiences with sexism. LinkedIn also allows members to call out actual or perceived hateful, prejudicial, or discriminatory conduct, provided such content doesn't otherwise violate its policies. In connection with a newsworthy event, LinkedIn may label such content instead of removing it from the platform. LinkedIn will not penalize members for posting content in these circumstances but may take steps to limit its distribution to protect other members who may not wish to see it.

If a member is sharing dismissive or derisive content within the LinkedIn Feed - content that discourages, mocks, ridicules, makes fun of or refutes other individuals or groups or their self-identification or identity - this content may not be eligible for broad distribution on the platform. Additionally, if other members are expressing disinterest in, dismissing or reporting a certain piece of content, that content may not receive broad distribution on the platform, as this indicates members potentially find the content to be encouraging gender-based violence or otherwise unhelpful or unwanted.

LinkedIn has internally developed classifiers with high levels of precision and recall that are trained a set of confirmed policy-violating content identified by human reviewers used to proactively enqueue potentially policy-violating content for human review. In addition to proactive review, it is straightforward for members to flag potentially policy-violating content for human review. Content review is conducted by a global team of

language and policy enforcement experts that adhere to strict limits with respect to the time in which the review must be completed. In addition, this team is subject to ongoing, mandatory training and is measured against established quality assurance standards.

European members can opt in (through their message settings) to an optional advanced safety feature that allows LinkedIn's automated machine learning models to detect sexual harassment and certain other abusive content (beyond CSAM, malware, and phishing) in messages received by the member. Where the setting is enabled and LinkedIn's automated systems detect potentially harmful content, depending on whether the sender and recipient have exchanged prior communications, the message will be directed to the recipient's spam folder or be hidden from the recipient by a warning. The warning can be dismissed, giving the recipient the ability to view and report the message to LinkedIn for review if desired. LinkedIn removes content that expresses physical or sexual attraction towards others, requests a romantic relationship, or subjects others to sexually suggestive or explicit content.

LinkedIn also warns members when photos or videos in Feed or Profiles are detected that may contain sensitive or graphic content. The content will initially be blurred with a warning, giving members control over whether they view this content, which can serve a newsworthy or educational purpose while complying with LinkedIn's policies.

Additionally, LinkedIn empowers members to control how their information is viewed by others and how others interact with them. For example, members can:

- Control how their Profile is shown on LinkedIn, such as on search engines;
- Control who can send them an invitation to connect on LinkedIn (e.g., members can choose to receive connection invitations only from people who already know their email address);
- Disallow non-connections from sending them messages (to avoid unwanted contact);
- Block particular members by using the "block" feature (see more [here](#));
- Restrict the visibility of their email address, list of connections, and last name; and
- Disallow others from finding their LinkedIn Profile even if the searching person knows the member's email address or phone number.

### **Discrimination and Hate**

Risks related to Discrimination and Hate include exposure to discriminatory content or activities on the platform, including hate speech, and potential bias and discrimination in LinkedIn's algorithmic systems. Absent sufficient mitigations, risks related to Discrimination and Hate may manifest on the platform in ways such as members attempting to share content in Feed that contains hate speech, job posters attempting to post a job that discriminates against certain races or genders, or LinkedIn's recommender systems recommending candidates to recruiters based on bias in algorithms.

The **Inherent Probability** of systemic risks concerning Discrimination and Hate stemming from the function, use, or misuse of LinkedIn's services is assessed as "Likely." There is a lower likelihood of members sharing content publicly on LinkedIn related to Discrimination and Hate, considering the professional nature of the platform. Additionally, the nature of the LinkedIn platform favors knowledge sharing rather than virality, and therefore, tends to not facilitate the rapid and wide dissemination of content that does not meet a minimum quality bar. In turn, this limits the potential for discriminatory or hateful content to spread in a viral nature on the platform. Absent sufficient mitigations, LinkedIn's recommender systems could operate with bias or make recommendations with discriminatory implications. Considering the relative frequency of content or activities related to Discrimination and Hate appearing in LinkedIn's content moderation systems as well as public discourse related to LinkedIn, the probability is assessed as "Likely."

The **Inherent Severity** for content and activities that negatively impact Discrimination and Hate on LinkedIn is rated as “High” primarily due to the gravity of risks within this category, considering the impact to wellbeing, societal, and economic impacts at the individual level and societal impacts up to a regional scale. Risks within this category, such as hate speech and discrimination, may be remediable but can cause serious damage, including loss of career advancement opportunities.

With the rating for Inherent Probability as “Likely” and Inherent Severity as “High,” the **Inherent Risk** associated with Discrimination and Hate across LinkedIn products and features is therefore assessed as “High.”

In addition to the DTSP Best Practices LinkedIn has effectively implemented across all Risk Areas, which are further described under “[Mitigations](#),” LinkedIn has implemented robust mitigations following industry best practices with defined, documented, and managed processes. As such, the **maturity of mitigation** efforts LinkedIn has applied to the Discrimination and Hate risk area is assessed as “Managed,” which brings the **Residual Risk** rating down to “Low.”

LinkedIn has invested resources into ensuring "two members who are equally qualified should have equal access to opportunity" on LinkedIn, as well as limiting hate speech on the platform. LinkedIn continues to prioritize the development and enhancement of strategies to further manage and mitigate risks related to Discrimination and Hate on the platform. The key mitigations currently implemented specific to Discrimination and Hate are described in the subsequent paragraphs of this section.

LinkedIn policy prohibits hate speech, symbols, and groups on the platform and removes any content that attacks, denigrates, intimidates, dehumanizes, and incites or threatens hatred, violence, prejudicial or discriminatory action against individuals or groups because of their actual or perceived race, ethnicity, national origin, caste, gender, gender identity, sexual orientation, religious affiliation, or disability status. LinkedIn also removes content that personally attacks, intimidates, shames, bullies, or disparages other LinkedIn members, or is otherwise abusive or uncivil. Additionally, members may not engage in trolling or other repetitive negative content that disrupts other members or conversations on the platform.

LinkedIn developed classifiers with high levels of precision and recall that are trained on a set of confirmed policy-violating content identified by human reviewers. Such classifiers are used to proactively enqueue potentially policy-violating content for human review. In addition to proactive review, it is straightforward for members to flag potentially policy-violating content for human review. Content review is conducted by a global team of language and policy enforcement experts that adhere to strict limits with respect to the time in which the review must be completed. In addition, this team is subject to exacting training and quality assurance standards.

If a member is sharing dismissive or divisive content within Feed - content that discourages, mocks, ridicules, makes fun of or refutes other individuals or groups or their views, self-identification, or identity - this content will not be eligible for broad distribution on the platform and may be removed. LinkedIn allows members to report actual or perceived hateful, prejudicial, or discriminatory conduct, provided such content doesn't otherwise violate its policies. Additionally, if other members are expressing disinterest in, dismissing or reporting a certain piece of content, that content may not receive broad distribution on the platform, as this indicates members potentially find the content to be discriminatory or otherwise unhelpful or unwanted.

For Jobs, LinkedIn prohibits discrimination in job posts based on protected characteristics, including age, gender, gender identity, religion, ethnicity, race, national origin, disability, sexual orientation, and any other basis protected under law; however, in a jurisdiction where it is legally permissible to do so, LinkedIn may permit language in posts expressing preference for members of certain groups historically disadvantaged in hiring in that location. LinkedIn automatically reviews job postings created by members for potential discrimination and flags potentially violative jobs for human review. If a decision is made that a specific job posting is

discriminatory, LinkedIn searches for and removes other copies of the job posting on the system. Members can also report for review by LinkedIn any job post on LinkedIn that they believe may be discriminatory.

In-product functionality is also utilized within Jobs to deter member posting of discriminatory content, such as an inclusivity nudge that brings potentially discriminatory language in a draft job post to a poster's attention before submission, so they can avoid posting jobs that run afoul of LinkedIn's Jobs Policies. In addition, Representative Results in Recruiter Search seeks to ensure that the results displayed in the first five pages of a Recruiter search contain the same binary gender distribution as the entire qualified search population. Products like Diversity Nudges suggest changes to search parameters to facilitate more equal gender balance in Recruiter search results. Filters on Recruiter do not allow customers to search for members along any protected characteristics. Customers also have the ability to prevent and avoid their own unintended bias by enabling in-product functionality like Hide Candidate Photos and Names. Additionally, all jobs can only be explicitly targeted by LinkedIn customers along very limited criteria – general region and skills. All jobs are visible to all members, regardless of targeting criteria identified by the job poster for promotion. In other words, all members could still find a particular job post on LinkedIn even if it was not specifically promoted to them (because of skills or location).

Specifically for Collaborative Articles posted in Pages, LinkedIn considers a variety of factors to invite an inclusive set of contributors to the articles. For instance, considering gender representation of the expert member collaborators that are writing the articles.

If an advertiser attempts to target an ad based on age or gender, LinkedIn first requires the advertiser to certify that they will not use LinkedIn to discriminate based on age, gender, or other protected characteristics in ads relating to employment, housing, education, and credit, and provides a link to a Help Center article detailing LinkedIn's policy. LinkedIn also has specific policies preventing ad targeting discrimination, as well as policies applicable to particular types of ads products on LinkedIn. For example, terms of use for Lead Gen Forms prohibit the collection of sensitive data. LinkedIn also prohibits the use of Insight Tag on Pages collecting specific health-related or financial related webpages (to protect against the inadvertent collection of data that could be used to facilitate discrimination). LinkedIn does not allow targeting on protected characteristics other than age and gender, and only allows ad targeting based on age or gender with the advertiser first providing this certification. To protect against abuse through exclusion, customers are not able to exclude ads based on age or gender. Features like Lookalike Audiences do not use sensitive demographic attributes, including attributes such as sex or age within the model.

European members can opt in (through their message settings) to an optional advanced safety feature that allows LinkedIn's automated machine learning models to detect sexual harassment and certain other abusive content (beyond CSAM, malware, and phishing) in messages received by the member. Where the setting is enabled and LinkedIn's automated systems detect potentially harmful content, depending on whether the sender and recipient have exchanged prior communications, the message will be directed to the recipient's spam folder or be hidden from the recipient by a warning. The warning can be dismissed, giving the recipient the ability to view and report the message to LinkedIn for review if desired. LinkedIn removes content that expresses physical or sexual attraction towards others, requests a romantic relationship, or subjects others to sexually suggestive or explicit content.

LinkedIn is also committed to measuring and preventing bias and discrimination in its algorithmic systems. Recommender systems including the LinkedIn Feed are assessed for unintended bias. Currently, this is only done across inferred binary gender. LinkedIn's approach to Responsible AI is forward leaning. LinkedIn examines the ethics and compliance of its AI-powered products and services as part of the Trustworthy Design Review (TDR) process, which it applies to LinkedIn product initiatives prior to launch. LinkedIn has also developed and open-sourced the LinkedIn Fairness Toolkit (LiFT), which makes use of common fairness definitions to enable the



measurement of fairness in large-scale machine learning workflows. LiFT can be deployed in training and scoring workflows to measure biases in training data, evaluate different fairness notions for ML models, and detect statistically significant differences in model performance across different subgroup LinkedIn conducts Fairness Checks on Feed distribution and ranking. LinkedIn evaluates its algorithms for binary gender bias, relying on a combination of member provided gender data (through [Self-ID](#)) and binary gender inference.

LinkedIn also conducts Fairness Checks on Job recommendations to ensure jobs are shown equally across demographics based on available data. LinkedIn conducts A/B testing of algorithmic systems and leverages long term holdouts for portions of the LinkedIn population to identify any potential undesirable impacts resulting from changes to algorithmic systems. LinkedIn also monitors for unexpected anomalies in distribution of content to track and avoid the introduction of bias into LinkedIn's algorithms. To prevent job discrimination, LinkedIn launched a Discrimination Prevention Model across all job types (including online, field, and ingested jobs) to improve model precision and recall. Now more discriminatory jobs will be detected and prevented, resulting in reduced harm to seekers (abusive views on discriminatory jobs) and more equitable hiring marketplace outcomes. LinkedIn employs repeat offender systems so that members that repeatedly post policy-violating content face escalating consequences, including potential permanent restriction from LinkedIn.

### **Protection of Personal Data**

Risks to Protection of Personal Data include the collection, processing, mishandling, or disclosure of user data without a lawful basis, including data mining, profile scraping, phishing, hacking, malware, and data breaches. Absent sufficient mitigations, risks related to Protection of Personal data may manifest on the platform in ways such as LinkedIn inappropriately using personal data without a lawful basis, inadequate transparency in how LinkedIn processes and stores user data, or improper leakage of data to the public.

The **Inherent Probability** of systemic risks concerning Protection of Personal Data is assessed as "Likely." Although a high level of sophistication is required to perpetrate harms such as hacking, malware, or data breaches, as well as internal misuse of member data, the financial incentives of fraudulent activities lead to a higher demand for theft of personal data. Considering the relative frequency of content or activities related to Human Dignity appearing in LinkedIn's content moderation systems as well as public discourse related to LinkedIn, the probability is assessed as "Likely."

The **Inherent Severity** for content and activities that negatively impact Protection of Personal Data is rated as "High" primarily due to the gravity of risks within this category, considering the potential for significant harm to economic, security, and societal systems at the individual and local level. Some risks within this category – particularly related to phishing, hacking, malware, or data breaches – may reach a broader scale of impact while others, such as collection of data without consent, may only impact the individual. For the most part, these risks are remediable, but may include some irremediable damage.

With the rating for Inherent Probability as "Likely" and Inherent Severity as "High," the **Inherent Risk** associated with Protection of Personal Data across LinkedIn products and features is therefore assessed as "High."

In addition to the DTSP Best Practices LinkedIn has effectively implemented across all Risk Areas, which are further described under "[Mitigations](#)," LinkedIn has implemented robust mitigations specific to the Protection of Personal Data following industry best practices with defined, documented, and managed processes. As such, the **maturity of mitigation** efforts LinkedIn has applied to the Protection of Personal Data risk area is assessed as "Managed," which brings the **Residual Risk rating** down to "Low."

LinkedIn takes the responsibility of protecting user data very seriously and is actively working to develop and enhance strategies to further manage and mitigate risks related to Protection of Personal Data, aiming for even greater protections for users. LinkedIn continues to invest in policies, policy enforcement mechanisms, and advanced detection capabilities, including capabilities to prevent Profile scraping and address other evolving



threats related to user privacy. The key mitigations currently implemented specific to Protection of Personal Data are described in the subsequent paragraphs of this section.

LinkedIn's core value is "members first." That value guides LinkedIn as it aims to provide clarity to platform members, act consistently, and give its members control. That means LinkedIn aims to be transparent in its policies and offers its members settings to help them control the data they provide on the platform. LinkedIn is compliant with the General Data Protection Regulation (GDPR) and allows users worldwide to access, correct errors in, erase, object to the processing of, and export their personal data. LinkedIn also requires its customers, suppliers, data processors, and subprocessors to enter into Data Protection Agreements ("DPAs") that contractually limit how the customer must handle data. The DPAs require data to be handled in accordance with GDPR and other applicable laws and include EU approved Standard Contractual Clauses which permit the transfer of data outside of the EU. Members also can exercise their data subject rights by accessing, objecting to and managing LinkedIn's use of their data or they can delete it from LinkedIn. LinkedIn has appointed a Data Protection Officer (DPO) to ensure that its processing of personal data is in compliance with GDPR. The DPO is the 'voice of the member' within LinkedIn. He makes recommendations in relation to new and existing data processing operations and handles complaints relating to the processing of personal data.

LinkedIn's Privacy Policy, information on [privacy settings](#), privacy FAQs, and region-specific privacy information is all made available to its users on the LinkedIn Privacy Hub and in the Help Center. LinkedIn uses a variety of in-product indicators to enhance member awareness and empowerment, including "Why am I seeing this ad?" (WAISTA).

Trust is an operating priority at LinkedIn that includes Safety, Privacy, and Security and is reported on at regular company-wide meetings and directly to the Chief Executive Officer. LinkedIn has invested in privacy protections and privacy enhancing technologies through the establishment of a Data Privacy Engineering team that works in line with the Privacy Legal team. A Privacy Impact Assessment is built into the review process for initiatives or features assessed as potentially likely to include personal data.

LinkedIn has robust practices to safeguard and cultivate a culture of respect for user privacy. LinkedIn has its own internal Data Privacy Principles for data processing across the business and maintains an InfoSec standard which support the proper handling of data within LinkedIn's data ecosystem. LinkedIn maintains a robust global information security team that protects its infrastructure and member data contained therein. This team is comprised of world-class technologists and investigators that ensure that anomalies (identified through automation) are investigated to ensure that external actors cannot access LinkedIn's internal systems. LinkedIn requires annual privacy and trust training for all employees, and regularly meets and engages with key data protection regulators to discuss privacy concerns and emerging trends.

Data processing operations involving personal data go through LinkedIn's Trustworthy Design Review (TDR) process. This process assesses the risk of each processing activity to ensure that all risks are assessed and documented and that appropriate mitigations are put in place to protect members' personal data. For higher-risk activities that involve the processing of personal data, LinkedIn completes Data Protection Impact Assessments ("DPIAs"). DPIAs document the type of data collected (data elements and whether sensitive data is included), risks associated with the processing (including impact to the fundamental rights and freedoms of data subjects), and mitigations that have been implemented to reduce such risks. Once completed, LinkedIn's Data Privacy Officer (DPO) reviews the DPIA, asks independent questions to ensure that the activity is in line with GDPR and member expectations, provides recommendations, and signs off once the risk is deemed to be acceptable.

To boost account security for members and Job posters, LinkedIn encourages the enablement of two-factor authentication (2FA). Clicking "Secure account" will redirect members to the settings page to complete enrollment.

To protect members' privacy, LinkedIn prohibits the use of third-party software or browser extensions that scrape, modify the appearance of, or automate activity on LinkedIn's website.

LinkedIn also employs a robust set of mitigations to identify and address external threats to the Protection of Personal Data, which have been described in further detail in the [Consumer Protection and Fraud](#) section of this Report. Most notably, LinkedIn reviews public-facing content for harmful or abusive indicators and flags potential violative content for review. It scans both public- and privately-facing content for phishing and malware that could offer bad actors access to member data.

### **Civic Discourse and Electoral Processes**

Risks to Civic Discourse and Electoral Processes include election misinformation, disinformation, coordinated manipulation and inauthentic activity, foreign interference, and the disruption of electoral processes, as well as political polarization, political bias, and the creation of echo chambers. Absent sufficient mitigations, risks related to Civic Discourse and Electoral Processes may manifest on the platform in ways such as members posting inaccurate or misleading information on electoral processes to Feed, advertisements hyper-targeting vulnerable groups with election misinformation, or recommender systems recommending political content in a biased manner or a manner that contributes to political polarization or the creation of echo chambers.

The **Inherent Probability** of systemic risks concerning Civic Discourse and Electoral Processes stemming from the function of LinkedIn's services is assessed as "Likely." There is a lower likelihood of members sharing content that could negatively impact Civic Discourse and Electoral Processes on LinkedIn due to its professional nature. There has been a high number of elections in Europe and globally during the assessment period, which increases the likelihood of bad actors attempting to negatively impact electoral integrity. The LinkedIn platform favors knowledge sharing rather than virality, and therefore, tends to not facilitate the rapid and wide dissemination of content that does not meet a minimum quality bar. In turn, this limits the systemic risk posed by the spread of harmful content. LinkedIn also does not allow political advertisements on the platform. Considering these factors and the relative frequency of content or activities related to Civic Discourse and Electoral Processes appearing in LinkedIn's content moderation systems as well as public discourse related to LinkedIn, the probability is assessed as "Likely" that users could share or be exposed to information negatively impacting Civic Discourse and Electoral Processes on the platform.

The **Inherent Severity** for content and activities that negatively impact Civic Discourse and Electoral Processes on LinkedIn is rated as "Critical" primarily due to the gravity of risks within this category, considering the potential for significant harm to political, societal, economic, and security systems at the local, country, and even regional levels. Some risks within this category have potentially irremediable consequences on political or security systems. The impact of other risks may be considered remediable but still have the potential to cause significant damage.

With the rating for Inherent Probability as "Likely" and Inherent Severity as "Critical," the **Inherent Risk** associated with negative impacts to Civic Discourse and Electoral Processes across LinkedIn products and features is therefore assessed as "High."

In addition to the DTSP Best Practices LinkedIn has effectively implemented across all Risk Areas, which are further described under "Mitigations," LinkedIn has implemented robust mitigations specific to Civic Discourse and Electoral Processes, following industry best practices with defined, documented, and managed processes. LinkedIn also considered the guidance provided by the European Commission (Commission 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 Article 35(3) of Regulation (EU) 2022/2065) in evaluating the maturity of mitigations for this Risk Area. While many of LinkedIn's existing practices aligned with the guidelines, LinkedIn made additional enhancements during this review period to risk mitigations for election integrity consistent with the referenced guidelines. As such, the **maturity of mitigation** efforts LinkedIn has applied to the Civic Discourse and Electoral Processes risk area is assessed as "Managed," which brings the **Residual Risk** rating down to "Low."

Considering the unprecedented number of elections occurring around the globe this year and the Critical nature of this risk, LinkedIn continues to prioritize the development and enhancement of strategies to further manage and mitigate risks related to Civic Discourse and Electoral Processes. The key mitigations currently implemented specific to this risk, which align with the European Commission Guidelines, are described in the subsequent paragraphs of this section.

LinkedIn conducted a global study to understand how the LinkedIn platform is used relative to electoral processes and to identify areas for additional focus or support. This analysis considered the presence and activity of political actors on the service; the likely use of the LinkedIn platform for election-related discourse; the number of LinkedIn members in a Member State; previously identified Tactics, Techniques, and Procedures (TTPs) for information manipulation in the context of elections; relevant details on parties or candidates, party programs, manifestos, or other political material; and relevant organizing events such as demonstrations or rallies, campaigning, fundraising, or other related political activities. From this analysis, LinkedIn concluded that users tend to search for, share, and / or access very little election-related content on LinkedIn due to its nature as a professional platform. Nevertheless, LinkedIn took a variety of measures to strengthen its management and mitigation of election-related risks.

LinkedIn has a dedicated Civic Integrity team that worked to operationalize the additional guidance set forth from the EC around election integrity. The Civic Integrity team is global and cross-functional, with representation across the organization including the Legal, Public Policy, Content Policy, Communications, Escalations, Enforcement Operations, and Product teams, as well as the DSA Compliance Function.

LinkedIn developed a quantitative and qualitative approach to assessing the priority of global elections in order to adequately and consistently stand up enhanced internal coverage of election risk based on relative impact. This quantitative and qualitative assessment is used to determine the length and severity of the risk mitigation measures that need to be implemented for each high priority election, including the language resourcing requirements. Depending on the election, risk mitigation measures at LinkedIn are in place and functioning one to six months before an electoral period and can continue for one month after the elections. As a result of this analysis, LinkedIn stood up a cross-functional tiger team in support of the EU Parliamentary elections including individuals with country-specific expertise.

In addition to the tiger teams, LinkedIn has developed incident response plans and escalation paths to address potential crisis events. During this period, LinkedIn hosted tabletop exercises to pressure test its election-related procedures and plans under various scenarios. LinkedIn worked cross functionally to run workshops and provide internal training to help ensure team members showcase a clear understanding of the proactive and reactive response and mitigation strategies available for many potential election-related scenarios. LinkedIn employs a "follow the sun" model for crisis events and has created a Follow the Sun Escalation directory and provides training on how to leverage the directory. LinkedIn engages in the appropriate encompassing bodies to address the exchange of election-related information including the Rapid Response Group, Europol, and other pan-EU bodies.

LinkedIn encourages civil conversation about all topics on the platform, including topics touching on Civic Discourse and Electoral Processes. However, LinkedIn does not allow that conversation to be expressed in a hateful or bullying manner. LinkedIn does not allow users to share false or misleading content or content to interfere with or improperly influence an election or other civic process. LinkedIn also prohibits political advertisements on its platform, including ads advocating for or against a candidate, party, or ballot proposition or ads designed for political fundraising or otherwise intended to influence an election outcome.

If potentially policy-violating content is posted, LinkedIn reviews that content for policy compliance. Potentially policy-violating content is identified for human review through classifiers trained through machine learning and member reports. Internally developed classifiers proactively enqueue potentially policy violating content for human review. In addition to proactive review, it is straightforward for members to flag potentially policy-violating content for human review. Content review is conducted by a global team of language and policy enforcement experts that adhere to strict limits with respect to the time in which the review must be completed. The review team is subject to exacting training and quality assurance standards.


Over this review period, LinkedIn invested in its elections metrics framework to more efficiently identify election-related content. LinkedIn monitors conversation via keywords in relevant elections during high priority elections to identify violations or when operational capacity breaches a threshold and requires additional support. LinkedIn established a Multi-Level Review Protocol for election-related content. Public content on LinkedIn is scanned for election-related information and undergoes additional human review as needed to ensure a high level of election integrity. LinkedIn works with external fact checkers and leverages a curated repository of reputable global fact-checking resources across multiple priority languages and localities to support accurate review of election-related content.

LinkedIn aims to educate its members about civic discourse, electoral processes, and public security through its global team of news editors. LinkedIn has provided a [Help Center article](#) to support media literacy, and LinkedIn has an internal team of global news editors that provides trustworthy and authoritative content to its member-base at all times.

Figure 16



## European elections: news and resources

 **Pieter Cranenbroek**  
Senior News Editor at LinkedIn  
Published Apr 18, 2024

+ Follow

Citizens in 27 EU member states voted over the past four days to elect the European Parliament. Results are still being counted, after the last polling stations closed at 11pm CET on Sunday. French President Emmanuel Macron **dissolved French parliament** and called snap elections after exit polls showed Marine Le Pen's National Rally were set to beat Macron's party. Voters in the 27 EU countries will elect 720 MEPs, 15 more than in previous elections.

You can find more information on the European Parliament elections on the EU's official website.

During the pre-election cycle, LinkedIn relies on trusted and reputable publisher sources for featured shares, focusing on the policy impact on businesses and professionals around the EU. LinkedIn also curates links to topical landing pages from trusted publishers to provide members with easy and reliable entry points to more detailed coverage. LinkedIn does not compete with trusted publishers for speed or depth of coverage, but instead aims to connect their existing coverage to LinkedIn members and their needs. During the European Elections, this team provided manually curated and localized storylines.

LinkedIn has a Nation State Threat program to account for state-sponsored malign actor relationships to content on the platform. LinkedIn's team of threat investigators and intelligence analysts works with peers and other stakeholders, including the AI modeling team, to identify and remove nation-state actors and coordinated inauthentic campaigns. LinkedIn conducts investigations into election-related influence operations and nation-state targeting including continued information sharing on threats with industry peers and Law Enforcement on a regular basis. LinkedIn works with peer companies and other stakeholders to receive and share indicators related to fake accounts created by state-sponsored actors, such as confirmed Tactics, Techniques, and Protocols (TTPs) and Indicators of Compromise (IOC). This exchange of information leads to a better understanding of the incentives of sophisticated and well-funded threat actors and how they evolve their TTPs to achieve those goals, which assists LinkedIn in their identification and removal. LinkedIn's teams also work closely with Microsoft's Threat Intelligence Center (MSTIC) and Democracy Forward teams on security issues.

To respond to the ever-changing threat landscape, LinkedIn's teams continually invest in new technologies for combating inauthentic behavior on the platform. LinkedIn committed to the Tech Accord to Combat Deceptive Use of AI in 2024 Elections. Aligned with this commitment, LinkedIn is investing in AI technologies such as advanced network algorithms that detect communities of fake accounts through similarities in their content and behavior; computer vision and natural language processing algorithms for detecting AI-generated elements in fake Profiles such as deep fakes; anomaly detection of risky behaviors; and deep learning models for detecting sequences of activity that are associated with abusive automation. LinkedIn has also adopted the Coalition for Content Provenance and Authenticity's industry-leading "Content Credentials" technology (C2PA) to include metadata labeling, including data about whether content is created using AI, on content containing the C2PA technology. Furthermore, LinkedIn acts vigilantly to maintain the integrity of all accounts and to ward off false account activity:

- Monitoring platform conversations regarding significant elections and establishing metrics for when election-related conversations, violations, or operational capacity breach a threshold and require additional support;
- Maintaining a dedicated Anti-Abuse team to research emerging trends and key risks and develop tools to address them;
- Using AI to detect inauthentic activity and communities of fake accounts;
- Conducting hash matching for known instances of deepfake content;
- Using automated systems to detect and block automated activity; and
- Maintaining 24/7 escalation paths to address any emerging issues.

### **Freedom of Expression and Information**

Risks to Freedom of Expression and Information include risk that freedom of expression, information, the press, or pluralism of the media are negatively impacted by the use, design, or functioning of a platform or its algorithmic systems, policies, or enforcement. Absent sufficient mitigations, risks related to Freedom of Expression and Information may manifest on the platform in ways such as LinkedIn's content moderation systems removing content that constitutes legally permissible speech, preventing legitimate jobs from being posted, or amplifying content from some but not all media viewpoints.

The **Inherent Probability** of systemic risks concerning Freedom of Expression and Information stemming from the use, misuse, or function of LinkedIn's services is assessed as "Likely." As a real-identity, specific-purpose platform, LinkedIn operates under standards of professionalism, which are reflected both in content policies and enforcement, as well as in content prioritization and amplification. LinkedIn strictly enforces both policies for removing content from the platform and for determining when content is eligible for broader distribution. Therefore, absent mitigations to ensure fair and appropriate application of these content moderation policies, the probability of negative impacts to user rights to Freedom of Expression and Information is assessed as "Likely."

The **Inherent Severity** for content and activities that negatively impact Freedom of Expression and Information on LinkedIn is rated as "High" primarily due to the gravity of risks within this category, considering the potential for significant harm to societal and wellbeing systems up to the country level and political systems up to the regional level. Depending on the local political environment, consequences related to the negative impact on Freedom of Expression and Information can be irremediable. But the impact of many risks within this category is not considered significant.

With the rating for Inherent Probability as "Likely" and Inherent Severity as "High," the **Inherent Risk** associated with negative impacts to Freedom of Expression and Information across LinkedIn products and features is therefore assessed as "High."

In addition to the DTSP Best Practices LinkedIn has effectively implemented across all Risk Areas, which are further described under "[Mitigations](#)," LinkedIn has implemented a robust set of mitigations following industry best practices with defined, documented, and managed processes. As such, the **maturity of mitigation** efforts LinkedIn has applied to the Freedom of Expression and Information risk area is assessed as "Managed," which brings the **Residual Risk rating** for this area down to "Low."

LinkedIn carefully weighs trade-offs between freedom of expression and online harms, erring on the side of freedom of expression where possible. Over the review period, authors appealed only about .6% of the content moderation decisions that LinkedIn made. LinkedIn granted roughly 24% of those appeals, representing .15% of moderation decisions. These numbers illustrate the priority LinkedIn places on freedom of expression considering 1) LinkedIn's receptivity to a meaningful percentage (24%) of appeal requests and 2) the very small percentage of total content moderation decisions overturned (.15%). That said, LinkedIn continues to invest in developing and enhancing strategies to further mitigate risks to Freedom of Expression and Information. The key mitigations currently implemented specific to Freedom of Expression and Information are described in the subsequent paragraphs of this section.

LinkedIn has internally developed classifiers with high levels of precision and recall that are trained on a set of confirmed policy-violating content identified by human reviewers are used to proactively enqueue potentially policy violating content for human review. Where content does not violate LinkedIn policies, subject to certain limited exceptions, LinkedIn generally will not remove or geo-block the visibility of content on LinkedIn without valid legal process requiring us to do so.

Content that would otherwise violate LinkedIn's policies may be allowed in cases where the content is being shared for awareness or to condemn. In these cases, LinkedIn may label and obscure the content for members who may find this content sensitive or disturbing, or otherwise do not want to view it. However, LinkedIn won't remove the content or penalize the author for posting it. When speech doesn't conclusively cross the boundaries set by LinkedIn policies, LinkedIn minimizes interference and gives the speaker the benefit of the doubt.

Content review is conducted by a global team of language and policy enforcement experts who adhere to strict limits with respect to the time in which the review must be completed. This team is subject to exacting training

and quality assurance standards, and LinkedIn has separate teams that administer these standards. LinkedIn has clear and well-documented policies and tracks edge cases for sensitive escalations to enhance or clarify policies as needed. LinkedIn employs a dedicated team of trainers, who not only support the onboarding of new content moderators, but also provide ongoing educational opportunities for all moderators. Moderators have direct access to global content policy managers through regular office hours and dedicated escalation pathways. For particularly complex decisions, content policy managers also have access to in-house lawyers who can consult regional legal experts as needed.

Users have access to report content on each LinkedIn surface. Both reporter and creator generally are notified when content is removed, and both are given an opportunity to appeal the decision.

LinkedIn responds to periodic evaluations by the third-party independent organization Ranking Digital Rights before it publishes its annual ratings on LinkedIn practices, governance, and leadership on the protection of freedom of expression and privacy. LinkedIn also engages with its parent company, Microsoft, in connection with Microsoft's commitments to human rights, including free expression.

LinkedIn carefully considers all government removal requests and works to mitigate any implications they may have on freedom of expression and human rights. For government demands, LinkedIn employs safeguards to ensure any actions taken are narrow, specific, submitted in writing, and based on valid legal orders. Through its parent company, Microsoft, LinkedIn also engages with broader civil society organizations on best practices related to government removal requests and participates in human rights impact assessments. LinkedIn generally provides prompt notice to the member(s) whose content is at issue in a government removal request, unless legally prohibited from doing so. This notice may enable the impacted member to challenge the request served on LinkedIn.

LinkedIn is also committed to measuring and preventing bias in its algorithmic systems that would negatively impact freedom of expression and pluralism. LinkedIn has developed and open-sourced the LinkedIn Fairness Toolkit (LiFT), which makes use of common fairness definitions to enable the measurement of fairness in large-scale machine learning workflows. LiFT can be deployed in training and scoring workflows to measure biases in training data, evaluate different fairness notions for ML models, and detect statistically significant differences in model performance across different subgroup LinkedIn conducts Fairness Checks on Feed distribution and ranking. LinkedIn evaluates its algorithms for binary gender bias, relying on a combination of member provided gender data (through [Self-ID](#)) and binary gender inference. LinkedIn also monitors for unexpected anomalies in distribution of content to track and ensure the introduction of bias into LinkedIn's algorithms.

## **Public Health**

Risks to Public Health include false information and the promotion of harmful, yet legal, substances or practices (e.g. weight loss medications, vaccine misinformation). Absent sufficient mitigations, risks related to Public Health may manifest on the platform in ways such as members posting misinformation related to health topics in Feed, News results including low quality information related to health events, or LinkedIn Ads including information on legal but harmful substances.

The **Inherent Probability** of systemic risks concerning Public Health stemming from the function, use, or misuse of LinkedIn's services is assessed as "Not Likely." The LinkedIn platform favors knowledge sharing rather than virality, and therefore, tends to not facilitate the rapid and wide dissemination of content that does not meet a minimum quality bar. In turn, this limits the potential for health misinformation to spread in a viral nature on the platform. Considering the relative frequency of content or activities related to Public Health appearing in LinkedIn's content moderation systems as well as public discourse related to LinkedIn, which has decreased this review period, the probability that members may share or be exposed to misleading health information is assessed as "Not Likely."



The **Inherent Severity** for content and activities that negatively impact Public Health on LinkedIn is rated as “Critical,” primarily due to the gravity of risks within this category, considering the potential for significant and irreparable harm to wellbeing, economic, and societal systems at the individual and up to global level. Some risks within this category such as health misinformation and exposure to harmful substances have potentially irreparable consequences.

With the rating for Inherent Probability as “Not Likely” and Inherent Severity as “Critical,” the **Inherent Risk** associated with the negative impacts to Public Health across LinkedIn products and features is therefore assessed as “Moderate.”

In addition to the DTSP Best Practices LinkedIn has effectively implemented across all risk areas, which are further described under “[Mitigations](#),” LinkedIn has implemented mitigations following industry best practices with defined and documented processes. As such, the **maturity of** mitigation efforts LinkedIn has applied to the Public Health risk area is assessed as “Defined,” which brings the **Residual Risk rating** down to “Low.”

Recognizing that the maturity score is not Managed, LinkedIn continues to prioritize the development and enhancement of strategies to further manage and mitigate risks related to Public Health, aiming for a safer user experience. The key mitigations currently implemented specific to Public Health are described in the subsequent paragraphs of this section.

LinkedIn has invested in new technologies to combat inauthentic behavior that may contribute to risk to Public Health and has robust policies and protocols in place around healthcare advertising and health misinformation.

LinkedIn’s Professional Community policies prohibit users from sharing content that directly contradicts guidance from leading global health organizations and public health authorities, including false information about the safety of efficacy of vaccines or medical treatments.

LinkedIn has policies related to healthcare advertisements that limit drug advertisements to approved retailers. Marketing Solutions’ policies only allow drug advertising in geographic jurisdictions where it is legal. LinkedIn Marketing prohibits the promotion of opioids, anabolic steroids, or other unsafe products on LinkedIn’s prohibited list. LinkedIn leverages a specialized vendor to certify pharmacy and telehealth service ads and limits their distribution to within the U.S. only. LinkedIn also uses classifiers to identify and prohibit content related to opioids, anabolic steroids, or other unsafe products on Feed and other surfaces.

LinkedIn policies prohibit direct-to-purchase links and/or private contact information to facilitate the sale of prescription drugs, pharmaceuticals, alcohol, and tobacco products. LinkedIn has developed classifiers to automatically detect this content or health misinformation on the platform and has also partnered with fact-checking organizations.

Regarding Media Literacy, LinkedIn does not prioritize any news sources in Feed, but in crisis situations, (e.g., COVID-19), search banners are used to point members to reputable sources of information (e.g., when members searched for “COVID,” LinkedIn pointed members to trustworthy information about that topic, including links to global health organizations).

### **Public Security**

Risk to Public Security includes risk that content or activities degrading public security occur on the platform including terrorist recruitment, funding, or training activities, terrorist imagery or content, or support for terrorists or glorification of terrorist acts as well as mis- or disinformation related to crisis events. Absent sufficient mitigations, risks related to Public Security may manifest on the platform in ways such as the members posted content glorifying terrorist acts in Feed, members posting a job that facilitates terrorist activities, or advertisers promoting content with misinformation on a public crisis event.



The **Inherent Probability** of systemic risks concerning Public Security stemming from the function, use, or misuse of LinkedIn's services is assessed as "Not Likely." While data analysis shows an increase in overall market demand for posting and consuming content related to world crisis events such as the Israel-Hamas war, there is a lower likelihood of members sharing such content on LinkedIn due to its professional nature. The LinkedIn platform also does not allow for rapid and wide dissemination of content that has not been selected for quality, which limits systemic spread of harmful content. Based on the relative frequency of content or activities related to Public Security appearing in LinkedIn's content moderation systems as well as public discourse related to LinkedIn, the probability is assessed as "Not Likely" that users would share or access information or perpetrate activities negatively impacting Public Security on the LinkedIn platform absent sufficient mitigations.

The **Inherent Severity** for content and activities that negatively impact Public Security on LinkedIn is rated as "Critical," primarily due to the gravity of risks within this category, considering the potential for significant harm to security, wellbeing, societal, political, and economic systems at the individual, country, and regional levels. The impact of many risks within this category are irremediable.

With the rating for Inherent Probability as "Not Likely" and Inherent Severity as "Critical," the **Inherent Risk** associated with negative impacts to Public Security across LinkedIn products and features is therefore assessed as "Moderate."

In addition to the DTSP Best Practices LinkedIn has effectively implemented across all Risk Areas, which are further described under "[Mitigations](#)," LinkedIn has implemented robust mitigations specific to Public Security following industry best practices with defined and documented processes. As such, the **maturity of mitigation** efforts LinkedIn has applied to the Public Security risk area is assessed as "Defined," which brings the **Residual Risk rating** down to "Low."

Considering the Critical nature of risks in this category, LinkedIn continues to invest in the development and enhancement of strategies to further mitigate and manage risks related to Public Security, aiming for an even safer user experience. The key mitigations currently implemented specific to Public Security are described in the subsequent paragraphs of this section.

LinkedIn Policy does not tolerate content promoting dangerous organizations or individuals. LinkedIn does not allow organizations or individuals that espouse violence on LinkedIn, including terrorist organizations, violent extremist groups, and individuals affiliated with such groups or activity. LinkedIn also restricts Profiles and Pages associated with these organizations and individuals regardless of whether they have posted violative content.

LinkedIn does not allow any terrorist organizations or violent extremist groups on the platform nor individuals who affiliate with such organizations or groups to have a LinkedIn Profile. Content that depicts terrorist activity, which is intended to recruit for terrorist organizations, or threatens, promotes, or supports terrorism in any manner is not tolerated on LinkedIn. LinkedIn's Professional Community Policies do not allow organizations or individuals that espouse violence on LinkedIn, including terrorist organizations, violent extremist groups, and individuals affiliated with such groups or activity. LinkedIn restricts Profiles and Pages associated with dangerous organizations and individuals regardless of whether they have posted violative content.

LinkedIn's [Professional Community Policies](#) detail that although it is allowed to have civil conversation about all topics, including those touching on Public Security, LinkedIn does not allow that conversation to be expressed in a hateful or bullying manner. For certain egregious violations to LinkedIn's Professional Community Policies (e.g., terrorism, extremely violent content), LinkedIn may permanently restrict a member's account after a single violation.

Content review is conducted by a global team of language and policy enforcement experts that adhere to strict limits with respect to the time in which the review must be completed (SLAs). The time limits are dependent on

the severity of member harm, and certain types of content, e.g., potential terrorist content, is human reviewed and removed as soon as possible. In addition, the Enforcement Operations (EO) team is subject to exacting training and quality assurance standards, and LinkedIn has separate teams that effectuate these standards. LinkedIn's team of internal content policy professionals also provide specific civic and election-related guidance to the EO team as appropriate given the circumstances.

LinkedIn uses hash matching to detect and remove previously identified terrorist or violent extremist content. LinkedIn also works closely with Global Internet Forum to Counter Terrorism (GIFCT) and other civil society groups to receive inputs to its content policies and additional information on how its platform can be abused to harm users. As a member of GIFCT, LinkedIn has access to GIFCT's Incident Response processes, including access to hash databases or Content Incident Protocols.

LinkedIn uses detection systems that identify the presence of a single bad actor creating or controlling multiple accounts, and it identifies anomalous activity that may be indicative of a fake account. LinkedIn's automatic detection filters for policy violating content prior to being posted on the platform. LinkedIn also relies on member reports and a team of investigators to uncover accounts and groups that may have evaded all other levels of defense. In accordance with applicable legal obligations, LinkedIn reports certain CSAM-related information to the National Center for Missing & Exploited Children (NCMEC) and provides member notifications in the event of significant privacy breaches. LinkedIn also publishes information regarding its responses to government removal requests and voluntarily discloses certain information to relevant law enforcement authorities, as permitted by law, in the event LinkedIn becomes aware of certain imminent threats.

#### **"Illegal" Content and Activities**

Risks related to "Illegal" Content and Activities include risk that "illegal" content and activities such as intellectual property (IP) infringement or defamation;<sup>4</sup> the promotion, concealment, acquisition, or sale of illegal, dangerous, or regulated goods (including counterfeit goods, animals, drugs, psychoactive substances, criminal property, firearms, and other weapons) and services relevant to the same; the facilitation of illegal immigration; or activities causing the suffering of animals are shared or perpetrated on the platform. Absent sufficient mitigations, risks related to "Illegal" Content and Activities may manifest on the platform in ways such as advertisers posting advertisements for illegal substances, members facilitating sale of illegal products via messages, members posting jobs that facilitate illegal immigration, or members posting content that constitutes IP infringement.

The **Inherent Probability** of systemic risks concerning "Illegal" Content and Activities stemming from the function, use, or misuse of LinkedIn's services is assessed as "Not Likely." Due to the professional nature of LinkedIn, it is "Not Likely" that members would publicly engage in illegal activities on the LinkedIn platform. The LinkedIn platform also does not allow for rapid and wide dissemination of content that has not been selected for quality, which limits systemic spread of illegal content. Considering the relative frequency of content or activities related to "Illegal" Content and Activities appearing in LinkedIn's content moderation systems as well as public discourse related to LinkedIn, the probability is assessed as "Not Likely" that users would share or access "Illegal" Content or perpetrate Illegal Activities on LinkedIn's platform absent sufficient mitigations.

The **Inherent Severity** of impact for content and activities that negatively impact "Illegal" Content and Activities on LinkedIn is rated as "Critical" primarily due to the gravity of risks within this category considering the potential for harm to security, environment, and wellbeing systems up to the regional and even the global level for environment systems. Some risks within this category have potentially irremediable consequences.

---

<sup>4</sup> Although defamation is not criminalized in all jurisdictions, LinkedIn has included defamation in this risk category for purposes of its systemic risk assessment.

With the rating for Inherent Probability as “Not Likely” and Inherent Severity as “Critical,” the **Inherent Risk** associated with negative impacts to “Illegal” Content and Activities across LinkedIn products and features is therefore assessed as “Moderate.”

In addition to the DTSP Best Practices LinkedIn has effectively implemented across all Risk Areas, which are further described under “[Mitigations](#),” LinkedIn has implemented robust mitigations specific to “Illegal” Content and Activities following best practices with defined, documented, and managed processes. As such, the **maturity of mitigations** LinkedIn has applied to the “Illegal” Content and Activities risk area is assessed as “Managed” which brings the **Residual Risk rating** down to “Low.”

Considering the Critical nature of this risk area, LinkedIn continued to invest in developing and enhancing strategies to further mitigate and manage risks related to “Illegal” Content and Activities, aiming for an even safer user experience. The key mitigations currently implemented specific to “Illegal” Content and Activities are described in the subsequent paragraphs of this section.

LinkedIn’s Ads Agreement prohibits advertising customers from “engag[ing] in illegal or fraudulent conduct”. Similarly, its Advertising Policies prohibit, among other things, illegal products, services and activities; harmful products and services; counterfeit goods; and intellectual property infringement.

LinkedIn’s User Agreement and Professional Community Policies specifically prohibit the violation of IP rights, including copyright, patents, trademarks, trade secrets or other proprietary rights. For example, users may not copy or distribute (except through available sharing functionality) the posts or other content of others without their permission.

LinkedIn has internally developed classifiers with high levels of precision and recall that are trained on a set of confirmed policy-violating content identified by human reviewers used to proactively enqueue potentially policy-violating content for human review. In addition to proactive review, it is straightforward for members to flag potentially policy-violating content for human review. Content review is conducted by a global team of language and policy enforcement experts who adhere to strict limits with respect to the time in which the review must be completed. In addition, this team is subject to exacting training and quality assurance standards.

LinkedIn has a responsive approach to requests for removal of content that a user believes is defamatory or infringes IP rights. Using dedicated online forms, users can submit detailed complaints. LinkedIn processes and investigates each claim, and takes, documents, and communicates the appropriate action to the affected user(s).

Potentially illegal content shared within the LinkedIn Feed may not be eligible for broad distribution on the platform. Additionally, if other members are expressing disinterest in, dismissing or reporting a certain piece of content, that content may not receive broad distribution on the platform, as this indicates members find the content to be potentially illegal.

For certain egregious violations of its Professional Community Policies, LinkedIn may permanently restrict a member’s account after a single violation.

### *Risk Areas of Minimal Residual Risk*

Considering the probability, severity, and maturity of mitigations, LinkedIn assessed that two of the assessed risk areas fell into a Residual Risk category of “Minimal:” Private and Family Life and the Rights and Protection of Minors.

#### **Private and Family Life**

Risks to Private and Family Life include malicious sharing or exploitation of sensitive or private data, doxing, or sharing of non-consensual private images. Absent sufficient mitigations, risks related to Private and Family life

may manifest on the platform in ways such as members posting or threatening to post nonconsensual private images or information in Messages or members posting private information about another member or individual in Feed.

The **Inherent Probability** of systemic risks concerning Private and Family Life stemming from the function, use, or misuse of LinkedIn's services is assessed as "Remote." While LinkedIn limits the demographic or sensitive data collected by the platform and members tend to self-limit personal and family information shared on the platform, some members may share personal information on LinkedIn that may be exploited by bad actors and bad actors may share personal information about individuals from other sources to the platform. However, there is a lower likelihood of members sharing non-consensual intimate imagery or content on LinkedIn due to its professional nature. The LinkedIn platform also does not allow for rapid and wide dissemination of content that has not been selected for quality, which limits systemic spread of harmful content. Considering the relative frequency of content or activities related to Private and Family life appearing in LinkedIn's content moderation systems or in public discourse related to LinkedIn, the probability is assessed as "Remote" that users would share or access information or perpetrate activities negatively impacting Private and Family life on LinkedIn's platform absent sufficient mitigations.

The **Inherent Severity** of impact for content and activities that negatively impact Private and Family Life on LinkedIn is rated as "High" primarily due to the gravity of risks within this category, considering the potential for harm to economic, security, societal, and wellbeing systems at the individual level with some broader reverberations. Some risks in this category, such as posts that contain sensitive data or disclosure of nonconsensual private or intimate images, have potentially irremediable impacts. Some risks in this category, such as users' inadvertent sharing of sensitive personal information (e.g. GDPR Article 9 information that includes racial or ethnic origin, political opinions, religious or philosophical beliefs etc.) may be less severe.

With the rating for Inherent Probability as "Remote" and Inherent Severity as "High," the **Inherent Risk** associated with negative impacts to Private and Family Life across LinkedIn features is "Low."

In addition to the DTSP Best Practices LinkedIn has effectively implemented across all Risk Areas, which are further described under "[Mitigations](#)," LinkedIn has implemented mitigations specific to Private and Family Life following industry best practices with defined, documented, and managed processes. As such, the **maturity of mitigation** efforts LinkedIn has applied to the Private and Family Life risk area is assessed as "Managed," which brings the **Residual Risk** rating down to "Minimal."

LinkedIn continues to invest in developing and enhancing strategies to further mitigate and manage risks related to Private and Family Life, aiming for an even safer user experience. The key mitigations currently implemented specific to Private and Family Life are described in the subsequent paragraphs of this section.

LinkedIn's Professional Community Policies prohibit the posting of harassing content, including targeted personal attacks, intimidation, shaming, or disparagement, as well as revealing others' personal or sensitive information (aka "doxing") or inciting other to do the same. LinkedIn policies also prohibit sharing materials depicting nudity or sexual activity, creating fake profiles, and falsifying information about yourself or others.

If LinkedIn becomes aware that potentially policy-violating content has been posted, LinkedIn reviews that content for policy compliance. Potentially policy-violating content is identified for human review by internally developed classifiers with high levels of precision and recall that are trained on a set of confirmed policy-violating content. In addition to this proactive review, members can flag potentially policy-violating content for human review by LinkedIn's Trust and Safety team. If other members are expressing disinterest in, dismissing, or reporting a certain piece of content, that content may not receive broad distribution on the platform, as this indicates members potentially find the content to be harmful, unhelpful, or unwanted.

In an effort to prevent fraudulent and malicious actors from accessing LinkedIn or from posting malicious content or scraping Profile data, LinkedIn has built automated detection systems to prevent bulk account creation attempts, identify the presence of a single bad actor creating or controlling multiple accounts and identify anomalous activity that may be indicative of a fake account.

To ensure that safety, security, and privacy remain at the core of the design and modification of LinkedIn's products, LinkedIn maintains a comprehensive pre-launch product review process. Among other things, this review considers how the product would collect and use data, to help LinkedIn measure privacy risk and ensure that appropriate data protection controls and safety measures are included.

Finally, LinkedIn helps protect members by empowering them to control the privacy of their own data, determining whether they can appear in search results on external search engines, how much of their Profile appears to LinkedIn members who are not first connections, and who can see their contact information. For European members who have affirmatively enabled certain enhanced safety features (through their settings), LinkedIn also displays safety reminders to such members if attempts to move private conversations away from the platform are detected in their incoming messages. The member can click through to view or report the message. This flow is designed to empower and educate members to make informed safety choices before they move certain conversations off platform.

### **Rights and Protection of Minors**

Risks related to Rights and Protection of Minors include child sexual exploitation, grooming, or sharing Child Sexual Abuse Material (CSAM); collection of minor's data without parental consent; exposure of children to harmful content or activities; cyberbullying and harassment of children; or promotion of self-harm to minors. Absent sufficient mitigations, risks related to Rights and Protection of Minors may manifest on the platform in ways such as members attempting to contact minors through InMail for grooming purposes, members soliciting or sharing child exploitation material through Messages, advertisers posting ads that promote dangerous or harmful products or services to minors, or members posting harmful or harassing content directed at minors within Feed.

The **Inherent Probability** of systemic risks concerning Rights and Protection of Minors stemming from the function, use, or misuse of LinkedIn's services is assessed as "Remote." LinkedIn, as a specific-purpose professional network is not designed for or used heavily by users under the age of 18. There is also a lower likelihood of members sharing material related to child exploitation on LinkedIn due to its professional nature. LinkedIn has historically and consistently detected only limited instances of child sexual abuse material on the platform. Considering the relative frequency of content or activities related to the Rights and Protection of Minors appearing in LinkedIn's content moderation systems as well as public discourse related to LinkedIn, the probability is assessed as "Remote" that users could share or access information or perpetrate activities negatively impacting the Right and Protection of Minors on LinkedIn's platform absent sufficient mitigations.

The **Inherent Severity** for content and activities that negatively impact the Rights and Protection of Minors on LinkedIn is rated as "Critical" due to the gravity of risk in this category, the vulnerable nature of the population impacted, the irremediability of impact of certain risks, namely CSAM, and the potential for societal impact broader than the individual scale for risks impacting youth. Not all risks within this category are assessed at the highest level of severity, namely collection, processing, and use of children's data without parental consent.

With the rating for Inherent Probability as "Remote" and Inherent Severity as "Critical," the **Inherent Risk** associated with negative impacts to the Rights and Protection of Minors across LinkedIn products and features is "Low."

In addition to the DTSP Best Practices LinkedIn has effectively implemented across all Risk Areas, which are further described under [“Mitigations,”](#) LinkedIn has implemented robust mitigations specific to the Rights and Protection of Minors following industry best practices with defined, documented, and managed processes. As such, the **maturity of mitigation** efforts LinkedIn has applied to the Rights and Protection of Minors risk area is assessed as “Managed,” which brings the **Residual Risk rating** for down to “Minimal.”

Considering the critical nature of this risk area, LinkedIn continues to invest in developing and enhancing strategies to further mitigate and manage risks related to the Rights and Protection of Minors, aiming for an even safer user experience. The key mitigations currently implemented specific to the Rights and Protection of Minors are described in the subsequent paragraphs of this section.

Only a subset of minors (i.e., 16 and 17 year olds) are even eligible to use LinkedIn. Anyone under the age of 16 is specifically prohibited by LinkedIn’s [User Agreement](#) from having an account. During the account creation process, new members who are listed as attending a secondary education institution are required to attest that they are over 16 years old. If they say “no,” then they are not allowed to proceed with creating an account. If LinkedIn otherwise becomes aware that a member is under the age of 16, LinkedIn will close the member’s account.

Although most minors are restricted from becoming LinkedIn members and those on the platform are not many, LinkedIn goes to great lengths to ensure their safety and security. LinkedIn has published robust policies prohibiting underage users on the platform, the sharing of CSAM, and the use of children in Editorial and Learning content.

LinkedIn has developed advanced detection capabilities related to harmful content that underage users might be exposed to on LinkedIn, including illegal or dangerous goods or services, dangerous organizations or individuals, and hateful, harmful, shocking, violent, or harassing content. Specifically, for CSAM, LinkedIn scans all public-facing and privately shared videos and images for CSAM (including content that is nearly identical to confirmed CSAM content), conducting hash matching against known CSAM imagery. Detected CSAM is removed or prevented from being shared on the platform, reported to the National Center for Missing and Exploited Children (NCMEC), preserved to enable law enforcement investigations, and the associated account is placed in restriction. Once CSAM has been identified, LinkedIn also applies algorithmic models to identify and remove near-duplicates from the platform. Although the numbers of confirmed CSAM cases on LinkedIn are quite small, as explained in LinkedIn’s [Transparency Report](#), LinkedIn remains diligent and focused on this anti-abuse work.

LinkedIn does not enable ad targeting to age groups below 18 in the EU and takes additional steps, out of an abundance of caution, to ensure that profiling ads are not received by EU members under 18. Specifically, LinkedIn excludes members who are listed as attending a secondary education institution on their Profile from all targeted advertising. They will still see ads, but they will not be personalized based on any profiling or activity information. LinkedIn will be transparent about this practice, including the fact that it may be overly broad and impact members that are not minors but are listed as attending a secondary education institution. Additionally, all members have additional controls to restrict the privacy of their information, including whether their Profile is shown in Search Engines.

LinkedIn also empowers members (including those under the age of 18) to control how their information is viewed by others and how others interact with them. For example, members can:

- control how their profile is shown outside of LinkedIn, such as on search engines;
- control who can send them an invitation to connect on LinkedIn (e.g., you can choose to receive connection invitations only from people who already know your email address);
- disallow non-connections from sending them messages (to avoid unwanted contact);

- block particular members by using the “block” feature (see more here);
- restrict the visibility of their email address, list of connections, and last name;
- disallow others from finding their LinkedIn profile even if the searching person knows the member’s email address or phone number; and
- turn on two-factor authentication.

## Scoring Rationale

### *Year 2 Changes in Risk Scores*

The results of the LinkedIn Systemic Risk Assessment this year are largely unchanged from the Year One assessment, with some notable shifts in Residual Risk and Inherent Probability.

- The Inherent Probability related to Consumer Protection and Fraud increased from “Highly Likely” last year to “Expected” this year based on the relative (i.e., comparable) percentage of flagged fraudulent accounts on LinkedIn as compared to other platforms (based on publicly available data). The increase in Inherent Probability also increased the Inherent Risk Rating from “High” last year to “Critical” this year and the Residual Risk rating from “Low” last year to “Moderate” this year. Therefore, LinkedIn is taking a conservative approach in scoring the Inherent Probability of Consumer Protection and Fraud as “Expected” since it is the most relevant systemic risk to LinkedIn based on the nature of the platform and historical data.
- Residual Risk related to Public Health decreased from “Moderate” last year to “Low” this year based on a decrease in the inherent probability of Public Health content on the platform due to the additional distance from the COVID health crisis.
- Residual Risk related to Private and Family Life decreased from “Low” last year to “Minimal” this year based on a decrease in the relative frequency of content or activities that may negatively impact Private and Family Life appearing in LinkedIn’s content moderation systems as well as public discourse related to LinkedIn.
- Inherent Probability related to “Illegal” Content and Activities decreased from “Likely” last year to “Not Likely” this year due to the updated Risk Area definitions, including shifting terrorist content out of this Risk Area into Public Security, shifting stalking and threats of violence into Mental and Physical Wellbeing, and shifting hate speech into Discrimination and Hate.
- Mental and Physical Wellbeing is a new Risk Area this year and received the same Residual Risk score this year as Gender-Based Violence (which was folded into this Risk Area) received last year.
- Public Security is a new Risk Area this year that received the same Residual Risk score this year as Civic Discourse, Electoral Processes, and Public Security received last year.

### *Probability Across Risk Areas*

While LinkedIn’s platform and services can host a variety of systemic risks, the Inherent Probability or likelihood of occurrence for most risk areas is assessed at relatively low levels. The professional nature of the platform discourages typical members from posting potentially harmful content that would be visible to their employers and potential or current business partners. This determination is supported by the fact that the volume of violative content on LinkedIn is consistently lower than industry norms.

Considering (1) content flagged and removed from the platform, by automated classifiers and human reporting, (2) the volume of public discourse related to LinkedIn and each risk area, and (3) trends in public concern and emerging threats across the industry, LinkedIn assessed Consumer Protection and Fraud as having the highest relative probability of occurrence on the platform, followed by Human Dignity, Mental and Physical Wellbeing, and Discrimination and Hate. Private and Family Life and the Rights and Protection of Minors were assessed at the lowest levels of probability.



### *Impact of Influencing Factors*

In the assessment of each risk area, LinkedIn considered the potential impact of all Article 34 factors on probability, severity, and mitigation effectiveness, including the design of recommender systems and algorithmic systems; content moderation systems; applicable terms and conditions and their enforcement; systems for selecting and presenting advertisements; LinkedIn data related practices; intentional manipulation of the service; amplification or rapid and wide dissemination; and specific regional or linguistic aspects.

**Recommender Systems and Rapid Dissemination:** LinkedIn strives to broadly distribute only high-quality content—content that meets specific quality and professionalism criteria – in Feed. In practice, this means content that shares knowledge or advice on professional topics that LinkedIn thinks its members will likely be interested in will get broader distribution, whereas divisive, unprofessional, or harmful content should not be amplified or broadly disseminated. Furthermore, the GenAI features that LinkedIn provides can only be used in limited, context-specific scenarios, thereby limiting the widespread use of AI systems capable of generating text, images, videos, or other synthetic content and the dissemination of such content on its platform. This aligns with LinkedIn’s view that the goal of LinkedIn is not virality or reach at all costs; instead, it’s about reaching the right professionals with the right information they care about that can help in their careers.

**Algorithmic Systems:** LinkedIn has a [robust process](#) in place to examine the fairness and potential bias of algorithmic systems deployed in any new feature or product. Algorithmic review is included as part of LinkedIn’s Trustworthy Design Review (TDR) process, the LinkedIn Fairness Toolkit (LiFT), and regular Fairness Checks of Feed and content distribution. LinkedIn is committed to upholding trust, promoting fairness and inclusion, providing transparency, and embracing accountability with the use of AI on its platform.

**Content Moderation:** LinkedIn’s approach to content moderation is risk-based and carefully weighs safety against freedom of expression, erring toward freedom of expression whenever possible. LinkedIn employs advanced detection techniques and sophisticated algorithms to detect [policy-violating content](#) and activity on the platform before it causes harm. LinkedIn uses a layered, multidimensional approach to moderate content within LinkedIn’s Trust ecosystem, including automatic and proactive prevention as well as human-led detection. In some cases, LinkedIn may permanently suspend an account after multiple policy violations or a single egregious content policy violation (e.g., in the case of child exploitation material).

Final decisions on content removal or account suspension made by LinkedIn’s human reviewers feed back into the system to further train the platform’s models. LinkedIn measures its preventive defense services regularly to improve accuracy in the filtering process.

**Advertisements:** It is important to note that a majority (roughly 80%) of LinkedIn’s advertisements are business-to-business rather than business-to-consumer, which has a significant impact on LinkedIn’s risk profile related to advertising systems. LinkedIn maintains and enforces [policies](#) related to advertising on the platform. These policies include, among other things, prohibitions on political advertisements; policies that limit health-care related advertisements to come from certified retailers and that limit distribution to jurisdictions where such advertisements are legal; and policies that prohibit discrimination based on protected characteristics. LinkedIn has also updated its advertising tools to remove the ability to target members in Europe using LinkedIn Groups, to prevent any misconception that ads could be targeted at members using special categories of data. LinkedIn also allows members in Europe to choose whether to connect their data across LinkedIn features and solutions.

**Intentional Manipulation of the Service:** As a real-identity platform, LinkedIn addresses intentional manipulation at the source by investing significantly to prevent the creation of [fake accounts](#). LinkedIn’s teams have created and deployed sophisticated models to block bulk account creation, identify clusters of bad accounts, detect anomalous account behavior, identify AI-generated Profile photos, and remove phishing links and malware. In the last Transparency Reporting period, LinkedIn’s automated defenses blocked 90.5% of the

fake accounts that were caught during the July - December 2023 period, with the remaining 9.5% stopped by manual investigations and restrictions. 99.6% of the fake accounts were stopped proactively, before a member report.

**Linguistic and Regional Considerations:** LinkedIn currently supports the following languages on its platform: English, Arabic, Chinese, Czech, Danish, Dutch, French, German, Hindi, Indonesian, Italian, Japanese, Korean, Malay, Norwegian, Polish, Portuguese, Romanian, Russian, Spanish, Swedish, Tagalog, Thai, Turkish, and Ukrainian. LinkedIn has enabled automated and user-led detection of policy-violating content in all supported languages. Regarding European elections, LinkedIn engages with stakeholders as needed to address rising concerns relative to election cycles and ensure appropriate linguistic and cultural coverage. LinkedIn does not allow political advertisements.

## Mitigations

DTSP participating companies commit to five fundamental areas of best practices, which are the foundation for the operation of trusted and safe products and services: (1) product development, (2) governance, (3) enforcement, (4) improvement, and (5) transparency. Underpinning these are the 35 specific best practices of the DTSP Best Practices Framework that provide concrete examples of different activities and processes that companies may have in place to mitigate risks associated with harmful content and conduct. By committing to these best practices, LinkedIn has aligned with industry-accepted Trust and Safety standards, enabling it to better identify and manage potential risks. By mapping these mitigations to each of the DTSP Best Practice Buckets, LinkedIn creates greater transparency and provides reliable and effective solutions to challenges associated with harmful content and conduct.

## Best Practice Highlights

### *Product Development*

**Trustworthy Design Review (TDR):** The TDR process is one of LinkedIn's primary tools for risk identification and mitigation during the product development lifecycle. TDR helps LinkedIn's features meet LinkedIn's high standards for protecting members' privacy and safety. At LinkedIn, new products and initiatives undergo a pre-launch review and assessment by cross-disciplinary specialists, with a focus on legal compliance, data protection, security, and safety. The teams first exchange certain key information about the new product or initiative to determine if a privacy impact assessment, security review, and/or safety assessment should be conducted as part of the TDR process. This means that Trust and Safety consultation is built into the review process for new products and initiatives. During this evaluation, LinkedIn reviews and documents any potential risks and abuse vectors associated with new products or features including, but not limited to, considerations of member trust, necessary tooling, policy, support requirements, member education, transparency, public relations risks, training needs, and trust design recommendations, along with their respective mitigations. The process provides feedback to enhance the product or initiative and improve relevant Trust and Safety measures and helps to ensure compliance with data protection commitments, security standards, and applicable laws, as well as alignment with LinkedIn's Guiding Principles for Trust. As part of TDR, LinkedIn's legal team also reviews any proposed new GenAI feature on the platform, including to ensure members are aware that a feature uses GenAI and - if GenAI content is suggested to the member - the member should review and revise such content, as they're responsible for what they ultimately publish or send.

**Abuse Pattern Analysis:** LinkedIn has a dedicated abuse prevention team that prioritizes, triages, and remediates numerous abuse vectors (both acute and systemic) in order to keep LinkedIn members safe. LinkedIn's Trust Product team orchestrates regular prevalence exercises where it manually reviews samples of accounts and content on platform to identify new patterns of abuse on the site, which informs both design of

future defenses and priorities in its Trust and Safety roadmap. The Trust Product team has also implemented several proactive defenses to detect fraudulent behavior and policy-violating content. LinkedIn monitors the performance of these defenses regularly to ensure they are working optimally and address any gaps as they are identified.

**Trust and Safety Consultation:** In addition to the TDR process, LinkedIn's Trust and Safety, Trust Engineering, Trust Product, and Trust Data Science teams work in concert through the entire product development, strategic planning, and budgeting lifecycles to ensure alignment of LinkedIn's priorities and initiatives with Trust Principles.

**Accountability:** LinkedIn has built accountability into the TDR process by managing cross-functional reviews in a centralized tracking and project management system to help ensure accountability and thoroughness of the review process.

**Feature Evaluation:** As described above, specific features are assessed holistically as part of the TDR process before launch. For algorithmic features, LinkedIn has also developed and open-sourced the LinkedIn Fairness Toolkit (LiFT), which makes use of common fairness definitions to enable the measurement of fairness in large-scale machine learning workflows. It can be deployed in training and scoring workflows to measure biases in training data, evaluate different fairness notions for ML models, and detect statistically significant differences in model performance across different subgroups.

Specifically, when building out GenAI features, Trust, InfoSec and Legal teams provide detailed list of requirements (e.g. information on Responsible AI Transparency and Labeling requirements) for product teams to abide by during the TDR process. Together, these teams work to comply with the law, ensure member trust, comply with LinkedIn's Responsible AI Principles, and ensure LinkedIn's applications, features, and data are secure.

LinkedIn is mindful that GenAI models can reflect the stereotypes that exist in society and actively works to avoid harmful stereotypes in GenAI tool outputs by deploying multi-layered tactics throughout the product development life cycle. LinkedIn conducts Fairness Checks on material changes to the algorithms for LinkedIn Feed and LinkedIn Job recommendations, conducts A/B testing, and holds out groups of members to assess impacts of algorithmic feature launch. For instance, to mitigate the risk of abuse on job descriptions (for instance through injection of harmful text into the prompt) in LinkedIn Jobs, the team has deployed drop down menus that include standardized information about the job. These measures enable the generation of job descriptions that are responsive to the customer's inputs while also reducing the potential for abuse and bias. Similarly, with personalized writing suggestions for your Profile (while it's still based on information already on the member's Profile) LinkedIn constructed the prompts in a manner that reduces the risk of problematic output (resulting from "jailbreaking" the AI tool through nefarious prompts that aim to end run restrictions in AI that are designed to maximize the chance that outputs will be reasonable and appropriate) and the member feedback menu includes the option to flag "biased content."

**Risk Assessment:** LinkedIn's Digital Safety function, partners with the relevant product and engineering organizations to conduct continuous assessments of risk during the full product development lifecycle. This ensures a "safety by design" approach throughout the process. Furthermore, LinkedIn's Compliance Team performs LinkedIn Compliance deep dives and assists with the Company's Internal Audit workstreams. LinkedIn also has a robust Trust Product planning process (including the TDR process described above) to address emerging threats and risks.

**Pre-Launch Feedback:** Pre-launch feedback from cross-functional stakeholders is incorporated as part of the TDR review process. This process enables LinkedIn to evaluate new products and features for risk and abuse vectors, including but not limited to, member trust, tooling needs, policy, support requirements,

communication, transparency, PR risks, potential abuse vectors, training needs and trust design recommendations. For GenAI features, LinkedIn has measures in place to help ensure that any model used and feature launched is compliant with LinkedIn's Responsible AI Principles. Model alignment focuses on fundamental risks (e.g. sexual content) and not on product specific risk (e.g. political content). User feedback metrics are also used to serve as the guardrails for and to improve all GenAI features and models.

**Post-Launch Evaluation:** In addition to the TDR process, the product team(s) responsible for the relevant product or initiative lead ongoing engagement with the other trust stakeholders from the TDR process to learn from real risk patterns and implement any additional safety features post-launch. This process may include Fairness Checks and other Quality Assurance activities, further described under Effectiveness Testing. If a high-risk situation is identified post-launch, LinkedIn teams take mitigating measures before the product or feature is fully ramped.

**User Feedback:** LinkedIn has a "feedback" button on the site for users to provide feedback. Members can also directly provide feedback using several in-product features like "Reporting," "Blocking," "I don't know this person," etc. Users can also provide feedback by directly contacting LinkedIn's customer support teams. This feedback is incorporated via various automated and manual means into its proactive defenses to help improve accuracy and coverage. An active external and internal communications team cascades member feedback to various internal teams. LinkedIn monitors metrics on these reports regularly to identify and address problematic trends. These learnings are then used to inform both the design of future defenses and priorities in its Trust and Safety roadmap. In addition, LinkedIn's User Experience and Research (UXR) and Market Research teams partner with product and design stakeholders to gather feedback directly from LinkedIn members, customers, and internal users of Trust tools throughout the innovation process; foundational research feeds ideation based on member Trust experiences and evaluative usability research on proposed Trust features.

**User Controls:** LinkedIn provides platform and feature specific controls that allow members to tailor their experience on the platform via user controls and settings.

- **Cross-Feature Controls:** Members in the EEA and Switzerland can choose whether to connect their core LinkedIn experience (including feed, messaging, and search) with other LinkedIn services (jobs, marketing solutions, and learning course services) to tailor these services to their Profile or keep them generic. Members can also access and manage their data or delete it from LinkedIn at any time.
- **Feed:** LinkedIn provides controls that enable members to select whether to view their Feed by personalized or non-personalized priority; hide posts; block, mute, or unfollow members; and report content. A "Sort by" feature is available in the user interface, allowing recipients of the service in the European Union based on IP to change their preferred sorting option for their Feed, thus providing an alternative to the recommender system. The "Most recent" setting arranges updates by publication date, while the "Most relevant" setting ranks them by relevance leveraging the Feed Recommender System. LinkedIn's "I don't want to see this" option in the overall Feed surface allows members to prevent specific posts from appearing in their feed again. Members also have granular control over notifications through their settings.
- **Ads:** All LinkedIn members have the option to select within their settings whether they would like their Profile information to be used to personalize the ads visible to them on LinkedIn. Europe-based members are automatically opted out of third-party targeted advertising. LinkedIn does not enable ads targeting to age groups below 18 in the EU and takes additional steps, out of an abundance of caution, to ensure that profiling ads are not received by EU members under 18. LinkedIn also updated its advertising tools to remove the ability to target members in Europe using membership in LinkedIn Groups. This change, now effective for all new advertising campaigns, was made to prevent any

misconception that ads to European members could be indirectly targeted based on special categories of data or related profiling categories.

- **Profile:** All members have additional controls to restrict the privacy of their information, including whether their Profile is shown in search engines, how much of their Profile appears to LinkedIn members who are not first connections, who can send them an invitation to connect, and who can see their contact information. Members who have voluntary verifications displayed on their Profile (e.g., identity, workplace or educational institution verifications) can remove such verifications from their Profile at any time through the member's Profile Settings.
- **Games:** Configurable settings exist to turn on/off the sharing of Games-related content. Games scores are private unless a member chooses to share them.
- **Jobs:** Members can opt out of personalized job experiences and not receive job recommendations.
- **Messages:** Members can choose to only receive communications from people they know, thus allowing them to control who can contact them.

### *Product Governance*

**Policies & Standards:** LinkedIn's content policies are developed by its Content Policy team, which is part of LinkedIn's Legal team, in close collaboration with other parts of Legal as well as Trust and Safety, Product, and Engineering teams to operationalize those policies with granular procedures. LinkedIn's terms of service and policies are updated periodically, as needed. The [Professional Community Policies](#) applies to all members and content. Job posts on LinkedIn must also comply with LinkedIn's [Job Policies](#).

Anyone under the age of 16 is specifically prohibited by LinkedIn's terms of service—its User Agreement—from having an account. During the account creation process, new members are required to attest that they are over 16 years old. If they say “no,” then they are not allowed to proceed with signing up. If LinkedIn otherwise becomes aware that a member is under the age of 16, LinkedIn will close the member's account.

Furthermore, for online advertisers, LinkedIn sets forth detailed policies and standards in the LinkedIn Ads Agreement regarding launching ad campaigns and all online advertisers are required to abide by the requirements listed to launch successful ad campaigns. LinkedIn also has global ad policies in place that do not allow targeting sensitive groups on the platform. This decision helps to protect LinkedIn members and customers.

**User focused product management:** In its Feed, LinkedIn strives to broadly distribute only high-quality content that meets specific quality and professionalism criteria. In practice, this means content that shares knowledge or advice on professional topics that LinkedIn thinks its members will likely be interested in will get broader distribution, whereas divisive, unprofessional, or harmful content will not be amplified or broadly disseminated. This aligns with LinkedIn's view that the goal of LinkedIn is not virality or reach at all costs; instead, it's about reaching the right professionals with the right information they care about that can help in their careers. LinkedIn leverages user feedback to limit the visibility of content that is less likely to lead to an enhanced professional experience. LinkedIn provides the opportunity for users to select whether to view their Feed in a personalized or non-personalized manner; hide posts; block, mute, or unfollow members; and report content. LinkedIn also leverages user reports to find and remove posts and comments that go against LinkedIn's Professional Community Policies which is explained more in the following Community Guidelines/Rules best practice.

**Community Guidelines/Rules:** LinkedIn's public-facing policies are set forth in the User Agreement, Privacy Policy, Professional Community Policies, Jobs Terms & Conditions, and Ads Policies, each of which (1) can easily be accessed on the LinkedIn site and within the product, and (2) are publicly available in a number of languages. LinkedIn's User Agreement and Privacy Policy also are presented to members upon registration. Users can

leverage LinkedIn's searchable Help Center for additional details about its policies and how LinkedIn enforces them, including how to report abusive content. LinkedIn's policies include provisions that outline what is and what is not allowed on the platform. LinkedIn Professional Community Policies prohibit a range of abusive content, including harassing and abusive content, violent and graphic content, child exploitation material, illegal and inappropriate commercial activity, dangerous organizations and individuals, false or misleading content, scams and fraud, hateful and derogatory content, sexual harassment and unwanted advances, nudity and adult content, and spam content.

**User Input:** LinkedIn has an active external and internal communications team that interviews users across a variety of topics and cascades member feedback to various internal teams. Users can also directly provide feedback about Trust issues using several in-product features like "Reporting," "Blocking," "I don't know this person," or by reaching out to LinkedIn's customer support teams. This feedback is incorporated via various automated and manual means into LinkedIn's proactive defenses to help improve accuracy and coverage. LinkedIn monitors metrics on these reports on a regular basis to identify and address problematic trends. These learnings are then used to inform its Trust and Safety roadmap.

LinkedIn offers members the ability to shape their feed based on the accounts they follow. LinkedIn offers users topic agnostic empowerment tools to control their feed including: Disinterest, hide, and reporting content (LinkedIn alters feed ranking based on these signals) and the ability to unfollow, block, or remove connections (which will also alter the feed that is recommended to the member). To mitigate the impact of recommender systems, LinkedIn also allows members to sort their feeds chronologically in the EU. The provided [Help Center article](#) contains more information regarding how the feed is shaped.

**External Consultation:** LinkedIn consults with non-governmental organizations, peer companies, and other external stakeholders to enhance policies and community standards as well as product design and enforcement, including the [Global Internet Forum to Counter Terrorism \(GIFCT\)](#), the [Global Network Initiative \(GNI\)](#), and the [Digital Trust and Safety Partnership \(DTSP\)](#). It also engages with key external stakeholders through the EU Code of Practice on Disinformation and the EU Code of Conduct on Countering Illegal Hate Speech.

LinkedIn is engaged with numerous EU bodies to address the exchange of information such as Rapid Response Group as part of the Code of Practice, the European Commission, Coimisiun na Mean, Arcom, BSI, Europol, and other pan-EU bodies. These groups provide an efficient mechanism by which LinkedIn can exchange information about issues and content that requires immediate attention.

In addition to LinkedIn's internal team of global fact checkers, LinkedIn also utilizes external fact checkers, such as Reuters, as needed, to further expand local context/language coverage in EU countries. Reuters is a global news organization with reporters all over the planet who are widely respected as unbiased since they have adhered to a strict code of conduct for decades.

**Document Interpretation:** LinkedIn's Content Policy team, which collaborates closely with other parts of Legal as well as the Trust and Safety, Product, and Engineering teams, maintains detailed internal policies and procedures related to enforcement of the [Professional Community Policies](#). LinkedIn's internal policies are accompanied by sets of enforcement protocols that act as a step-by-step guide for consistent decision making. While the framework may vary across domains, LinkedIn leverages a common set of policy principles to define the spirit of each policy and drive ongoing research. LinkedIn documents its research efforts separately and cites them where appropriate; further, where novel policy challenges arise or precedents are set during escalations, LinkedIn leverages cross-functional issue briefs to document and refer to decision drivers.

**Community Self-regulation:** Several of LinkedIn's product features allow for community-led moderation. For example, Group Admins can approve or deny member posts prior to the posts being published. Admins or Managers can also remove posts from the Group. Group Admins also have the ability to remove or block a

member from the Group, which automatically removes all of their content. Members can report the post or comment, which is subject to LinkedIn's Professional Community Policies. With respect to Pages, Page admins can delete comments that are posted on their Page's posts. Admins can also add and remove other admins, depending on their admin role. Considering the professional nature of LinkedIn, self-regulation is a frequently used mechanism.

### *Product Enforcement*

**Roles & Teams:** LinkedIn's Trust apparatus includes a Legal team, covering Digital Safety, Product, Platform, & Partnerships, and Privacy; Trust and Safety Operations; Engineering, Product, and Data Science, and similar teams covering unique issues related to LinkedIn Marketing Solutions and Jobs. LinkedIn has a large content moderation team with hundreds of employees and correspondingly large managed service provider that help enforce LinkedIn's Professional Community Policies and other policies. LinkedIn's Anti-Abuse team creates tools to enforce prohibition of false account activity.

**Operational Infrastructure:** LinkedIn operates in 26 languages (including 12 EU languages) and has 36 offices around the world. LinkedIn engages cross-functional teams in policy enforcement and maintains staff around the globe to provide linguistically and culturally relevant content moderation.

**Tooling:** LinkedIn uses various technology to support its Trust operations, including tools to process flagged and removed content, facilitate investigations, combat fake accounts, detect and prevent spam, and communicate with members. For example, LinkedIn has developed internal tools to allow its content enforcement teams to review items that have been flagged for review by both human and automated means for potentially violating the Professional Community Policies. LinkedIn's flagship content moderation tool helps reviewers evaluate each item enqueued for human review (e.g., post, Job, ad, or message) and determine whether it violates a LinkedIn policy. Reviewers label the content appropriately and proceed to another item in the review queue. In addition, if classifiers locate and have sufficient confidence that a particular piece of content is policy violating, LinkedIn may, depending on the severity of the violation, limit the content's visibility or remove it entirely. LinkedIn leverages content moderation technology created in-house by LinkedIn's Trust Engineering team as well as third-party enterprise tools. LinkedIn also builds tools such as the LinkedIn Fairness Toolkit (LiFT), which enables the measurement of fairness in large-scale machine learning workflows.

**Training and Awareness:** LinkedIn has implemented robust training and quality assurance programs for content moderators, including regular audits on sample sets of content reviewed by moderation teams, regular group calibration sessions to address common error trends, and coaching for lower performers. LinkedIn also employs a dedicated team of trainers, who not only support the onboarding of new content moderators, but also provide ongoing educational opportunities for all moderators. LinkedIn has an intensive onboarding and training program for enforcement personnel, including topical trainings on key risk areas like elections. Content moderators have direct access to content policy managers through regular office hours and dedicated escalation pathways. For particularly complex decisions, content policy managers also have access to in-house lawyers who can consult regional legal experts as needed. LinkedIn also operates a Privacy Champions program, hosts Data Privacy week events, and requires annual privacy training for all employees. Finally, specifically for GenAI features, LinkedIn maintains an internal wiki for all product teams to use as a resource, which includes the legal/trust requirements and other information about using GenAI. Regarding external training, LinkedIn periodically runs in-product educational campaigns on LinkedIn to increase member awareness of key safety features and member control options. For example, LinkedIn provides job search safety reminders within its product to educate members on how to recognize and report suspicious jobs. This aims to increase job seeker awareness about common fraud schemes and empower LinkedIn members to (1) play a proactive role in avoiding scams and (2) contribute to LinkedIn's efforts to safeguard all members through the reporting of suspicious activities.



**Wellness and Resilience:** LinkedIn Trust and Safety has a dedicated position focused on wellness and has established wellness programs in place globally to provide on demand - linguistically and culturally aligned - support to moderators who may need assistance after reviewing abusive or extreme content. LinkedIn has numerous resources available to employees that support various dimensions of wellness—from top notch medical benefits to a robust Employee Assistance Program (EAP) service, as well as comprehensive fitness and mental health programs on site and virtually. LinkedIn also has vendors in each region to support the mental health of LinkedIn content reviewers. These vendors provide group education and 1:1 wellness coaching sessions with a trauma-informed lens that is geared towards the most at-risk employees. LinkedIn's Trust and Safety team also regularly conducts "pulse checks" and surveys to understand the wellness needs of the team and shift resources as necessary to meet those needs. Separately, content moderation tooling enhancement requests are also collected and implemented on a regular basis to improve the moderator experience, both in terms of efficacy and wellness. LinkedIn also takes into consideration the physical environment of content moderators to enhance design of working locations to promote mental wellbeing, tracks turnover and attrition as an additional indicator of wellness and creates opportunities for team members to offboard to other roles within LinkedIn as needed.

**Advanced Detection:** LinkedIn uses a three-layer, multidimensional approach to moderate content within its Trust ecosystem. The first layer of protection is automatic and proactive prevention. When a member attempts to create a piece of content on LinkedIn, various calls (or signals) are sent to LinkedIn's machine learning services. These services aim to automatically filter out certain policy violating content within 300 milliseconds of creation, meaning the content is visible only to the author and is not shown to anyone else on the platform. As part of this process, AI tools play a key role in helping LinkedIn proactively filter out potentially harmful content. LinkedIn uses content (like certain key words or images) that has previously been identified as violating its Professional Community Policies to help inform AI models and better identify and restrict similar content from being posted in the future.

The second layer of protection is a combination of automatic and human-led detection. LinkedIn's second layer of moderation detects content that's likely to be violative but for which the algorithm is not sufficiently confident to warrant automatic removal. This content is flagged by LinkedIn's AI systems for further human review. The moderation management system automatically creates review items completely and accurately in the appropriate queues upon receipt of moderation requests from upstream systems. If the human review team determines that the content violates LinkedIn's policies, it is removed from the platform. LinkedIn's human review team is instrumental in this process and in helping train the platform's models. The third layer is addressed under User Reporting. The third layer of detection is covered in the following section "User Reporting."

LinkedIn's GenAI features are focused on the member's Profile and not general-purpose content creation. GenAI recommendations follow professional guidelines, enforced through various classifiers. For example, if a member prompts the AI to write a headline for their Profile including that they are an immigrant, the AI recommendation will omit that information and only include profession-relevant descriptors of the member (based on the provided prompt and the member's Profile information).

LinkedIn applies to GenAI inputs and outputs a binary moderation process (blocking or not blocking) to minimize harmful and inappropriate content within GenAI features, including inputs and outputs containing hate, violence or sexual content. Responses are moderated in chunks of 100 tokens at a time, so - if any portion of a response is found to be harmful or inappropriate, even if some portions of the response have already been provided - the entire response is retracted. LinkedIn also identifies inputs relating to self-harm, and redirects members who provide these inputs to self-harm resources. LinkedIn also has in place jailbreak moderation, which is designed to detect and prevent attempts to bypass these input and output protections.

In relation to messaging, members can enable automated detection of harmful content within messages in their settings to have messages scanned for potentially harmful or harassing content.

Advertisements are automatically evaluated against LinkedIn policies at the time of their creation.

**User Reporting:** The third layer of LinkedIn’s content moderation approach is human-led detection. If members identify content they believe violates the Professional Community Policies, LinkedIn encourages them to report it using the in-product reporting mechanism represented by the three dots in the upper right-hand corner of the content itself on LinkedIn or on the “More” button on the Profile headline. Reported content is then generally enqueued for human review by LinkedIn’s Enforcement Operations team or by LinkedIn’s automated system and is removed or restricted if found to be in violation of LinkedIn’s policies. The automated system is based in part on past decisions human reviewers have made regarding whether content violates LinkedIn’s policies. When users receive notification that their report has been resolved, the notice indicates whether the report was resolved by human review or LinkedIn’s automated system. Logged-out users in the EU are also able to report content visible to them using the guest reporting functionality. Reporting is available for all user-generated content on LinkedIn.

**Enforcement Prioritization:** Reported and classifier-detected content are enqueued based on the reporting reason chosen by the reporter or the classifier at issue, which determines the timeframe within which the content should be reviewed. Other factors considered when prioritizing reviews are language, region of the originating content, and type of media.

The label applied to content by classifiers or human reviewers determines how the content is actioned on platform. Depending on the severity of violation, LinkedIn may limit the visibility of certain content or remove it entirely. Repeated or egregious offenses may result in account restriction. For certain egregious violations of LinkedIn’s Professional Community Policies (e.g., child sexual abuse material, terrorism, extremely violent content, egregious sexual harassment), LinkedIn may permanently restrict an account after a single violation. Users can leverage the searchable Help Center for additional details about LinkedIn policies and how they are enforced, including how to report abusive content. LinkedIn monitors edge cases for sensitive policy escalations to document decisions and recommend policy adjustments as needed.

**Appeals:** Users have access to report content on each LinkedIn surface. Both reporter and creator generally are notified when content is removed or an account is restricted, and both are given an opportunity to appeal the decision. Notices are typically sent by email and contain a link to a notice page containing certain additional information (e.g., about the content at issue, the policy violated, the action LinkedIn has taken, redress info and, in most instances, a link to allow the individual to appeal LinkedIn’s decision). LinkedIn reviews those appeals and notifies the member of its appeal decision.

**External Reporting:** In accordance with applicable legal obligations, LinkedIn reports certain CSAM-related information to NCMEC and provides member notifications in the event of significant privacy breaches. As noted in greater detail in the “Product Transparency” section below, LinkedIn also publishes information regarding its responses to government removal requests. It also voluntarily discloses certain information to relevant law enforcement authorities, as permitted by law, in the event LinkedIn becomes aware of certain imminent threats (including, for example, imminent threats of self-harm and election-related influence operations).

**Flagging Processes:** LinkedIn has developed internal tools to allow content enforcement teams to review items that have been flagged for review by both human and automated means for potentially violating LinkedIn’s Professional Community Policies. Workbench, LinkedIn’s flagship tool for content moderation, helps reviewers look at each item (e.g. a LinkedIn post, job, ad, or message), evaluate it, and decide if it violates a LinkedIn policy. If it does, they label it appropriately and proceed to another item in the review queue.

Job posters are notified through email when their job posting is flagged for review, taken down, or paused due to trust review concerns. When notified, posters can review the job posting status, understand the reason for the flag, and take corrective actions as necessary.

**Third Parties:** LinkedIn works with both globally recognized independent fact checkers and human rights groups to mitigate content- and conduct-related risks. For example, LinkedIn works with an external global news organization when its internal fact checkers are not able to confirm the veracity of user-generated content. LinkedIn also works with civil society groups to understand how its platform can be abused to harm members, particularly in conflict zones.

LinkedIn has provided a [Researcher Access Program](#) which allows independent qualified researchers access to data that is publicly accessible on LinkedIn. LinkedIn selects APIs available for approved research projects that provide access to advertising and public post data, and as appropriate, supports research requests for other datasets containing public-facing LinkedIn data.

**Industry Partners:** LinkedIn works closely with Microsoft's Democracy Forward team on civic integrity. LinkedIn collaborates with industry partners are part of the [Global Internet Forum to Counter Terrorism \(GIFCT\)](#), the [Global Network Initiative \(GNI\)](#), and the [Digital Trust and Safety Partnership \(DTSP\)](#).

### *Product Improvement*

**Effectiveness Testing:** LinkedIn has a robust QA (Quality Assurance) program designed to help improve reviewer performance and consistency. Regular QA audits are performed on a sample of all content items reviewed by LinkedIn's moderation teams. Results are shared weekly with more detailed summaries and reports provided monthly. Lower performers receive one-on-one coaching, common error trends are addressed during group calibration sessions, and policy refresher trainings are delivered as needed. And moderators have direct access to global content policy managers through regular office hours and dedicated escalation pathways. For particularly complex decisions, content policy managers also have access to in-house lawyers who can consult country law experts as needed.

LinkedIn measures its preventive defense services regularly to improve accuracy in the filtering process. This is done by sending some positive samples for human review to measure the precision of LinkedIn's automated defense system. This reduces the likelihood that LinkedIn's auto-filtering process removes content that complies with LinkedIn's policies. LinkedIn also conducts Fairness Checks on Feed distribution and ranking, health checks, and long-term hold outs to make sure the algorithms are working as intended. LinkedIn looks at prevalence metrics via human review and will change the algorithm based on rising and dropping trends on the platform.

LinkedIn conducts vendor-led and internal red-teaming (both manual and automated) efforts to test GenAI capabilities for various harms and identify mitigations for such harms. Each product team launching a GenAI feature is also responsible for undertaking quality assurance testing on response quality, accuracy, and hallucinations (where AI confidently provides a wrong or unsupported response), with goals to remediate and prevent further inaccuracies.

**Process Alignment:** A number of factors figure into LinkedIn planning and process alignment for continuous improvement. LinkedIn has both a top-down and a bottom-up approach. Trust product management sets out priorities to address risks as part of the planning process, and in addition, product managers across the organization propose solutions to emerging issues as they arise. Trust planning is a robust process. The success of LinkedIn's large and cross-functional Trust Product team is measured on their ability to lower UMI (unique members impacted) related to content- and conduct-related risks.

**Resource Allocation:** LinkedIn rolled out a more sophisticated metric to measure the impact of risk, which can be used to allocate resources to address these systemic risks. LinkedIn has an active external and internal communications team that cascades member feedback to various internal teams within its broader Trust organization. It monitors metrics on these reports on a daily/weekly basis to identify and address problematic trends. LinkedIn also works with civil society groups to understand how its platform can be abused to harm members, particularly in conflict zones. In advance of each high priority election, LinkedIn reviews and determines whether an increase in resourcing is required to meet the demands of the upcoming election response.

**External Collaboration:** LinkedIn engages with external organizations to enhance understanding of potential risks and emerging trends as well as best practices for risk mitigation. External engagements include the [Global Internet Forum to Counter Terrorism \(GIFCT\)](#), the [Global Network Initiative \(GNI\)](#), and the [Digital Trust and Safety Partnership \(DTSP\)](#). LinkedIn also engages with key external stakeholders through the EU Code of Practice on Disinformation and the EU Code of Conduct on Countering Illegal Hate Speech. As a member of GIFCT, LinkedIn has access to the GIFCT's Incident Response processes, including ingesting hashes related to an event activated as Content Incidents or Content Incident Protocols. This allows LinkedIn to quickly become aware of, assess, and address potential content circulating online resulting from a terrorist or violent extremist event. For further information, see the [GIFCT's annual Transparency Report](#), which includes information on the hash-sharing database. LinkedIn also works with GIFCT and fellow members of that organization to solicit input on their content policies.

**Remedy Mechanisms:** Members can submit general complaints to LinkedIn by creating a support ticket in LinkedIn's [Help Center](#). Members can submit Privacy complaints through LinkedIn's [Privacy Hub](#). Where LinkedIn removes a member's content or suspends their account, LinkedIn generally provides the member notice of the action LinkedIn has taken and provides them an entry point through which to appeal LinkedIn's decision. LinkedIn reviews those appeals and notifies the member of the appeal decision.

### *Product Transparency*

**Transparency Reporting:** A cross-functional team of stakeholders from across LinkedIn's broader Trust organization (including members of its Legal, Trust and Safety and Data Science and Communications teams) work closely to produce a periodic global Transparency Report. LinkedIn's global Transparency Report is issued twice per year and is published in LinkedIn's Transparency Center ([linkedin.com/transparency](https://www.linkedin.com/transparency)). Additional country-specific Transparency Reports are also published depending on local requirements. LinkedIn's Transparency Reports currently report on the following metrics: fake accounts (including the number removed and percentages stopped at different stages); spam and scams (including percentages stopped by automated defenses and number removed proactively/after member reports); content removed under its Professional Community Policies (including harassment or abusive, misinformation, hateful or derogatory, violent or graphic, adult, and child exploitation); and copyright removals (including number of requests, total infringements reported, reported infringements removed/rejected). LinkedIn's Transparency Reports also include certain data on the number and types of government requests received and actioned by LinkedIn (including requests for user data and for content removal).

LinkedIn also publishes a [global Ads Library](#) where users can access information about ads that have run on LinkedIn. LinkedIn has also expanded the 'Why am I seeing this ad' feature to include additional information about the type of ads members see and other options to manage the advertising experience.

**Notice to Users:** LinkedIn's publicly available [Professional Community Policies](#) apply to all content surfaces. The content creator is generally notified when content is removed or an account restricted and given the

opportunity to appeal the decision. Notices are typically sent by email and contain a link to a notice page. LinkedIn reviews those appeals and notifies the member of its appeal decision. At the time a content creator (including ad publishers) is notified their content was moderated or restrictions were applied to their Profile, LinkedIn publishes the statement of reasons, which do not include personal data, to the European Commission's (EC's) DSA database.

Material updates to the User Agreement and Privacy Policy are communicated directly to members – by means of emails and in-product notifications in advance of the update, together with a preview of the update – and LinkedIn provides persistent links that allow users to access the terms.

**Complaint Intakes:** Members have several options for addressing concerns on LinkedIn. For general complaints, members can create support tickets by following the instructions on LinkedIn's [Help Center](#). Privacy-related complaints can be submitted through LinkedIn's [Privacy Hub](#). Additionally, members can report potentially violative content directly on the platform, with [reporting options](#) available on each surface.

**Researcher & Academic Support:** LinkedIn has long explored various mechanisms to support research on the issues related to LinkedIn's platform, including emerging skills and skill gaps (e.g., enabling academic research through the economic graph challenge, partnering with the World Bank). LinkedIn and Microsoft are collaborating with OpenMined on the "Initiative on Algorithms" to explore how to enable further research collaboration by exploring research infrastructure that protects user privacy and proprietary information, while achieving reproducibility, affordability, and scale for independent researchers. In furtherance with its commitments under the Code of Practice on Disinformation, LinkedIn is working with Trustlab to enable reporting on misinformation. More recently, LinkedIn has been supporting new research mechanisms related specifically to its ads and Feed, including by making a tool available that supports research into ad content shown on LinkedIn. It is also launching new mechanisms that will enable qualifying researchers to access public data (such as Feed) and ads.

Furthermore, LinkedIn supports researchers through the [\[beta\] Research Access Program](#). Researchers interested in using LinkedIn's publicly available data may fill out an [application form](#). Researcher applicants who satisfy the criteria set forth in the DSA are notified and asked to enter into terms governing their access to the public data (and/or use of APIs, as applicable) for the purpose of the approved research, including terms related to the security and legal use of personal data.

**In-Product Indicators:** LinkedIn uses a variety of in-product indicators to enhance member awareness and empowerment, including "Why am I seeing this ad?" as well as "Kindness Reminders" (currently deployed in English) to encourage members to think twice before posting comments that may be harmful or violate their Professional Community Policies.

European members can opt in (through their message settings) to an optional advanced safety feature<sup>5</sup> that allows LinkedIn's automated machine learning models to detect sexual harassment and certain other abusive content in messages received by the member. Where the setting is enabled and LinkedIn's automated systems detect potentially harmful content, depending on whether the sender and recipient have exchanged prior communications, the message will be directed to the recipient's spam folder or be hidden from the recipient by a warning. The warning can be dismissed, giving the recipient the ability to view and report the message to LinkedIn for review if desired.

In LinkedIn's commitment to uphold Responsible AI Principles and legal frameworks, LinkedIn has implemented a set of guidelines to ensure transparency and clear labeling across all AI-driven tools, products, or features. This approach involves consistent iconography and language throughout the member experience, context-

---

<sup>5</sup> LinkedIn automatically scans all content for CSAM, malware, and phishing.

appropriate labeling to distinguish AI from non-AI-generated content, adherence to legal and trust guidelines to avoid deceptive practices, and the inclusion of feedback mechanisms for GenAI experiences. Content containing the Coalition for Content Provenance and Authenticity's industry-leading "Content Credentials" technology (C2PA) will be automatically labelled on LinkedIn. Users will see the "Cr" label. By clicking the label, users will be able to trace the origin of the AI-created media, including the source and history of the content, and whether it was created or edited by AI.

LinkedIn offers members the choice to engage with AI-driven features, accompanied by prominent AI disclosures regarding potential inaccuracies. LinkedIn encourages user review, editing, and approval of AI-suggested texts, and provides direct links for members to Help Center articles. This approach underscores LinkedIn's dedication to fostering a responsible, trustworthy digital environment, leading by example in the Responsible AI space and ensuring the platform remains a safe, transparent, and empowering space for professional growth and connection.

## Monitoring the Effectiveness of Mitigations

LinkedIn has robust processes in place to monitor the effectiveness of mitigations implemented to address systemic risks on the service.

LinkedIn's Trust team maintains and regularly monitors a range of dashboards to determine the health of the platform, including dashboards tracking that the following perform as intended:

- classifiers designed to auto-action or enqueue content for human review;
- LinkedIn's hide, mute, unfollow, block, and disinterest functionalities; and
- LinkedIn's content moderation decision quality assurance audit process.

Designated teams within LinkedIn's broader Trust organization monitor these dashboards regularly to measure their performance against certain target or acceptable thresholds. Additionally, a number of the dashboards are configured to send real time alerts to LinkedIn's Trust team whenever an issue with a functionality is detected. Where an alert is triggered or a dashboard's metrics deviate from acceptable levels, the Trust team investigates and remediates any issues or deviations. Trust leadership also receives regular updates on the monitoring and performance of these dashboards, providing leadership regular visibility into the ongoing effectiveness of these mitigations.

AI plays a key role in helping LinkedIn proactively filter out policy-violating content. LinkedIn uses such content to help inform AI models and better identify and restrict similar content from being posted in the future. LinkedIn's data science team monitors the content filtering process to measure how many content violations are successfully prevented and how much remains on the platform. Content that is proactively removed at creation is tracked through a data pipeline, and LinkedIn measures its preventive defense services regularly to improve accuracy in the filtering process. This is done by sending a set of positive samples for human review to measure the precision of LinkedIn's automated defense system.

The Trust and Safety team also regularly reviews content moderation metrics to monitor the efficacy of its mitigations and to identify where additional mitigations may be needed to account for new abuse vectors. LinkedIn regularly conducts manual reviews to estimate the prevalence of violative content on the site across abuse types, categorized by policy type, and uses this metric to determine the relative priority for resource allocation.

LinkedIn's Trust and Safety processes are periodically audited by the Company's Internal Audit team, which conducts thorough and objective evaluations of key risks, processes, and mitigates, helping LinkedIn identify areas for improvement and maintain high standards of compliance and controls. LinkedIn also follows Management Action Plans mandated by the Internal Audit team.

## Progress Made on Enhanced Mitigations

In last year's Systemic Risk Assessment report, LinkedIn outlined areas of focus for enhancement of safety mitigations during the 2023 – 2024 reporting period. LinkedIn has made enhancements in each identified area as described below:

1. Fulfilling its commitment to Advance Responsible AI Innovation, including the roll-out of its AI Governance Framework and publication of explainable models.
  - a. LinkedIn rolled-out its [AI Governance Framework](#) and publication of [how bias is tested](#) in its algorithms to enhance explainability and transparency in this process.
2. Continued investment in internal tooling, classifier refinement, multilingual models, and enhanced automation capabilities.
  - a. As part of the Year Two DSA Systemic Risk Assessment, LinkedIn continued to adjust the mapping and categorization of various internal data sources to enable additional data-driven assessment of probability and mitigation effectiveness. This has benefited LinkedIn's transparency reporting efforts specific to DSA. In combination with this year's data, LinkedIn has demonstrated progress year over year on efforts to mitigate systemic risks stemming from use or misuse of the LinkedIn platform.
  - b. LinkedIn expanded and refined its classifiers for detecting harmful content across the 12 risk areas, with specific focus on EU Elections and hate speech around the Israel-Hamas war.
  - c. LinkedIn also continues to monitor incidents, user reports, and social feedback to ensure that LinkedIn's policies and principles address emerging risks and trends.
3. Expansion of Jobs' automated scanning and risk ranking to include discrimination and ingested jobs.
  - a. LinkedIn employs machine learning models specifically designed to detect and remove policy-violating discriminatory content. During job post creation, LinkedIn's systems analyze the job post to determine whether it includes discriminatory phrases. If it does, LinkedIn highlights the language to job posters in-product and invites them to review both the language and LinkedIn's policies before continuing.
  - b. LinkedIn regularly updates and improves its automated systems in light of new fraud trends and hiring practices that vary market to market.
  - c. During the review period, LinkedIn expanded this scanning to include ingested jobs, not just those created on the LinkedIn platform.
4. LinkedIn will invest further to expand upon the sophisticated models built to detect fake accounts to achieve greater fidelity on detection of additional inauthentic activity, such as fake connection requests and messages.
  - a. Considering the rapidly evolving nature of threats related to Consumer Protection and Fraud and the need to continually assess and enhance detection techniques, LinkedIn continues to prioritize the development and enhancement of strategies to further manage and mitigate risks related to Consumer Protection and Fraud, aiming to maintain a high level of consumer safety



on the platform. The key mitigations currently implemented specific to Consumer Protection and Fraud are described in the [Consumer Protection and Fraud](#) section.

5. Refinement of Trustworthy Design Review (TDR) Process to map more closely to DSA requirements. LinkedIn plans to refine its robust TDR process to align even more closely with DSA product design requirements.
  - a. To further align to the DSA requirements, LinkedIn continues to refine its measurement of systemic risk materiality and the relationship of that process with the TDR process.
6. Expansion of researcher access to data.
  - a. During the assessment period, LinkedIn launched its [\[beta\] Research Access Program](#). Researchers interested in using LinkedIn's publicly available data may fill out an [application form](#). Researcher applicants who satisfy the criteria set forth in the DSA are notified and asked to enter terms governing their access to the public data (and/or use of APIs, as applicable).

## Planned Enhanced Mitigation Efforts

In accordance with DSA Article 35, LinkedIn has included in [Appendix B](#) a list of focus areas for the coming year to further enhance the reasonable, effective, and proportionate mitigation of systemic risks stemming from the use, misuse, or functioning of the LinkedIn platform. The mitigation focus areas include, for example:

- Continued enhancements to and refinement of the pre-launch product review process;
- Expanded and refined monitoring of systemic risk mitigation efforts through a more direct mapping of classifiers and internal risk taxonomy to systemic risk areas; and
- Continued investment in efforts to identify fraudulent accounts, malware, and other Tactics, Techniques, and Procedures that may be used for fraud and/or degrading consumer protection.

## Conclusion

The objective of the Systemic Risk Assessment is to assess areas of greater potential risk stemming holistically and objectively from the function, use, or misuse of the LinkedIn platform. This exercise enables LinkedIn, as a Very Large Online Platform, to better plan for and prioritize enhancements related to people, policies, processes, and technologies that will further reduce risk of negative impact on the fundamental rights of users. While overall systemic risk related to LinkedIn's products and services are assessed as low, LinkedIn has leveraged this assessment to identify opportunities to continue refining, enhancing, and expanding its efforts to keep its platform and services safe, trusted, and professional.

# Appendices

Appendix A: Methodology Expanded

Appendix B: List of Article 35 mitigations

## Appendix A: Methodology Expanded

### Risk Areas

LinkedIn defined 12 key Risk Areas to facilitate an assessment of systemic risks stemming from the use or misuse of LinkedIn platform or services as required by DSA Article 34.

Figure 17

Risk Area	Risk Definition	Risk Scenarios
<b>Civic Discourse and Electoral Processes</b>	Risk that content or activities negatively impacting civic discourse or electoral processes occur on the platform.	<ul style="list-style-type: none"> <li>Coordinated manipulation, inauthentic activity, and foreign interference</li> <li>Election misinformation</li> <li>Political polarization and the creation of echo chambers</li> </ul>
<b>Consumer Protection and Fraud</b>	Risk that activities that are fraudulent occur on the platform.	<ul style="list-style-type: none"> <li>Scamming, defrauding, spamming, fraudulent business or Flagship</li> <li>False representation, abuse of position, falsifying information, including the creation of fake accounts, use of bots, or other deceptive commercial practices</li> </ul>
<b>Discrimination and Hate</b>	Risk that discriminatory content or activities that target or exclude individuals, groups, or businesses based on protected grounds or promote bias or discriminatory practices occur on the platform.	<ul style="list-style-type: none"> <li>Bias and discriminatory practices</li> <li>Hate speech and discriminatory content including racially or religiously aggravated public order offences and harassment</li> </ul>
<b>Freedom of Expression and Information</b>	Risk that freedom of expression, information, the press, or pluralism of the media are negatively impacted by the design or functioning of a platform and its algorithmic systems, policies, or enforcement.	<ul style="list-style-type: none"> <li>Over moderation and methods for silencing expression</li> <li>Recommendations that suppress pluralism</li> </ul>
<b>Human Dignity</b>	Risk that content or activities degrading human dignity occur on the platform.	<ul style="list-style-type: none"> <li>Human trafficking</li> <li>Sexual exploitation, pornographic content, prostitution, and extreme pornography</li> <li>Violence, gore, vulgarity, and profanity</li> </ul>
<b>“Illegal” Content and Activities</b>	Risk that illegal content or activities are shared or perpetrated on the platform.	<ul style="list-style-type: none"> <li>Intellectual property infringement, defamation</li> <li>Promotion, concealment, acquisition, or sale of illegal, dangerous, or regulated goods</li> </ul>

		<p>(including counterfeit goods, animals, drugs, psychoactive substances, criminal property, firearms, and other weapons) and services relevant to the same</p> <ul style="list-style-type: none"> <li>• Facilitating unlawful immigration</li> <li>• Causing suffering of animals</li> </ul>
<b>Mental and Physical Well-being</b>	Risk that activities potentially causing harm to mental or physical well-being occur on the platform.	<ul style="list-style-type: none"> <li>• Gender based violence</li> <li>• Threats of violence or to kill, harassment, stalking, public order offenses, fear or provocation of violence, and controlling or coercive behavior</li> <li>• Encouraging or assisting suicide or attempted suicide or serious self-harm</li> <li>• Behavioral addictions impacting mental health</li> </ul>
<b>Private and Family Life</b>	Risk that content or activities negatively impacting private and family life occur on the platform.	<ul style="list-style-type: none"> <li>• Malicious sharing or exploitation of sensitive or private data and doxing</li> <li>• Disclosing or threatening to disclose nonconsensual private or intimate images</li> <li>• Privacy intrusions</li> </ul>
<b>Protection of Personal Data</b>	Risk that activities that negatively impact protection of personal data occur on the platform.	<ul style="list-style-type: none"> <li>• Collection or processing of information without consent, including data mining or targeted advertisements without consent</li> <li>• Data breaches or other unauthorized disclosure or insufficient protection of data</li> </ul>
<b>Public Health</b>	Risk that content or activities negatively impacting public health occur on the platform.	<ul style="list-style-type: none"> <li>• Promotion of legal but harmful substances or practices</li> <li>• Health misinformation</li> </ul>
<b>Public Security</b>	Risk that content or activities degrading public security occur on the platform.	<ul style="list-style-type: none"> <li>• Terrorist recruitment, funding, or training, terrorist imagery or content, or support for terrorist acts</li> <li>• Mis- or disinformation related to crisis events</li> </ul>
<b>Rights of and Protection of Minors</b>	Risk that content or activities negatively impacting the rights, privacy, safety and security of minors occur on the platform.	<ul style="list-style-type: none"> <li>• Child sexual exploitation and abuse, grooming, and sharing of Child Sexual Abuse Material</li> <li>• Collection of children's data without parental consent or targeting children with personalized advertisements</li> <li>• Exposure of children to harmful content or activities, including gambling, exploitation, crime, violence, terrorism, or violent extremism</li> <li>• Discrimination, cyberbullying, and harassment</li> <li>• Promotion of self-harm</li> </ul>

## Assessment Process

LinkedIn followed the DTSP Safe Assessment framework process to guide the conduct of the risk assessment.

### [DTSP Safe Assessment step descriptions](#)

Figure 18

Step	Description	Objective
<b>Discover</b> relevant information	Engage key product stakeholders and perform initial information discovery on the company’s practices across the 5 DTSP commitments and identify the practices to be evaluated for their use in mitigating content and conduct risks	Establish baseline understanding of the operational landscape and identify the specific DTSP practices used to mitigate content- and conduct-related risks
<b>Identify</b> and prioritize relevant risk considerations	Using the artifacts and information collected during the “Discover” stage - identify, document, and prioritize risks about the ways that content- and conduct-related risks are identified and mitigated.	Prioritize risks about the ways that content- and conduct-related risks are identified and mitigated to inform focus areas for the assessment.
<b>Assess</b> practices and risk mitigation	For the relevant risks about the ways that content- and conduct-related risks are identified and mitigated at the company and focus areas identified in the previous step, analyze the practices employed to control for, or protect against, Trust and Safety risks	Understand current state processes, practices, and tools in relation to a common maturity scale
<b>Test</b> control strength and effectiveness [Level 2 and Level 3 only]	Perform a control strength evaluation, including control design and effectiveness testing.	Understand, at a granular level, the operational effectiveness of risk mitigation processes, procedures, and tools.
<b>Report</b> results and findings	Compile all analysis results and report out on findings, observations, and future opportunities for improvement on the ways that content- and conduct-related risks are identified and mitigated at the company moving forward.	Share key observations and findings with partners to facilitate collaborative development of industry standards and perspective.

## Risk Assessment Inputs

LinkedIn gathered and reviewed the following inputs to inform and substantiate ratings assigned in the systemic risk assessment.

- **DTSP questionnaire responses:** LinkedIn modified the April 2023 Digital Trust and Safety Partnership (DTSP) questionnaire to incorporate additional DSA systemic risk considerations and solicited responses to the questionnaire from internal stakeholders across the Trust and Safety, Legal, Privacy, Product, and Engineering teams.

- **Product and Feature Changes:** LinkedIn collated an inventory of material changes to in-scope features during the assessment period.
- **Internal Consultations:** Following review of the DTSP questionnaire responses, LinkedIn conducted in-depth workshop sessions with internal stakeholder groups to solicit more detailed information on specific topics or issues related to risk prevalence as well as best practices and mitigating factors with greater potential to impact LinkedIn's risk profile.
- **External Consultations:** LinkedIn regularly engages, both individually and with parent company Microsoft, with external stakeholders, including civil society organizations, to receive feedback on platform Trust and Safety as well as to discuss best practices for addressing risk. LinkedIn conducted additional civil society engagement specific to this systemic risk assessment and the feedback from these engagements informed the assessment.
- **Internal Content Moderation Metrics:** LinkedIn considered internal metrics related to content moderation to inform both assessment of prevalence as well as effectiveness of mitigations.
- **Transparency Reporting Metrics:** LinkedIn considered metrics reported through Transparency Reporting to inform the assessment of both probability and effectiveness of mitigations.
- **Open-Source Data on Public Discourse Related to LinkedIn and Systemic Risk Areas:** LinkedIn reviewed a collection of social and digital media articles and conversations around the LinkedIn platform and each systemic risk area to identify trends in areas of public discourse and/or concern.
- **Risk and Mitigation Summaries:** LinkedIn stakeholders developed brief summaries of LinkedIn's controls and mitigations specific to each risk area. LinkedIn documented one-pagers for each risk area detailing risk manifestations and mitigations.
- **Policies and Publications:** LinkedIn reviewed relevant external policies, public practices, and other authoritative sources including blogs to identify additional policies and initiatives most relevant to the risk assessment.
- **Authoritative sources:** LinkedIn reviewed publicly available sources considered reliable due to their expertise and reputation, such as regulatory sources, reputable public opinion polling, and think tank research, to inform the objective assessment of severity.
- **DSA Risk and Control Matrix:** LinkedIn considered the DSA risk and control matrix as well as the results from the DSA audit to inform the existence and effectiveness of mitigations and Residual Risk scores.

## ***Inherent Risk***

**Inherent Risk** is determined by two factors - probability and severity. Calculating Inherent Risk involves assessing both the likelihood of an event occurring and the potential impact of that event. The greater the probability and severity of an event, the higher the Inherent Risk associated with it.

**Probability:** LinkedIn conducted a **data-driven probability assessment** to evaluate the likelihood of certain events occurring on the platform absent mitigations by analyzing relevant data, including public incident data, Transparency Report metrics, and internal metrics.

- Probability is the likelihood of the probable worst case of the risk occurring if management action and controls / mitigations fail or do not exist.
- Probability depends on factors such as volume of content flagged or removed for identified risk events.
- Each score assigned to probability in calculating Inherent Risk, ranging from 1 to 5, corresponds to a specific range of likelihood of the event occurring, with higher scores indicating a higher likelihood of the event.
- Items that substantiate probability include External consultations, Internal content moderation metrics, Transparency reporting metrics, and Open-source data on public discourse related to LinkedIn and systemic risk areas

*Probability Rating Scale*

Figure 19

Description	Score	Rating
The risk event or circumstance is relatively certain to occur	5	Expected
The risk event or circumstance is highly likely to occur	4	Highly Likely
The risk event or circumstance is likely to occur	3	Likely
The risk event or circumstance occurring is possible but not likely	2	Not Likely
The risk event or circumstance is only remotely probable	1	Remote

**Severity:** LinkedIn conducted an objective systems-based analysis to determine severity, considering the complexity, scale, and gravity of impact to assign an overall severity rating.

- Severity is the probable worst case of a risk's impact on the world or society if management action and controls / mitigations fail or do not exist.
- Severity is calculated once for each risk; and takes consideration of complexity (impact on economic, security, political, societal, wellbeing, and environmental systems), scale (global, regional, country, locality, individual), and gravity (remediability)
- Items that substantiate severity include External consultations, Internal consultations, and Authoritative sources

*Severity Rating Scale*

Figure 20

Description	Score	Rating
Impact that could cause critical, irremediable harm, damage, or loss	5x	Critical
Impact that could cause significant irremediable harm, damage, or loss	4x	High
Impact that could cause some harm or disruption but is generally manageable or remediable	3x	Moderate
Impact that could cause limited harm or disruption	2x	Low
Impact that has little or no consequence	1x	Minimal

*Inherent Risk Rating Scale*

Figure 21

Description	Score	Rating
A risk that is relatively certain to occur, would have a severe impact if it occurred, and requires immediate action to manage or mitigate.	$\geq 17$	Critical
A risk that is highly likely to occur, would have a significant impact if it occurred, and requires urgent action to manage or mitigate.	11 – 17	High

A risk that is likely to occur, would have a noticeable impact if it occurred, and requires some action to manage or mitigate.	6 – 11	Moderate
A risk that is not likely to occur, would have a limited impact if it occurred, and may require minimal action to manage or mitigate.	2 – 6	Low
A risk that is only remotely probable, would have a minimal impact if it occurred, and may not require any action to manage or mitigate.	<2	Minimal

## Residual Risk

**Residual Risk:** By multiplying the Inherent Risk score and mitigation score, a Residual Risk score is calculated that reflects the amount of risk remaining in a given situation. This provides a final assessment of the potential risk once mitigations have been applied.

**Mitigation:** LinkedIn leveraged the DTSP Safe Assessment Framework – aligning controls and mitigations to the 35 Best Practices - and Maturity Scale for mitigations, while evaluating if they are reasonable, proportionate, and effective. Items that substantiate mitigation include DTSP questionnaire responses, Internal consultations, External consultations, Internal content moderation metrics, Mitigation summaries, and Policies and publications.

Figure 22

DTSP Inventory of 35 Best Practices				
Product Development	Product Governance	Product Enforcement	Product Improvement	Product Transparency
PD1: Abuse Pattern Analysis	PG1: Policies & Standards	PE1.1: Roles & Teams	PI1: Effectiveness Testing	PT1: Transparency Reports
PD2: Trust & Safety Consultation	PG2: User Focused Product Management	PE1.2: Operational Infrastructure	PI2: Process Alignment	PT2: Notice to Users
PD3: Accountability	PG3: Community Guidelines/Rules	PE1.3: Tooling	PI3: Resource Allocation	PT3: Complaint Intakes
PD4: Feature Evaluation	PG4: User Input	PE2: Training & Awareness	PI4: External Collaboration	PT4: Researcher & Academic Support
PD5: Risk Assessment	PG5: External Consultation	PE3: Wellness & Resilience	PI5: Remedy Mechanisms	PT5: In-Product Indicators
PD6: Pre-Launch Feedback	PG6: Document Interpretation	PE4: Advanced Detection		
PD7: Post-Launch Evaluation	PG7: Community Self Regulation	PE5: User Reporting		
PD8: User Feedback		PE6.1: Enforcement Prioritization		
PD9: User Controls		PE6.2: Appeals		
		PE6.3: External Reporting		
		PE7: Flagging Processes		
		PE8: Third Parties		
		PE9: Industry Partners		

[https://dtspartnership.org/wp-content/uploads/2022/07/DTSP\\_Report\\_Safe\\_Assessments.pdf](https://dtspartnership.org/wp-content/uploads/2022/07/DTSP_Report_Safe_Assessments.pdf)



DTSP Maturity Rating Scale

Figure 23

Description	Score	Rating
A rating of Ad Hoc is assigned when execution of best practices is incomplete, informal, or inconsistent.	10%	Ad Hoc
A rating of Repeatable is assigned when execution of best practices occurs without standardized processes. Organizations aim to document more formalized practices.	30%	Repeatable
A rating of Defined is assigned when execution of best practices occurs with defined and documented processes. Processes are more proactive than reactive and are implemented across the organization.	50%	Defined
A rating of Managed is assigned when execution of best practices is defined, documented, and managed through regular reviews. Organizations use feedback to continuously mitigate process deficiencies.	65%	Managed
A rating of Optimized is assigned when execution of best practices promotes Trust and Safety in every aspect. Processes are continuously improved with innovative ideas and technologies.	80%	Optimized

[https://dtspartnership.org/wp-content/uploads/2022/07/DTSP\\_Report\\_Safe\\_Assessments.pdf](https://dtspartnership.org/wp-content/uploads/2022/07/DTSP_Report_Safe_Assessments.pdf)

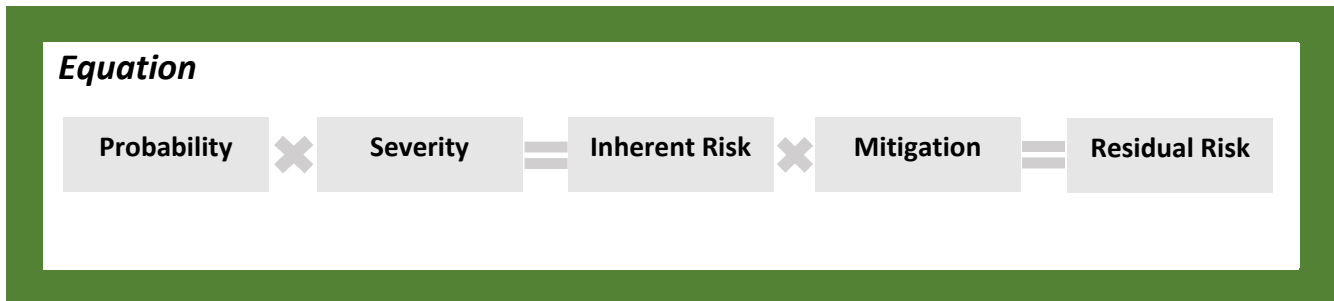
Residual Risk Rating Scale

Figure 24

Description	Score	Rating
A risk that is relatively certain to occur, would have a severe impact if it occurred, and requires immediate action to manage or mitigate.	>=17	Critical
A risk that is highly likely to occur, would have a significant impact if it occurred, and requires urgent action to manage or mitigate.	11 – 17	High
A risk that is likely occur, would have a noticeable impact if it occurred, and requires some action to manage or mitigate.	6 – 11	Moderate
A risk that is not likely to occur, would have a limited impact if it occurred, and may require minimal action to manage or mitigate.	2 – 6	Low
A risk that is only remotely probable, would have a minimal impact if it occurred, and may not require any action to manage or mitigate.	<2	Minimal

## Sample Scoring Calculation

Figure 25



### Inherent Risk

- If the probability of occurrence of a risk is Highly Likely -> probability score of 4
- If the severity of a risk is Moderate -> severity score of 3
- Based on the probability score of 4 and severity score of 3 -> Inherent Risk score of 12 (= 4 x 3 ) -> inherent rating of High

### Residual Risk

- If the inherent risk rating score is 12
- If the mitigating maturity rating is "Managed" -> mitigation score of 65%
- Based on the inherent rating score of 12 and mitigation score of 65% -> Residual Risk score of 4.2 (= 12 x ( 1 – 0.65 ) ) -> Residual Risk rating of Low

## Appendix B: List of Article 35 mitigations

The following areas have been identified for heightened focus for the coming year, in addition to the ongoing risk mitigation monitoring and refinement processes. LinkedIn is committed to addressing each of the topics identified below to (1) further define the associated risks, (2) explore enhanced mitigation options, (3) vet mitigation options as appropriate with internal and external subject matter experts, (4) develop action plans as needed to address focus areas, and (5) track implementation of the developed action plans.

Figure 26

Mitigation Title	Mitigation Description
Monitoring of Mitigations	LinkedIn plans to identify opportunities to more directly map classifiers and internal risk taxonomy to systemic risk areas in order to enhance the monitoring of mitigations.
Refined Models to Detect Additional Fake Behaviors	LinkedIn will invest in further expanding upon the improvements made its models built to detect fake accounts to achieve greater fidelity on detection of additional inauthentic activity.
Continuous Tuning of Classifiers	LinkedIn plans to continue tuning and moderating its classifiers, multilingual models, and enhanced automation capabilities, particularly for risks related to Human Dignity and Consumer Protection and Fraud.
Continuous Improvement of Election Response	LinkedIn plans to conduct post-mortem reviews following high priority elections to identify opportunities to tighten and refine LinkedIn's election response capabilities.
Enhancements to Verification	LinkedIn plans to enhance protective measures to reduce the impacts of verification abuse, as appropriate, based on additional review criteria.
Fraudulent Accounts	LinkedIn plans to invest in efforts to identify fraudulent accounts, malware, and other Tactics, Techniques, and Procedures that may be used for fraud and/or degrading consumer protection.
Product Launch	LinkedIn plans to explore opportunities to further enhance the pre-launch product review and launch process.
Researcher Access	LinkedIn plans to continue investing in and being transparent in its practices related to researcher access to data to researchers, members, and the public.