Systemic Risk Assessment

August 2023

Assessment Report

Introduction

LinkedIn Ireland Unlimited Company—the provider of LinkedIn's services in the European Union—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 one assessment, through August 2023, 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 European Union. Although the assessment considered the specific risk categories and influencers outlined in Article 34 with a focus on their impact in the European Union, LinkedIn also evaluated these risks and influencers 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. 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. However, as a members-first platform committed to delivering the best member and customer experience with a focus on trust, LinkedIn focuses significant resources on ensuring that what members see on LinkedIn is productive, successful, and free of inappropriate and unwanted content or behavior. To that end, LinkedIn invests heavily in numerous trust and safety domains to proactively enhance the safety, security, privacy, and quality of the LinkedIn user experience.

Notable LinkedIn investments include

- The continued development of advanced models for detecting and preventing the creation of fake or inauthentic accounts, fraudulent activities, or sharing of critically harmful content, including child sex 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 and additional testing and trials of additional 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) fraud, (2) negative impact on the right to human dignity, and (3) the dissemination of "illegal" content. Accordingly, LinkedIn has developed 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. 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. In the last transparency reporting period, LinkedIn's automated defenses blocked 87.4% of detected fake accounts during the account registration process and 99.6% of detected fake accounts before a member report.

- LinkedIn is innovating to reduce hate speech, bullying, harassment, and other harmful content linked to Human Dignity on the platform. These initiatives include the initial roll-out of "Kindness Reminders" to members that encourage members to rethink potentially harmful comments before sharing as well as an optional advanced safety feature for European members that allows LinkedIn's automated machine learning models to detect sexual harassment and certain other abusive content[1] in messages received by the member, giving the recipient the ability to view and report the message to LinkedIn for review if desired.
- LinkedIn has robust policies and processes in place to address illegal content both proactively, in the case of CSAM, illegal and regulated commercial activity, hate speech, and content promoting terrorism, and reactively, in the case of intellectual property infringement and defamation. 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.

In the upcoming year, an area of priority for LinkedIn enhancements is artificial intelligence (AI) and generative AI in addressing risk presented or amplified by generative AI, in harnessing AI's capability to enhance content moderation processes, and in ensuring continued and enhanced responsible implementation of AI at LinkedIn.

AI is not new to LinkedIn. LinkedIn has long used AI to enhance its members' professional experiences. By leveraging the power of AI, LinkedIn helps 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 risks and potential for harm. That's why, consistent with its commitment to build a trustworthy platform, LinkedIn is committed to using AI responsibly. Earlier this year, inspired by, and aligned with, Microsoft's leadership in Responsible AI, LinkedIn shared the Responsible AI Principles that it uses 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.

[1] LinkedIn automatically scans all content for CSAM, malware, and phishing.

- **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.

Over the next year, LinkedIn will continue to advance its Trust and Safety efforts aligned with its AI principles as well as Microsoft's commitments to advancing responsible innovation.

Background

LinkedIn is a real-identity, specific-purpose platform, designed specifically for professionals. It is a networking tool that enables members to establish their professional identities online, connect with other professionals, and build meaningful 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:

1. LinkedIn is a real identity platform, where members must use their real or preferred professional name, and the content they post is visible to the member's 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.
2. LinkedIn operates under standards of professionalism, which are reflected in both 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.
3. LinkedIn services are tailored toward professionals and businesses. The majority of advertisements are business-to-business rather than business-to-consumer. Members must attest to being over the minimum age to apply for 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 significant resources in developing techniques to prevent the creation of fake profiles and is providing additional identity verification options 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 is what 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 to ensure its employees make the best possible decisions to protect LinkedIn members and maintain their trust. These principles are

LinkedIn provides its members with **clarity, consistency**, and **control** over their data. Simply, 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 of 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.

Finally, 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.

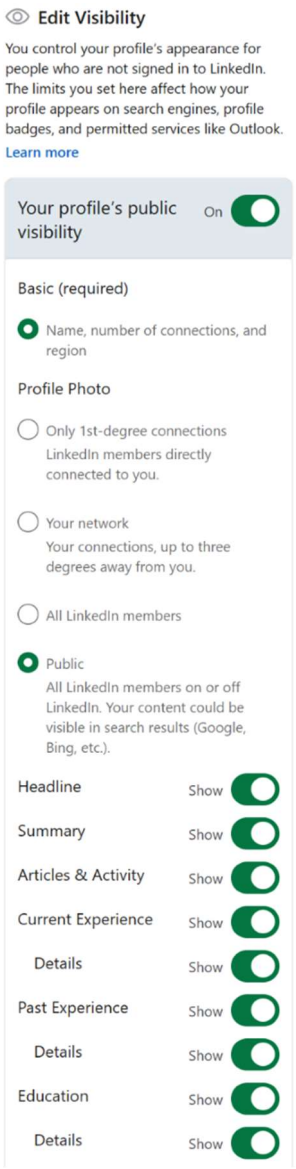
Products At a Glance

LinkedIn has members in more than 200 countries and territories worldwide and operates a diversified business model with revenue coming from its Talent Solutions, Marketing Solutions, Sales Solutions, and Premium Subscriptions products. While the 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.

Profile: Each member has a LinkedIn "profile" associated with their LinkedIn account. Functioning as an online resume 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 name and include accurate information within their profile. 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, 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 key portions of profiles are by default accessible via search engines to non-members or members in a logged-out state, members can control this via settings like those displayed in Figure 1.

Figure 1



Feed: LinkedIn focuses on providing its members with a relevant Feed that encourages them to engage with people and topics they value. The goal is to provide each professional with professionally-relevant content. The LinkedIn Feed is the place on the platform where members go to view content, most of which is provided by other members. A piece of 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, and 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 violates 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 their settings).

A member's Feed experience is dynamic and specific to that logged-in member, based on, among other things, their 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 connect with those organizations and learn more about them, such as their respective brand, business or school, employee culture, and career opportunities. 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 their organization, and job posts) to build and promote the organization's brand to the professional audience on LinkedIn.

Groups: A Group is a shared place that 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 a discussion or a conversation 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 a message to a connection or multiple connections in a Group chat, 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 a connection request with a personal message to other members. In addition, "InMail" messages are a premium (for a fee) feature that require the use of a credit 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 the job via LinkedIn, though some only offer the option to be taken to the employer's application website to apply directly there. Recommended jobs are curated for the

individual member based on their skills, job titles, education, interests, and various other attributes, including the 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.

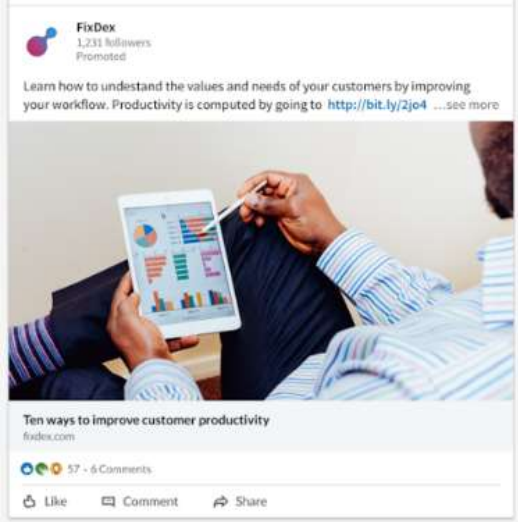
Search: Using LinkedIn's in-product search feature located at the top of most LinkedIn Pages, LinkedIn members can conduct a search across most of the platform, including profiles of other LinkedIn members. A searching member has the option to "See all results" for a keyword search or to narrow a search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

Figure 2

LinkedIn Marketing Solutions: 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:[2]

- *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" or "Sponsored" that appear in the upper-left corner of the post.

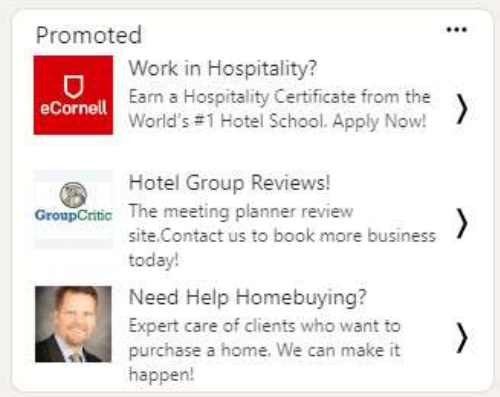
Figure 3



- *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.

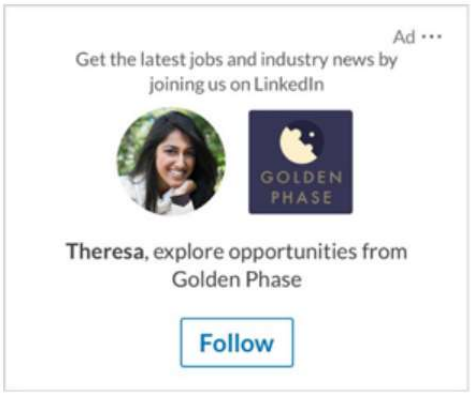
[2] There is a fourth type shown to members outside of the EU known as Sponsored Messaging. These appear as messages in the member's LinkedIn inbox that are clearly labelled as "sponsored."

Figure 4



- *Dynamic Ads*: These are personalized ads using the member's profile image, name, or job function to create customized ads for each 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.

Methodology

LinkedIn aligned its risk assessment process and methodology with the Digital Trust & Safety Partnership (DTSP) Safe Framework, as an 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 to consider probability and severity as well as reasonableness, proportionality, and effectiveness of mitigations informed by the best available information and scientific insights.

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 eleven risk areas identified in Article 34. The methodology is summarized here and further expanded in Appendix A: Methodology Expanded.

Discovery

As part of Discovery, LinkedIn modified the Safe Assessment questionnaire to meet the requirements of the DSA Systemic Risk Assessment; conducted stakeholder engagement and awareness-building; collected responses to the questionnaire; reviewed internal supporting documentation; examined external authoritative sources and case studies; collected relevant public policies and publications; and summarized relevant internal controls.

Inputs to Risk Assessment:

1) LinkedIn internal stakeholders' responses to the modified Safe Assessment questionnaire
2) Guidance from authoritative sources, such as regulators, Think Tanks, or non-governmental organizations, on severity of systemic risks
3) Collection of LinkedIn published policies and official communications
4) Summarized mitigations by risk area

Identification

As part of the Identification phase, LinkedIn: conducted workshops with key stakeholders across the organization to deep dive into key risk areas and mitigation efforts and to understand specific processes, tools, and operational practices; conducted tailored engagements with civil society organizations familiar with digital rights concerns in the European Union to discuss the risk assessment process, areas of concerns, and best practices for mitigation; gathered key metrics related to trust and safety policy enforcement; deconstructed Transparency Reporting metrics; and gathered open source information related to trends, patterns, and potential systemic risk related to LinkedIn.

Inputs to Risk Assessment:

5) Notes from Internal Consultations
6) Notes from External Consultations
7) Internal content moderation metrics
8) Transparency Reporting metrics
9) Open-source data on public discourse related to LinkedIn and systemic risk areas

Assessment

LinkedIn assessed the probability and severity of each of the eleven 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 nine 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 by the systemic risk area (geographic, political, security, environmental, 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. 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; linguistic and cultural considerations; 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.

Testing

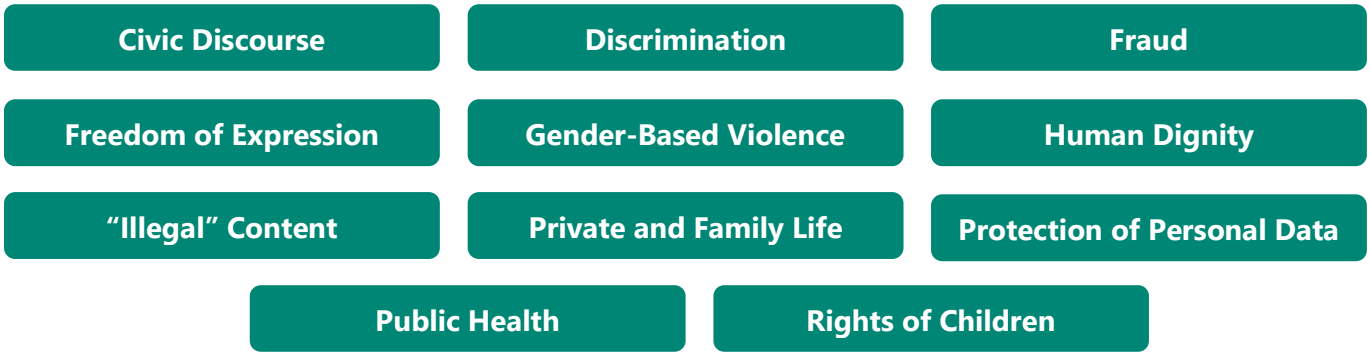
LinkedIn routinely conducts quality assurance activities and tests the effectiveness of its algorithmic systems and content moderation processes. LinkedIn relied on existing mitigation testing procedures in place for the purposes of this Risk Assessment.

Reporting

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.

Summary of Results

LinkedIn assessed systemic risk stemming from the use or misuse of the platform and services within eleven key risk areas. The full names and definitions are included in Appendix A: Methodology Expanded.



Probability across Risk Areas

While LinkedIn's platform and services can be host to a variety of systemic risks, the inherent probability or likelihood of occurrence for most risk areas is assessed at relatively low levels, considering the 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. The volume of violative content on LinkedIn is consistently lower than industry norms.

Considering a review of content flagged and removed from the platform, by automated classifiers and human reporting, as well as the volume of public discourse related to LinkedIn and each risk area, as well as trends in public concern and emerging threats across the industry, LinkedIn assessed Fraud as having the highest relative probability of occurrence on the platform, followed by "Illegal" Content, Human Dignity, and Public Health. Personal Data, Civic Discourse, Freedom of Expression, Discrimination, and Gender-Based Violence were assessed at a slightly lower level of probability. Private and Family Life was assessed at a lower level still. And Rights of Children was assessed with the lowest level of probability, considering both the low numbers of instances of violative content found on the platform as well as the low number of members under the age of 18.

Impact of Influencers

In the assessment of each risk area, LinkedIn considered the potential impact of all Article 34 factors on probability, severity, and mitigation effectiveness, including: 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: In its Feed, LinkedIn strives to broadly distribute only high-quality content—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.

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 artificial intelligence 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. 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. LinkedIn has implemented robust training and QA programs for content moderators including regular audits on a sample of content reviewed by moderation teams, regular group calibration sessions to address common error trends, and coaching for lower performers.

Advertisements: LinkedIn maintains and enforces policies related to advertising on the platform, including LinkedIn prohibits political advertisements; LinkedIn has policies that limit health-care related advertisements to come from certified retailers, to target healthcare professionals, and to limit distribution to jurisdictions where such advertisements are legal; LinkedIn requires Ads customers to certify they will not use LinkedIn to discriminate based on gender or other protected characteristics before accessing the limited demographic form fields (some of which are not available at all for talent-related ad campaigns). 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.

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 87.4% of detected fake accounts during the account registration process and 99.6% of detected fake accounts 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. With regard to European elections, LinkedIn engages with stakeholders as needed to address rising concerns relative to election cycles. LinkedIn does not allow political advertisements.

Alignment of Investments to Risk

Considering the higher likelihood of fraudulent accounts, content, or activity on the platform, LinkedIn has invested significant resources in developing advanced techniques to detect and address fraudulent behaviors.

Fraud

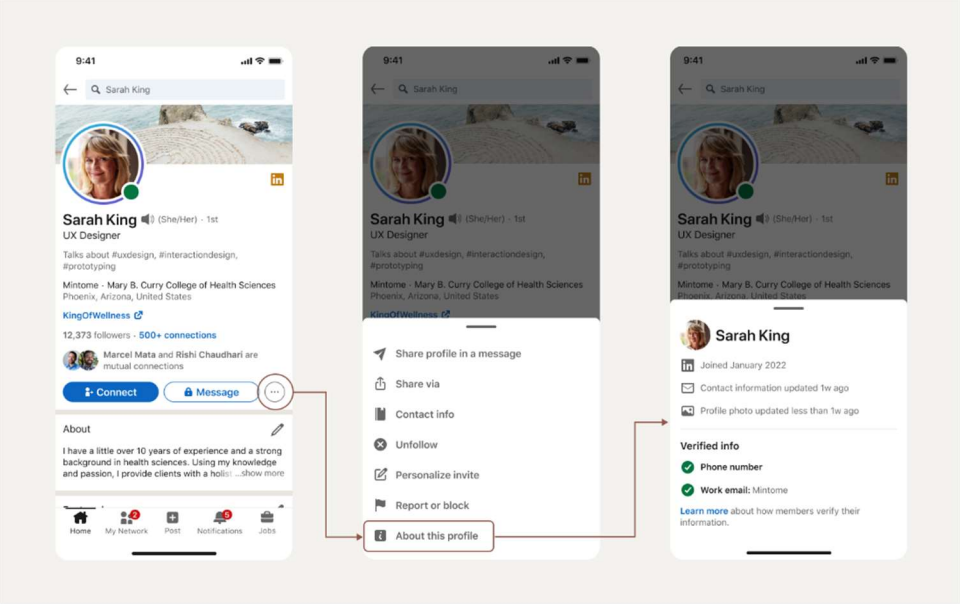
LinkedIn has invested in robust policies, policy enforcement mechanisms, and advanced detection capability to mitigate the risk of Fraud on the platform, which includes risk related to scamming, defrauding, spamming, or falsifying information, including the creation of fake accounts and use of bots or other deceptive commercial practices on LinkedIn.

Mitigations applied to reducing systemic risk related to Fraud are assessed as Managed on the DTSP maturity scale as LinkedIn best practices related to this risk area are defined, documented, and managed through regular reviews. LinkedIn uses feedback to continuously mitigate process deficiencies in its approach to detecting and preventing Fraud on platform. Despite significant investments, LinkedIn categorizes its approach to Fraud as Managed rather than Optimized given the rapidly evolving nature of Fraud risks as they manifest on platform and the need to continually assess and enhance detection techniques.

In addition to the DTSP Best Practices LinkedIn has effectively implemented across all risk areas, which are further described under Mitigation, LinkedIn has invested in additional capabilities to mitigate the rapidly evolving risks associated with Fraud on the platform.

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 represents themselves, 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 is strictly prohibited, as is promoting pyramid schemes or romance scams, sharing malicious software, or engaging in phishing attempts.

Figure 6



LinkedIn has implemented several features to protect its users from inauthentic accounts. Members can choose to only receive communications from people they know, thus allowing them to control who can contact them. Additionally, LinkedIn created an "About this profile" feature that will show users when a profile was created and last updated, along with whether the member has verified a phone number and/or work email associated with their account, which can help to identify the authenticity of the members with whom they are interacting.

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. LinkedIn is also in the process of implementing a variety of verification features where a member verifies certain data on their profile. For example, ID verification currently is ramping in some countries outside of the European Union, where a member can verify their identity with third-party partners for use on LinkedIn to show the member's profile name matches their identification documents. Work email verification has already been globally implemented, allowing members to verify their workplace using their current company email address and confirming with a code sent to that email.

LinkedIn also provides the option to companies to require job posters to verify their work emails. This feature significantly reduces potential fraud within Jobs. LinkedIn also maintains Notice of Copyright Infringement and Notice of Trademark infringement forms for use when users in good faith believe that content posted by a member has infringed upon their rights.

To prevent fake accounts on LinkedIn, the Anti-Abuse team has built automated detection systems. The first line of defense is registration scoring, which uses machine-learned models to evaluate every new registration attempt and give it an abuse risk score. This score determines if the account can be created immediately, needs to be challenged for verification, or is prevented from being created altogether. This system is effective at blocking bulk account creation attempts. However, not all fake accounts are created in bulk, so LinkedIn uses downstream models that group accounts together by shared attributes and identify clusters with statistically abnormal data distributions, which can indicate the presence of a single bad actor creating or controlling multiple accounts. Additionally, activity-based models look for specific types of bad behavior or anomalous activity that may be indicative of a fake account. Despite these automated defenses, a human element is still necessary to catch any fake accounts that have evaded the models. LinkedIn also deploys deep learning models to identify AI-generated profile images. LinkedIn relies on member reports and a team of investigators to uncover accounts that may have evaded all levels of defense.

Higher Residual Risk

Considering the probability, severity, and maturity of mitigations, the risk assessment identified the areas of Public Health and Human Dignity as carrying slightly higher residual risk relative to other risks on the LinkedIn platform.

Public Health

Public Health includes the risk of health misinformation, promotion of legal but harmful substances or practices, behavioral addictions, and content encouraging self-harm. Inherent Severity for Public Health was assessed at the highest level considering the potential gravity of risks within these categories, as well as the potential to impact multiple systems of society (wellbeing, societal, environmental) at the country and regional level.

While still low due the professional nature of LinkedIn, the inherent Probability of risks to Public Health on the platform were assessed at a slightly higher likelihood based on minimal content that was removed from the platform related to vaccine misinformation.

Mitigations applied to reducing systemic risk related to Public Health are assessed as Defined on the DTSP maturity scale as LinkedIn processes related to Public Health are defined and documented, more proactive than reactive, and implemented across the organization.

In addition to the DTSP Best Practices LinkedIn has effectively implemented across all risk areas, which are further described under Mitigation, 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 self-harm and suicide, as well as healthcare advertising.

LinkedIn enables members to report content that portrays an intent to commit suicide/self-harm or violence against others. These reports receive priority review and are evaluated by specialists to determine how credible and imminent the threat is. When a report is evaluated as credible and imminent, the specialist will notify appropriate law enforcement authorities, typically within hours of evaluating the report. LinkedIn also provides communications to members who report the content or whose content is reported.

LinkedIn has policies related to healthcare advertisements that limit drug advertisements to approved retailers and limit the targeting of such advertisements to healthcare-related professionals, instead of consumers or a general audience. 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 provides guidance to healthcare providers to enhance digital healthcare advertising campaigns. 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.

Risks related to Public Health are not high, considering the nature and functioning of LinkedIn's platform does not allow for rapid and wide dissemination of content that has not been selected for quality. Content shared by members is typically only available to members within their network, which, for the purposes of any inauthentic accounts that may slip past LinkedIn's proactive defenses is likely to be small.

Human Dignity

Human Dignity includes the risks of threats or promotion of violence, hate speech, extortion, bullying & harassment, or promotion of human trafficking. Inherent Severity for Human Dignity was also assessed at the highest level considering the potential gravity of risks within these categories, namely threats of violence and promotion of human trafficking, which can severely impact individuals and locality in the wellbeing, security, and societal systems.

While still low due the professional nature of LinkedIn, the inherent Probability of risks to Human Dignity on the platform were assessed at a slightly higher likelihood than other risks based on content that was detected and removed from the platform related to harassment and hate speech.

In addition to the DTSP Best Practices LinkedIn has effectively implemented across all risk areas, which are further described under Mitigation, LinkedIn has invested in new technologies to detect and address hate speech and harassment. LinkedIn scans public-facing content for hate speech, harassment, and other content that may impact Human Dignity and flags potential violative content for human review. LinkedIn 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, disparages other LinkedIn members, or is otherwise abusive or uncivil.

LinkedIn allows members to call out actual or perceived hateful, prejudicial, or discriminatory conduct, provided such content doesn't otherwise violate its Harassing and inflammatory content policies.

LinkedIn has also deployed "Kindness Reminders" to the platform for comments posted in English. For this feature, if LinkedIn detects potentially harmful content in the drafted comment, it will prompt the creator with a warning asking if they are sure they would like to proceed with sharing the potentially harmful content.

Risks related to Human Dignity are not high, considering the nature and functioning of LinkedIn's platform does not allow for rapid and wide dissemination of content that has not been selected for quality. Content shared by members is typically only available to members within their network, which, for the purposes of any inauthentic accounts that may slip past LinkedIn's proactive defenses is likely to be small.

Lower Residual Risk

Considering the probability, severity, and maturity of mitigations, the risk assessment identified the area of Rights of Children as carrying significantly lower residual risk on the LinkedIn platform.

Rights of Children

Risk to the Rights of Children includes risk that content promoting child sexual abuse appears on the platform, that LinkedIn uses children's data without parental consent, and that children are exposed to harmful content on LinkedIn. Inherent Severity of this risk was assessed at the highest level considering 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.

Inherent Probability of this risk area was assessed at the lowest level considering the low number of LinkedIn members under the age of 18 and the very low numbers of CSAM content identified on the platform.

Mitigations applied to reducing systemic risk related to Rights of Children are assessed as Managed on the DTSP maturity scale as LinkedIn best practices related to this risk area are defined, documented, and managed through regular reviews and continuous enhancement. In addition to the DTSP Best Practices LinkedIn has effectively implemented across all risk areas, which are further described under Mitigation, LinkedIn has invested in robust policies, policy enforcement mechanisms, and advanced detection capabilities to effectively mitigate the risk to Rights of Children on the platform.

Anyone under the age of 16 is specifically prohibited by LinkedIn's terms of service 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.

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. Additionally, all members have additional controls to restrict the privacy of their information, including whether their profile is shown in Search Engines.

LinkedIn has a robust policy against 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), and the associated account is placed in restriction.

LinkedIn has also published robust policies and developed advanced detection capabilities related to other 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. Considering the professional focus of the platform, LinkedIn removes content depicting adult nudity or sexual activity.

Additional Highlights

The remaining systemic risk areas were assessed as having a low residual risk, considering both the lower likelihood of occurrence on the platform as well as the assessment of the applied mitigation measures as Managed. In addition to the DTSP Best Practices, further elaborated in the Mitigations section, LinkedIn has applied unique measures to address each of the following potential systemic risk areas.

Civic Discourse

To mitigate risks related to Civic Discourse, which can include election misinformation, disruption of electoral processes, and incitement to violent riots and considers the influence of coordinated manipulation and inauthentic activity, LinkedIn partners with stakeholders and has invested in new technologies to combat inauthentic behavior that may contribute to this risk.

LinkedIn prohibits political advertisements on its platform. LinkedIn works closely with Microsoft's Democracy Forward team, which is an industry leader in developing tools to further civic integrity. LinkedIn works with peer companies and other stakeholders to receive indicators related to fake accounts created by state-sponsored actors. Its Trust and Safety team uses information from a variety of sources to detect and remove fake accounts and, when necessary, work with peers and other stakeholders to find and remove state-sponsored activity.

To respond to the ever-changing threat landscape, LinkedIn's teams continually invest in new technologies for combating inauthentic behavior on the platform. LinkedIn is investing in artificial intelligence 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. More specifically, LinkedIn acts vigilantly to maintain the integrity of all accounts and to ward off false account activity through

- Maintaining a dedicated Anti-Abuse team to research emerging trends and key risks and develop tools to address them;
- AI to detect inauthentic activity and communities of fake accounts;
- Hash matching for known instances of deepfake content;
- Detecting whether members have installed known prohibited automation software; and
- Using automated systems to detect and block automated activity.

LinkedIn's Trust and Safety team proactively monitors for threats, including those around elections and civic integrity. LinkedIn also works with third parties to enforce its policies. For example, it utilizes external fact checkers in a variety of languages as needed. Furthermore, LinkedIn works with peer technology companies to exchange information about the tactics, techniques, and procedures (TTPs) of sophisticated abusive accounts and advanced persistent threat (APT) actors. LinkedIn aims to educate its members about civic discourse, electoral processes, and public security through its global team of news editors. These editors provide each member with relevant, timely information sourced from credible news sources that educate LinkedIn members about facts and news events.

Discrimination

To mitigate risks related to Discrimination, which include risk of discriminatory content on the platform, discriminatory activities (e.g., in Jobs), and potential bias and discrimination in LinkedIn's algorithmic systems, LinkedIn has invested significant energy into ensuring "two members who are equally qualified should have equal access to opportunity" on LinkedIn.

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 any lookalike job postings on the system. Users can also report for review by LinkedIn any job post on LinkedIn that they believe may be discriminatory.

In-product functionality is also utilized 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.

LinkedIn is also committed to measuring and preventing bias and discrimination in its algorithmic systems. 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 subgroups.

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 conducts AB 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 ensure the introduction of bias into LinkedIn's algorithms.

7Freedom of Expression

LinkedIn carefully weighs trade-offs between freedom of expression and online harms, erring on the side of speech where possible. LinkedIn operates under standards of professionalism, which are reflected in both content policies and enforcement, as well as in content prioritization and amplification. LinkedIn has policies that prohibit anything but safe, trusted, and professional content. And LinkedIn chooses to prioritize only content that meets LinkedIn's high-quality standards that will enhance members' professional experiences.

LinkedIn prioritizes member feedback with regard to content visibility and offers members the opportunity to curate their own experiences in Feed. LinkedIn provides the opportunity for members 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 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.

As noted, LinkedIn conducts Fairness Checks on Feed distribution and ranking, and monitors for unexpected anomalies in distribution of content to track and ensure the introduction of bias into LinkedIn's algorithms. LinkedIn is also biannually audited for its commitments to human rights, including free expression, by the Global Network Initiative and 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 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 country law experts as needed.

LinkedIn also has a robust QA program designed to help improve reviewer performance and consistency and conducts QA audits, provides coaching to lower performers, and addresses error trends in group calibration sessions. LinkedIn is also proactively revamping quality assurance efforts on human moderator review of harmful content, conducting AB testing to determine the effectiveness of various tactics such as jury review and masking previous reviewer information.

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 grants roughly 40% of requested appeals, which makes up only .05% of content decisions. This means both that roughly 99.9% of the time LinkedIn content moderation decisions are not appealed, but when users do appeal, LinkedIn is responsive to user voice.

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.

A cross-functional team of stakeholders from across LinkedIn's broader Trust organization (including members of the Legal, Trust & Safety, 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 its Transparency Center (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 the 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 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).

Gender-Based Violence

LinkedIn is committed to addressing risk related to Gender-Based Violence on the platform, which may include gender-based hate speech, sexual harassment, promotion of human trafficking or criminal sexual acts, or other forms of violence against women and girls. LinkedIn scans public-facing content for gender-based hate speech, sexual harassment, and other content that may impact Gender-Based Violence and flags potential violative content for human review.

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 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 Harassing and inflammatory content policies.

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.

In 2022 and 2023, LinkedIn conducted surveys to better understand member awareness of what constitutes unwanted sexual advances and sexual harassment and conducted an extensive in-product (and in language) educational campaign to introduce various relevant safety-related features and increase awareness regarding acceptable user behavior. LinkedIn has also deployed "Kindness Reminders" to the platform in English. For this feature, if LinkedIn detects potentially harmful content in draft comments, it will prompt the creator with a warning asking if they are sure they would like to proceed with sharing the potentially harmful content.

Dissemination of "Illegal" Content

LinkedIn has invested in robust policies, policy enforcement mechanisms, and advanced detection capability to effectively mitigate the risk to dissemination of "Illegal" Content on the platform, which includes intellectual property infringement; defamation; CSAM; hate speech; terrorist or violent extremist content; unlawful non-consensual sharing of private images; online stalking; or promotion or sale of illegal, dangerous, or counterfeit goods, services, or animals.

LinkedIn has a responsive approach to member requests for removal of content which they believe infringes upon their intellectual property or which is defamatory. Users can submit complaints, LinkedIn processes and investigates each claim, and takes, documents, and communicates the appropriate action to the affected user(s).

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 uses hash matching to detect and remove previously identified terrorists 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 has a zero-tolerance policy for CSAM. LinkedIn scans all public-facing and privately shared videos and images for CSAM, 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), and the associated account is placed under restriction. Once CSAM has been identified, LinkedIn also applies algorithmic models to identify and remove near-duplicates from the platform.

For certain egregious violations of its Professional Community Policies (e.g., child sexual abuse material, terrorism, extremely violent content, egregious sexual harassment), LinkedIn may permanently restrict a member's account after a single violation.

Private and Family Life

LinkedIn's Managed mitigations for Protection of Personal Data carry over to many of the risks identified under Risks to Private and Family Life, which include malicious sharing of or exploitation of sensitive or private data, doxing, or sharing of non-consensual private images. While LinkedIn limits the demographic or sensitive data collected by the platform and users tend to self-limit personal and family information shared on the platform, it recognizes that some members may share personal information on LinkedIn that may be exploited by bad actors or that bad actors may share personal information about individuals from other sources to the platform.

The primary way LinkedIn addresses this risk is through its efforts 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, it uses downstream models to 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 also relies on member reports and a team of investigators to uncover accounts that may have evaded all other levels of defense.

LinkedIn also reviews public-facing content for harmful or abusive indicators and flags potential violative content for human review. It scans public and privately facing content for phishing and malware that could offer bad actors access to member data.

And 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. Members also have the ability to access and manage their data or delete it from LinkedIn.

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.

Protection of Personal Data

Risk to protection of personal data includes risk related to collection, processing, or release of user data without consent, including data mining, profile scraping, targeted advertisements without appropriate lawful basis; phishing, hacking, malware, and data breaches. Inherent Severity of this risk was assessed at a slightly lower level considering, while the harm due to insufficient protection of personal data can be significant at the individual or business level, the scale of impact is typically contained within economic and societal systems and does not typically extend to country, regional, or global levels.

Mitigations applied to reducing systemic risk related to protection of personal data are assessed as Managed on the DTSP maturity scale as LinkedIn follows industry best practices with defined, documented, and managed processes and is actively working to enhance protections in place to prevent profile scraping and other evolving threats related to user privacy. In addition to the DTSP Best Practices LinkedIn has effectively implemented across all risk areas, which are further described under Mitigation, LinkedIn has invested in robust policies, policy enforcement mechanisms, and advanced detection capability to effectively mitigate the risk of fraud on the platform, which may result in users sharing personal data with malicious actors or scraping of profile data.

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 offer 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'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. LinkedIn uses a variety of in-product indicators to enhance member awareness and empowerment, including "Why am I seeing this ad?"

Trust is an operating principle 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 considerations.

LinkedIn has robust practices to safeguard and cultivate a culture of respect for user privacy. LinkedIn has a Privacy Champions program and hosts an annual privacy event during Data Privacy Week, requires annual privacy training for all employees, and regularly meets with key regulators to discuss privacy concerns and emerging trends.

Mitigation

DTSP participating companies commit to five fundamental areas of best practices, which are the foundation for trusted and safe products and services. These fundamental areas are product development, governance, enforcement, improvement, and 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 Trust and Safety practices with industry-accepted standards, enabling it to better identify and manage potential risks. By mapping these mitigations to each of the DTSP Best Practice Buckets, LinkedIn can achieve greater transparency and provide reliable and effective solutions to challenges associated with harmful content and conduct.

Product Development

Trustworthy Design Review (TDR) process: LinkedIn's TDR process is the focal point for risk identification and mitigation during the product development lifecycle. At LinkedIn, new products and initiatives undergo a review and assessment by cross-disciplinary specialists, with a focus on legal compliance, data protection, security, and safety. As part of this process, LinkedIn identifies and documents potential risks, abuse vectors, and mitigations, and provides feedback to enhance the product or initiative as well as to improve relevant trust and safety measures. Included among the cross-disciplinary specialists partaking in the TDR process are LinkedIn's product, engineering, legal, information security, and trust and safety teams. Those 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. Including this determination at the outset of the TDR process helps to ensure compliance with data protection commitments, security standards, and applicable laws, as well as alignment with LinkedIn's Trust Principles of putting LinkedIn members first; providing clarity, consistency, and control over data; maintaining a safe, trusted, and professional platform; providing a product with no unfair bias; and respecting the laws that apply to LinkedIn. This process is tracked in a centralized tracking and project management system to ensure accountability and completeness of the review process. LinkedIn regularly reviews and conducts quality checks on the TDR process itself to continually enhance and refine its Trust and Safety reviews.

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 also has several proactive defenses in place 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 ensure accountability and completeness of the review process.

Feature Evaluation: Specific features are assessed holistically as part of the TDR process before product or initiative 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. Additionally, LinkedIn conducts Fairness Checks on changes related to LinkedIn Feed, conducts AB testing, and holds out groups of members to assess impact of algorithmic feature launch.

Ongoing Risk Assessment: LinkedIn's Digital Safety function, which, among other things, serves as LinkedIn's Article 41 Compliance 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.

Pre-Launch Feedback: Pre-launch feedback from cross-functional stakeholders is incorporated as part of the TDR review process.

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.

User Feedback: LinkedIn has a "feedback" button on the site for users to provide feedback. This feedback is incorporated via various automated and manual means into its 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 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 at all stages of the innovation process; foundational research feeds ideation based on member Trust experiences and evaluative usability research on proposed Trust features.

User Controls: 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 provides the opportunity for members to: select whether to view their Feed by personalized or non-personalized priority; hide posts; block, mute, or unfollow members; and report content. Members also have granular control over notifications through their settings. 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, and who can see their contact information. Members can choose to only receive communications from people they know, thus allowing them to control who can contact them. Members also have the ability to access and manage their data or delete it from LinkedIn.

Product Governance

Policies & Standards: LinkedIn's content policies are written 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 & Safety, Product, and Engineering teams to operationalize those policies with granular procedures. LinkedIn's terms of service and policies are updated periodically, as needed. Its Professional Community Policies apply to all members. 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. LinkedIn monitors edge cases for sensitive policy escalations to document decisions and recommend policy adjustments as needed. 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.

User focused product management: In its Feed, LinkedIn strives to broadly distribute only high-quality content - 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 visibility of content 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.

Community Guidelines/Rules: LinkedIn's 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 and nudity, 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 policy, violent and graphic content, 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. 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.

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.

Document Interpretation: LinkedIn's Content Policy team 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 back 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 Company Pages, Page admins have the ability to 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 and Safety 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.

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 tools to handle flagged and removed content, investigations, and communication with users. 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. The tools give LinkedIn's reviewers information about the content and the author to conduct an in-depth analysis and practice due diligence as needed to make a decision on whether the content should remain on the LinkedIn platform or be removed due to violating the Professional Community Policies. LinkedIn's content moderation teams rely on 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: With regard to internal training, 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. LinkedIn has an intensive onboarding and training program for enforcement personnel. 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 also maintains a Privacy Champions program, hosts Data Privacy week events, and requires annual privacy training for all employees. LinkedIn also periodically runs educational campaigns in-product on LinkedIn to increase member awareness of key safety features and member control options.

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 & 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, artificial intelligence (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. 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.

User Reporting: The third layer of protection is human-led detection. If members locate 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. Reported content is then enqueued for review and is removed if found to be in violation of LinkedIn's policies. 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 that are taken into consideration when prioritizing reviews are language, region of the originating content, and type of media.

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 we've 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).

Flagging Processes: LinkedIn has an informal Trusted Flagger program and is in the process of developing a more formal Trusted Flagger program in accordance with Article 22 of the DSA.

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 global news organizations 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.

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 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.

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 recently rolled out a more sophisticated metric to measure the impact of risk, which will 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.

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.

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.

Product Transparency

Transparency Reporting: A cross-functional team of stakeholders from across LinkedIn's broader Trust organization (including members of its Legal, Trust & 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). 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).

User Notice: LinkedIn's publicly available Professional Community Policies apply to all content on platform. Material updates to its 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 people to access the terms. LinkedIn generally notifies authors and reporters (as applicable) on the content at issue, the policy violated, the action it has taken, redress info and, in most instances, a link to allow the individual to appeal LinkedIn's decision.

Complaint Intakes: Members can submit general complaints to LinkedIn customer support by creating a support ticket by following the instructions in LinkedIn's Help Center. Members can submit Privacy complaints through LinkedIn's Privacy Hub. Members can also report potentially violative content directly through the platform. Reporting is 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: LinkedIn has recently made 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.

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[3] 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.

LinkedIn utilizes Inclusivity nudges within Job postings. If algorithmic systems detect that a job title may be offensive, LinkedIn will notify the member in-product that what they are posting may violate LinkedIn policies.

LinkedIn requires Ads customers to certify they will not use LinkedIn to discriminate based on gender or other protected characteristics before accessing the limited demographic form fields (some of which are not available at all for talent-related ad campaigns). Recruiters are required to first certify via a checkbox that they will not use LinkedIn to discriminate based on gender or other protected characteristics before they are provided with demographic form fields.

LinkedIn also provides other members' profile information to members, including the date the profile was created and the last time the profile picture was changed, which can help to identify the authenticity of the members they are interacting with.

Conclusion

The objective of the Systemic Risk Assessment is to holistically and objectively assess areas of greater potential risk stemming from use or misuse of the LinkedIn platform. This exercise will enable Very Large Online Platforms to more effectively 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.

Additional Planned Mitigation Efforts

In response to DSA Article 35, LinkedIn is including in Appendix B a list of additional mitigations that it assesses to be reasonable, effective, and proportional to further mitigate associated systemic risks. LinkedIn's mitigations include:

- Fulfilling its commitment to Advance Responsible AI Innovation, including the roll-out of its AI Governance Framework and publication of explainable models.
- Continued investment in internal tooling, classifier refinement, multilingual models, and enhanced automation capabilities.
- Expansion of Jobs automated scanning and risk ranking to include discrimination and ingested jobs.

Year Two Risk Assessment Enhancements

As part of the year one DSA Systemic Risk Assessment, LinkedIn has adjusted mapping and categorization of various internal data sources to enable additional data-driven assessment of probability and mitigation effectiveness. This will also benefit LinkedIn's transparency reporting efforts specific to DSA. Next year's data will serve as a more complete baseline, which will enable LinkedIn to demonstrate progress year over year on efforts to mitigate systemic risks stemming from use or misuse of the LinkedIn platform.

[3] LinkedIn automatically scans all content for CSAM, malware, and phishing.

Appendices

Appendix A: Methodology Expanded

Risk Areas

LinkedIn defined 11 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 8

Risk Area	Risk Definition
Civic Discourse	Risk that content or activities negatively impacting civic discourse, electoral processes, or public security, including through coordinated manipulation and inauthentic activity, occur on LinkedIn.
Discrimination	Risk that discriminatory content or activities that target or exclude individuals, groups, or businesses based on protected grounds or promote discriminatory practices or gender inequality occur on LinkedIn.
Fraud	Risk that fraudulent activities undermining consumer protection, such as scamming, defrauding, spamming, or falsifying information, including the creation of fake accounts, use of bots or other deceptive commercial practices, are perpetrated on LinkedIn.
Negative Impact on Freedom of Expression, including Pluralism	Risk that freedom of expression, information, the press, and pluralism of the media are negatively impacted by the design or functioning of the LinkedIn platform and algorithmic systems, its policies, or enforcement, including through submission of abusive notices and methods for silencing speech and hampering plurality.
Gender-Based Violence	Risk that content or activities encouraging or perpetrating gender-based violence, such as gender-based hate speech, sexual harassment, promotion of human trafficking or criminal sexual acts, or other forms of violence against women and girls, occur on LinkedIn.
Negative Impact on the Right to Human Dignity	Risk that content or activities degrading human dignity, such as threats or promotion of violence, hate speech, extortion, bullying & harassment, or promotion of human trafficking occur on LinkedIn.
Dissemination of "Illegal Content"	Risk that illegal content or activities, such as intellectual property infringement; defamation; Child Sexual Abuse Material; hate speech; terrorist or violent extremist content; unlawful non-consensual sharing of private images; online stalking; or promotion or sale of illegal, dangerous, or counterfeit goods, services, or animals occur on LinkedIn.
Negative Impact on the Right to Private and Family Life	Risk that content or activities negatively impacting private and family life, including malicious sharing of or exploitation of sensitive or private data, doxing, or sharing of non-consensual private images, occur on LinkedIn.
Negative Impact on Protection of Personal Data	Risk that personal data on LinkedIn is collected, processed, or released without user consent (either by LinkedIn or LinkedIn users), including data mining or targeted advertisements without consent; phishing, hacking, malware, or data breaches; or other unauthorized disclosure or insufficient protection of data by LinkedIn.
Risks to Public Health	Risk that content or activities negatively impacting public health and user physical and mental wellbeing, such as the behavioral addictions, promotion of legal but harmful

	substances or practices including self-harm, health misinformation, or other serious negative effects to physical, mental, social, and financial wellbeing, occur on LinkedIn.
Rights of Children	Risk that content or activities negatively impacting the rights of the child, such as Child Sexual Abuse Material or promotion of child abuse; collection of children's data without parental consent; discrimination, cyberbullying, and harassment; exposure to gambling, exploitation, crime, violence, terrorism, or violent extremism; promotion of self-harm; or other harms to children, occur on LinkedIn.

Assessment Process

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

DTSP Safe Assessment step descriptions
Figure 9

Step	Description	Objective
Discover 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
Identify 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.
Assess 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 & safety risks	Understand current state processes, practices, and tools in relation to a common maturity scale
Test 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.
Report 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.

https://dtspartnership.org/wp-content/uploads/2022/07/DTSP_Report_Safe_Assessments.pdf

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 & 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.

- **Internal Consultations**: Following review of the DTSP questionnaire responses, LinkedIn conducted in-depth workshop sessions with the same 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 as well and the feedback from these engagements informed the assessment.
- **Mitigation Summaries**: LinkedIn stakeholders developed brief summaries of LinkedIn's controls and mitigations specific to each risk area.
- **Policies and Publications**: LinkedIn reviewed relevant external policies, public practices, and other publications 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.
- **Transparency report 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**: 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.
- **Internal metrics**: LinkedIn considered internal metrics related to content moderation to inform both assessment of prevalence as well as effectiveness of mitigations

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:
    - Public incident data
    - Transparency report metrics
    - Internal metrics
    - External stakeholder input

Probability Rating Scale
Figure 10

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 (economic, security, political, societal, wellbeing, and environmental), scale (global, regional, country, locality, individual), and gravity (remediability)
- Items that substantiate severity include:
    - External stakeholder input
    - Deep dive workshops
    - Authoritative sources

Severity Rating Scale
Figure 11

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

Inherent Risk Rating Scale
Figure 12

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

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
    - External stakeholder input
    - Deep dive workshops
    - Mitigation Summaries
    - Policies, blogs, releases
    - DTSP questionnaire responses
    - Internal Metrics

DTSP inventory of 35 best practices

Figure 13

https://dtspartnership.org/wp-content/uploads/2022/07/DTSP_Report_Safe_Assessments.pdf

DTSP Maturity Rating Scale
Figure 14

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 & Safety in every aspect. Processes are continuously improved with innovative ideas and technologies.	80%	Optimized

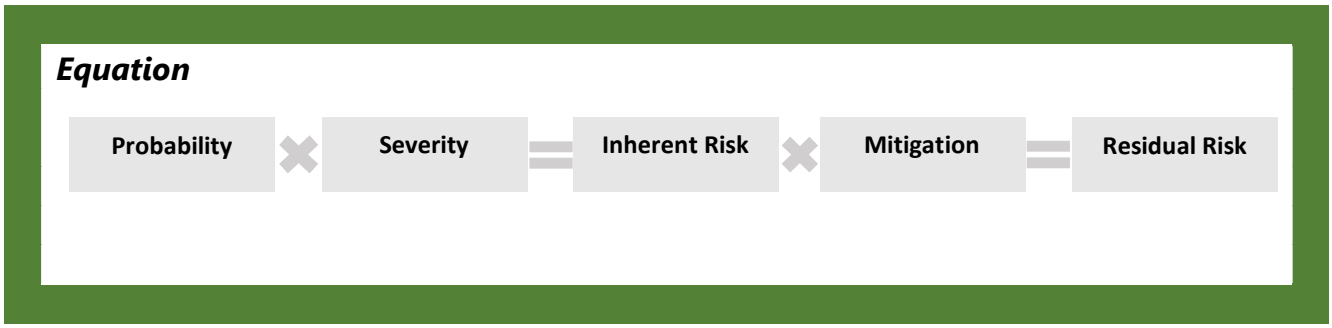
https://dtspartnership.org/wp-content/uploads/2022/07/DTSP_Report_Safe_Assessments.pdf

Residual Risk Rating Scale
Figure 15

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 16



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 rating score of a risk is 12
- If the mitigating rating is "Managed" -> mitigation score of 65%
- Based on the inherent rating score if 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

Mitigation Title	Mitigation Description
AI Governance Framework: Improved Transparency	LinkedIn plans to roll out its AI governance framework to provide additional transparency around privacy and security to the public and publish further on its algorithmic systems to include explainability of models.
Researcher Access to Data: Information and Tools	LinkedIn will continue to invest in collaboration with OpenMined on the "Initiative on Algorithms" to potentially enable privacy and security forward collaboration with researchers. LinkedIn also plans to provide more information on its practices related to researcher access to data to researchers, members, and the public.
Refined Models to Detect Additional Fake Behaviors	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.
Continuous Tuning of Classifiers	LinkedIn plans to continue tuning and moderating its classifiers. LinkedIn regularly reviews and updates its machine learning algorithms to detect and remove violating content. By refining these classifiers to better reflect its policies, LinkedIn can more effectively prevent harmful content from spreading on the platform, providing a safer user experience for LinkedIn members.
Improvements to Jobs Automated Detection	LinkedIn plans to improve and expand current automated detection for policy-violating job posts, including in particular potential discriminatory and fraudulent jobs. LinkedIn regularly updates and improves its automated systems in light of new fraud trends and hiring practices that vary market to market.
Expanded Scanning of Ingested Jobs	LinkedIn plans to improve its systems for automated detection of potentially harmful "ingested jobs," which refers to third-party jobs that LinkedIn pulls onto the platform. Ingested jobs have high quantity but low visibility and are considered lower risk as they are pulled from job sites selected by LinkedIn.
Refinement of TDR Process to map more closely to DSA requirements	LinkedIn plans to refine its robust Trustworthy Design Review process to align even more closely with DSA product design requirements.