Digital Trust
& Safety Partnership

The Safe Assessments

An Inaugural Evaluation of Trust & Safety Best Practices

July 2022

Digital Trust
& Safety Partnership

Table of Contents

Digital Trust
& Safety Partnership

A Message from the Executive Director

The internet enables much of the best in our world, from education and economic opportunities, to staying connected with friends and family amid a global pandemic. Although it is hard to imagine getting through the past few years without digital services, we have also seen the way they can be misused and abused to harm people.

The Digital Trust & Safety Partnership (DTSP) brings together technology companies providing a wide range of products and services around a common approach to increase Trust & Safety across the internet. In less than a year and a half since DTSP launched, we have moved rapidly toward creating and executing industry-wide assessments of how companies are implementing best practices, providing a roadmap to meaningfully increase Trust & Safety online.

Our results show that there are many key functions fundamental to Trust & Safety where industry practices have matured and converged. This is particularly true when it comes to building teams responsible for establishing, updating, and enforcing the rules for the use of digital services.

We have also identified substantial room for improvement across the commitments participating partners make to product development, governance, enforcement, improvement, and transparency. This is particularly the case when it comes to engaging the perspectives of users and external organizations in Trust & Safety, including human rights groups and academic researchers. Although companies have been working to address Trust & Safety for years, these operations are still new compared to other functions, and are evolving in response to new and changing risks. Consequently, we found some functions and practices across our framework occurring on an ad hoc basis.

This report continues the conversation we have begun with a wide range of advocates and policymakers. Through sharing these insights, we hope to promote the collaborative development of industry standards and encourage continuous improvement. We welcome feedback and expect that these interactions will enhance our efforts.

The culmination of much hard work, this report also marks the beginning of a new phase of effort. Looking ahead, we will be engaging the wider world of voices concerned with Trust & Safety as we iterate our framework, institute an objective and measurable third-party assessment process, and more widely and systematically engage interested parties in our work. We look forward to collaborating with you as we take these efforts forward.

David M. Sullivan
Executive Director

Executive Summary

This inaugural report synthesizes the results of the Safe Framework assessments conducted by ten DTSP partners during the first half of 2022. Participating partners for initial assessments were Discord, Google, LinkedIn, Meta Platforms, Inc., Microsoft, Pinterest, Reddit, Shopify, Twitter, and Vimeo. Other DTSP partners, including those that have joined the organization more recently, will participate in a future cycle of assessments.

DTSP launched in February 2021, bringing together companies of different sizes and business models, to develop industry best practices and verify their implementation through independent third-party assessments, to ensure consumer Trust & Safety when using digital services. Our goal is a safe online experience that continues to enable worldwide users of digital services to benefit from the social, economic, and political value they provide.

Our participating companies have committed to five fundamental areas of best practices ("the DTSP Commitments") to which a digital service must adhere to promote a safer and more trustworthy internet. These five commitments are the foundation for trusted and safe products and services: product development, governance, enforcement, improvement, and transparency. They are underpinned by 35 specific best practices, known as the DTSP Best Practices Framework, which provide concrete examples of the variety of activities and processes that organizations may have in place to mitigate risks from harmful content and conduct.

DTSP Inventory of 35 Best Practices

Product Development

- PD1: Abuse Pattern Analysis
- PD2: Trust & Safety Consultation
- PD3: Accountability
- PD4: Feature Evaluation
- PD5: Risk Assessment
- PD6: Pre-Launch Feedback
- PD7: Post-Launch Evaluation
- PD8: User Feedback
- PD9: User Controls

Product Governance

- PG1: Policies & Standards
- PG2: User Focused Product Management
- PG3: Community Guidelines/Rules
- PG4: User Input
- PG5: External Consultation
- PG6: Document Interpretation
- PG7: Community Self Regulation

Product Enforcement

- PE1.1: Roles & Teams
- PE1.2: Operational Infrastructure
- PE1.3: Tooling
- PE2: Training & Awareness
- PE3: Wellness & Resilience
- PE4: Advanced Detection
- PE5: User Reporting
- PE6.1: Enforcement Prioritization
- PE6.2: Appeals
- PE6.3: External Reporting
- PE7: Flagging Processes
- PE8: Third Parties
- PE9: Industry Partners

Product Improvement

- PI1: Effectiveness Testing
- PI2: Process Alignment
- PI3: Resource Allocation
- PI4: External Collaboration
- PI5: Remedy Mechanisms

Product Transparency

- PT1: Transparency Reports
- PT2: Notice to Users
- PT3: Complaint Intakes
- PT4: Researcher & Academic Support
- PT5: In-Product Indicators

A summary of the 35 best practices categorized by Commitment

Participating partners conducted internal assessments using the DTSP assessment methodology, the Safe Framework. The assessments examined the people, processes, and technology that contribute to managing content- and conduct-related risks for participating companies. Companies assessed their practices against a common maturity scale ranging five levels, from ad hoc to optimized:

DTSP MATURITY RATING SCALE

(1) Ad Hoc	(2) Repeatable	(3) Defined	(4) Managed	(5) Optimized
A rating of Ad Hoc is assigned when execution of best practices is incomplete, informal, or inconsistent.	A rating of Repeatable is assigned when execution of best practices occurs without standardized processes. Organizations aim to document more formalized practices.	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.	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.	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.

Safe Framework Assessments in a Nutshell

DTSP has **defined goals** (the DTSP Commitments), **identified actions** to achieve them (the DTSP Best Practices Framework), and established a **means of measurement** (the DTSP Maturity Rating Scale). We have aimed to be as concrete as possible, while taking a proportionate, risk-based approach where companies focus on addressing the risks particular to their products and practices.

For example, under the Product Development commitment, there are nine listed best practices, one of which is risk assessment.[1] A company evaluating the strength of its risk assessment practice would rank itself on the Maturity Scale anywhere from "ad hoc" to "optimized."

[1] Specifically, "Use in-house or third-party teams to conduct risk assessments to better understand potential Risks." See the DTSP Best Practices Framework.

This report reflects our first use of the Safe Framework in practice. We are encouraged by the results, and know that our learnings from this initial effort will help us to iterate and make our framework more effective over time. For this inaugural round, each company chose particular aspects of their operations to examine. Some focused on individual products, for example, or focused on specific kinds of prohibited content. Some assessed all five DTSP Commitments, while others focused on a subset of key commitments and best practices.

As an inaugural application of the DTSP assessment framework, the results are encouraging. Participating companies were able to substantively assess where they stand and develop future opportunities for improvement. The results are encouraging, not only in terms of the insights shared in this report, but because they show that companies of different sizes and business models have been able to implement a common assessment framework, which will evolve as our work moves forward.

Because of the differences in approach to assessment that were deliberately built into the Safe Framework, we cannot offer sweeping conclusions about all industry practices. However, we do offer some key insights and trends that can inform future action by companies, governments, and civil society.

Key Insights

Successes: many companies reported a mature state of development for core content moderation practices

Eight best practices were assessed at an overall maturity level of Managed:

Maturity Rating: (4) Managed	Commitment	Managed Practices
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.	Product Governance	Establish a team or function that develops, maintains, and updates the company's corpus of content, conduct, and/or acceptable use policies
	Product Governance	Develop user-facing policy descriptions and explanations in easy-to-understand language
	Product Enforcement	Constitute roles and/or teams within the company accountable for policy creation, evaluation, implementation, and operations

Continued on next page ⟶

Maturity Rating: (4) Managed	Commitment	Managed Practices
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.	Product Enforcement	Develop and review operational infrastructure facilitating the sorting of reports of violations and escalation paths for more complex issues
	Product Enforcement	Invest in wellness and resilience of teams dealing with sensitive materials, such as tools and processes to reduce exposure, employee training, rotations on/off content review, and benefits like counseling
	Product Enforcement	Implement method(s) by which content, conduct, or a user account can be easily reported as potentially violating policy (such as in-product reporting flow, easily findable forms, or designated email address)
	Product Enforcement	Ensure relevant processes exist that enable users or others to "flag" or report content, conduct, or a user account as potentially violating policy, and enforcement options on that basis
	Product Transparency	Provide notice to users whose content or conduct is at issue in an enforcement action (with relevant exceptions, such as legal prohibition or prevention of further harm)

These practices show that Trust & Safety teams and functions across the partners have performed relatively well when it comes to core practices and activities that fall squarely within their domain and can be implemented unilaterally, to some degree. These include constituting the teams responsible for content policies and developing public facing policy descriptions, as well as developing enforcement infrastructures that span people, processes, and technology, and notifying users whose content is subject to an enforcement action for violating policies.

Areas for improvement: many of the least mature practices relate to user feedback and external collaboration

Seven best practices were assessed at an overall level of maturity of Repeatable:

Maturity Rating: (2) Repeatable	Commitment	Repeatable Practices
A rating of Repeatable is assigned when execution of best practices occurs without standardized processes. Organizations aim to document more formalized practices.	Product Development	Provide for post-launch evaluation by the team accountable for managing risks and those responsible for managing the product or in response to specific incidents
	Product Governance	Institute processes for taking user considerations into account when drafting and updating relevant Product Governance
	Product Governance	Create mechanisms to incorporate user community input and user research into policy rules
	Product Governance	Facilitate self-regulation by the user or community to occur where appropriate, for example by providing forums for community-led governance or tools for community moderation and find opportunities to educate users on policies, for example, when they violate the rules
	Product Improvement	Develop assessment methods to evaluate policies and operations for accuracy, changing user practices, emerging harms, effectiveness and process improvement
	Product Improvement	Use risk assessments to determine allocation of resources for emerging content- and conduct-related risks
	Product Transparency	Create processes for supporting academic and other researchers working on relevant subject matter (to the extent permitted by relevant law and consistent with relevant security and privacy standards, as well as business considerations, such as trade secrets)

Three of the practices deemed least mature, according to the self-assessments, related to incorporating user and third-party perspectives into Trust & Safety policy and practices. This illustrates the internal focus of Trust & Safety functions. As a discipline, Trust & Safety has developed with less external engagement outside of companies until recently.

The least mature of all assessed practices is the creation of processes to support academic and other researchers working on relevant subject matter. This is an area of great interest, with pending regulations in Europe and proposed legislation in the United States.[2] While several assessments indicated planned improvements to mature this practice in the coming year, improvements to legal frameworks to address concerns around security and privacy, among other concerns, may be worthy of consideration.

Areas of ongoing improvement: integrating Trust & Safety into product development

The majority of assessments indicated companies were in the process of formalizing the relationship between Trust & Safety and product teams to better integrate these perspectives into product development. By continuing to adopt and enhance practices in this commitment, including the "Safety by Design" approach, companies anticipate improvement across the Product Development commitment in the future, with some companies planning concrete improvements in the coming year.

Where We Go From Here

This report marks the beginning of our collective effort to evolve and evaluate Trust & Safety practices across the industry. Partner companies are using the results of their self-assessments to enhance their practices, and DTSP will facilitate learning to identify collective opportunities to mature key practices.

Looking ahead, we foresee the following lines of effort:

- **Evolving the DTSP Best Practices Framework:** we will use the insights generated through this process and external input to review and improve the DTSP Best Practices Framework;
- **Moving to third-party assessment:** we are working with experts with deep understanding of both Trust & Safety and assessment frameworks to articulate our approach to independent third-party assessment and will share more information on this in the coming months; and
- **Engaging stakeholders globally:** as we raise awareness of Trust & Safety best practices, we will put in place specific mechanisms for stakeholder input and engagement and provide opportunities for dialogue.

[2] For example, see the EU Digital Services Act and the Platform Accountability and Transparency Act, announced by a bipartisan group of U.S. Senators.

1. Introduction

About the Digital Trust & Safety Partnership

The Digital Trust & Safety Partnership (DTSP) is focused on promoting a safer and more trustworthy internet.

We bring together participating companies to monitor and assess their people, processes, and technology against the five DTSP Commitments, as they identify and mitigate content- and conduct-related risks for their products and services. DTSP's current membership includes Apple, Bitly, Discord, Google, LinkedIn, Meta Platforms, Inc., Microsoft, Patreon, Pinterest, Reddit, Shopify, Twitter, Vimeo, and Zoom.

Although technology companies have been working to address Trust & Safety for years, these operations are relatively new compared to other company functions, and face rapidly changing risks. Until now, the field of Trust & Safety has not yet developed the kinds of best practices and model assessments that have been crucial to maturing and organizing other tech disciplines like cybersecurity.

Key Terms

Trust & Safety refers to the field and practices that manage challenges related to content- and conduct-related risk, including but not limited to consideration of safety-by-design, product governance, risk assessment, detection, response, quality assurance, and transparency.

Content- and conduct-related risk(s) refers to the possibility of certain illegal, dangerous, or otherwise harmful content or behavior, including risks to human rights, which are prohibited by relevant policies and terms of service.

The DTSP Best Practices Framework

All participating companies in the DTSP have agreed to promote a safer and more trustworthy internet and have committed to best practices in five areas:

- **Commitment 1: Product Development**
  Identify, evaluate, and adjust for content- and conduct-related risks in product development.

- **Commitment 2: Product Governance**
  Adopt explainable processes for product governance, including which team is responsible for creating rules, and how rules are evolved.

- **Commitment 3: Product Enforcement**
  Conduct enforcement operations to implement product governance.

- **Commitment 4: Product Improvement**
  Assess and improve processes associated with content- and conduct- related risks.

- **Commitment 5: Product Transparency**
  Ensure that relevant trust & safety policies are published to the public, and report periodically to the public and other stakeholders regarding actions taken.

Across the Commitments, 35 best practices have been identified, also known as the DTSP Best Practices Framework, that are non-exhaustive examples of the kinds of activities and processes that a company could have in place to mitigate risk and ensure the safety of the service.

DTSP Inventory of 35 Best Practices

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

A summary of the 35 best practices categorized by Commitment

All DTSP partners embrace the commitments. Not every practice, however, is suitable for every product or service.[3] Each company is responsible for implementing a combination of the best practices that is most appropriate to their individual products or services to mitigate content- and conduct-related risks and ensure adherence to these commitments.

[3] For example, a service that does not have a community function, such as file storage, may not need to use the "community self regulation" practice.

The Purpose of the DTSP Best Practices Framework

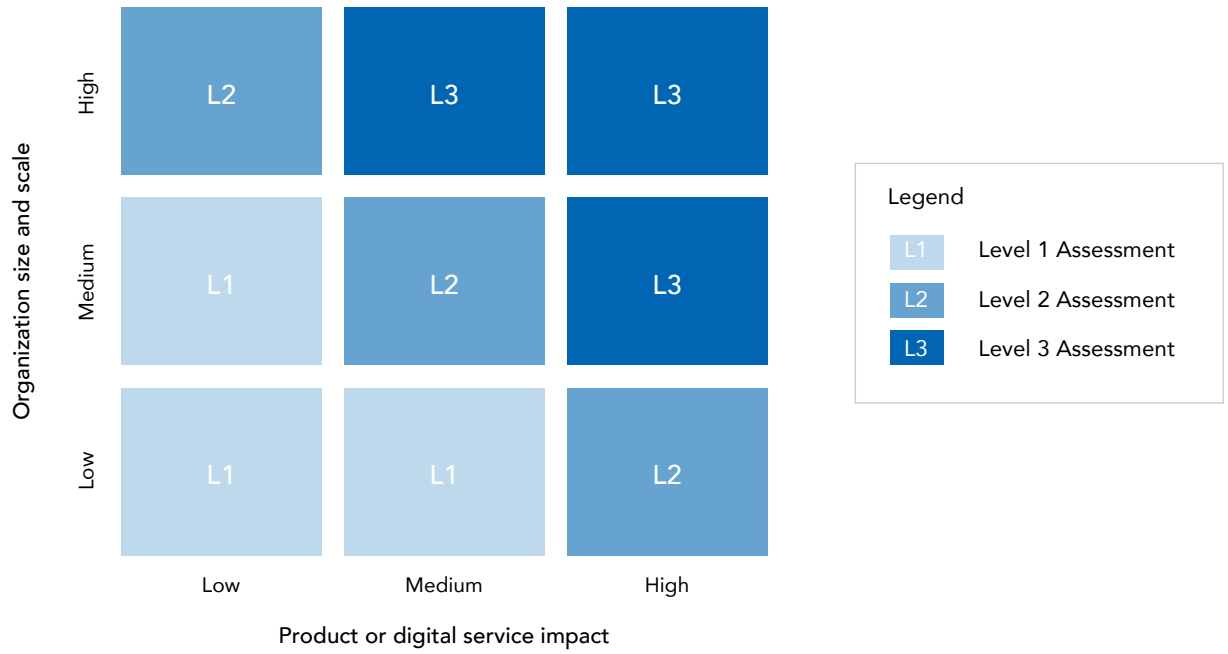
The DTSP Best Practices Framework focuses on considerations in the development, governance, enforcement, improvement, and clear documentation of digital products and services.

Each organization in the DTSP is guided by its own values, product aims, and experiences with user behavior. Each brings digital tools and blended machine and human processes to make decisions about how to enable a broad range of human expression and activity, while working to mitigate as much risk as possible by identifying and preventing harmful content or conduct. Despite the individual approaches, DTSP members agree on the need for a shared framework of best practices to help raise the bar on Trust & Safety operations across industry and create meaningful and robust standards for assessment.

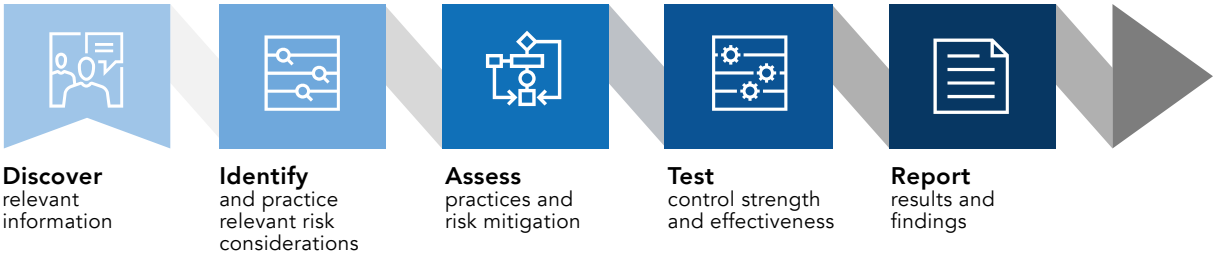
The Safe Framework

The Safe Framework examines the people, processes, and technology that contribute to managing content- and conduct-related risks for member companies. DTSP uses three levels of assessment that a company may undertake to examine Trust & Safety practices in support of a particular product, digital service, or function. The Level 3 assessment is designed as the most in-depth in terms of the breadth and depth of assessment procedures, while Level 1 is less detailed and provides for more summary-level analysis, with Level 2 falling in the middle.

Using a risk-based approach, the depth of assessment is then determined by evaluating the size and scale of the organization, as well as the potential impact of its product or service.



The assessment is designed to help organizations understand how DTSP practices will help them manage content- and conduct-related risks, using a five step methodology:



A full explanation of the Safe Framework Methodology can be found here.

ABOUT THIS REPORT

This report is a first attempt to provide industry-level insight into the Trust & Safety practices at leading technology companies. It is an assessment of how companies mitigate and manage risks related to online content and conduct. We intend for this report to serve as a baseline for future assessments, and therefore report on general trends and themes and not on specific conclusions; we want this report to be a guide for industry, government, and other key stakeholders rather than an inclusive set of considerations.

Information in this report is the direct output of assessments conducted by DTSP partner companies using the Safe Framework. The following DTSP partners contributed information reflected in this report: Discord, Google, LinkedIn, Meta Platforms, Inc., Microsoft, Pinterest, Reddit, Shopify, Twitter, and Vimeo.[4]

Report Scope and Methodology

The Safe Framework embraces the variety of methods taken by DTSP companies to fulfill their commitments. Rather than taking a narrow approach that could be overly rigid, we intend to use the learnings derived from this flexible approach to inform our future efforts.

Using the Safe Framework methodology, the companies took individual approaches to scoping their assessments and evaluating their Trust & Safety practices. Some companies assessed one or more products, while others assessed their Trust & Safety function, or a particular component of that function. Some companies assessed all commitments and best practices, whereas others focused on particular areas, either because they were most relevant to particular risks, or to focus on recent work to define and iterate particular practices.

To prepare this report, DTSP compiled insights from Assessment Results Surveys (see Appendix IV) completed for each assessment, and complemented those results with insights from DTSP member discussions held under the Chatham House rule.

[4] Other DTSP partners, including those that have joined the organization more recently, will participate in a future cycle of assessments.

This methodology presents certain challenges and limitations including, but not limited to:

- Comparability constraints from the assessment of different products and functions;
- Consistency challenges that arise from self-assessments and different approaches to implementation. Although all companies used the Safe Framework materials, each company developed its own documentation to implement the assessment; and
- Applicability limitations related to aggregating results, as not all insights included in this report apply to all participating companies.

We encourage readers to be mindful of these limitations and to approach this report as a first attempt to develop industry-wide analysis of the state of Trust & Safety practices, recognizing this snapshot will enable learning and inform the evolution of future assessments.

The DTSP Maturity Scale

For the inaugural assessments, DTSP partners agreed to a common maturity scale against which the DTSP Best Practices Framework would be assessed. Maturity models are commonly used in other disciplines, from software development to privacy, security, and corporate responsibility.

DTSP MATURITY RATING SCALE

(1) Ad Hoc	(2) Repeatable	(3) Defined	(4) Managed	(5) Optimized
A rating of Ad Hoc is assigned when execution of best practices is incomplete, informal, or inconsistent.	A rating of Repeatable is assigned when execution of best practices occurs without standardized processes. Organizations aim to document more formalized practices.	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.	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.	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.

Maturity in Context

Safe Framework assessments are deliberately designed to accommodate assessing practices across a wide variety of services, each with unique risks and which may be at very different maturity levels. Objectives may vary by product and by practice, depending on risk tolerance. The ultimate objective is not for all practices at all partner companies to be Optimized, and for some products, a practice that is assessed to be Managed may suffice.

2. Key Insights

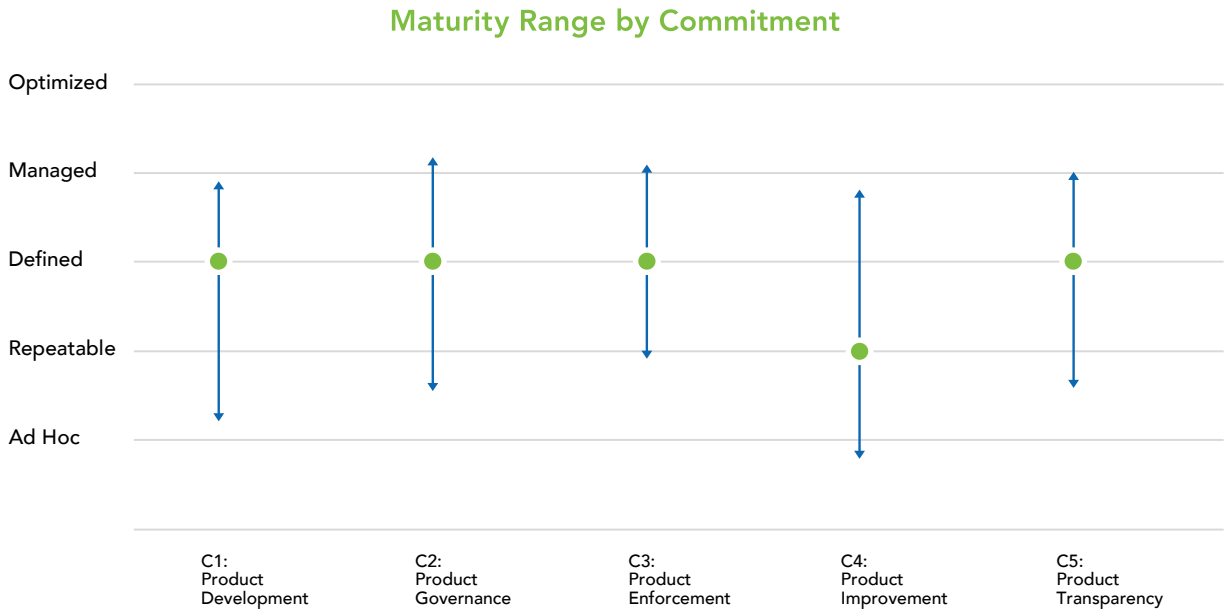
This section provides details about key insights, themes, and observations from the results of assessments completed by ten DTSP partner companies.

Partner companies are making significant use of the DTSP Best Practices Framework

DTSP conducted a baseline survey of partners in November 2021 which found that the companies participating in this first round of internal assessments were using at least 80 percent of the best practices. This indicates that the industry is generally aligned with the DTSP Best Practices Framework, and that it is well positioned to further mature within each of the five DTSP Commitments. It is important to note that companies may focus on assessing particular commitments and practices based on the risks they identify during the assessment. For this reason, not all assessments address all of the practices used by each company.

Best practices are more proactive than reactive, but with wide maturity range

Participating partners identified an overall level of maturity of "defined" across the assessments.[5] This aggregate maturity level belies both achievements and room for improvement, shown by the maturity range across the commitments:



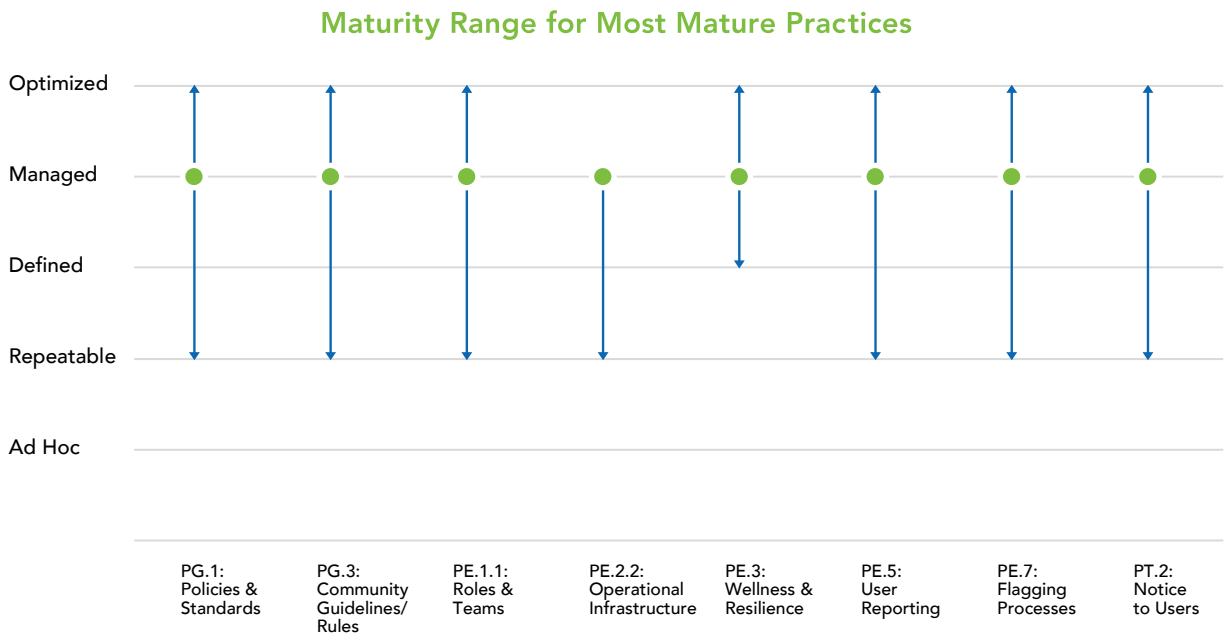
[5] A rating of Defined is assigned when execution of best practices occurs with defined and documented processes. Processes are more proactive than reactive, and implemented across the organization.

The Product Improvement commitment has the widest maturity range, followed by Product Development and Product Governance. These are the areas where there is the greatest divergence in self-identified levels of maturity, reflecting opportunities for improvement and shared learning within industry to understand different approaches.

Notably Product Governance, Product Enforcement, and Product Transparency are the only commitments where one or more companies assessed any practices to be at an Optimized maturity level.

More mature best practices show industry converging around core functions

Eight best practices were assessed at an overall maturity level of Managed.[6]



These practices fall under the Product Governance, Product Enforcement, and Product Transparency Commitments. They show that Trust & Safety across the partners has prioritized:

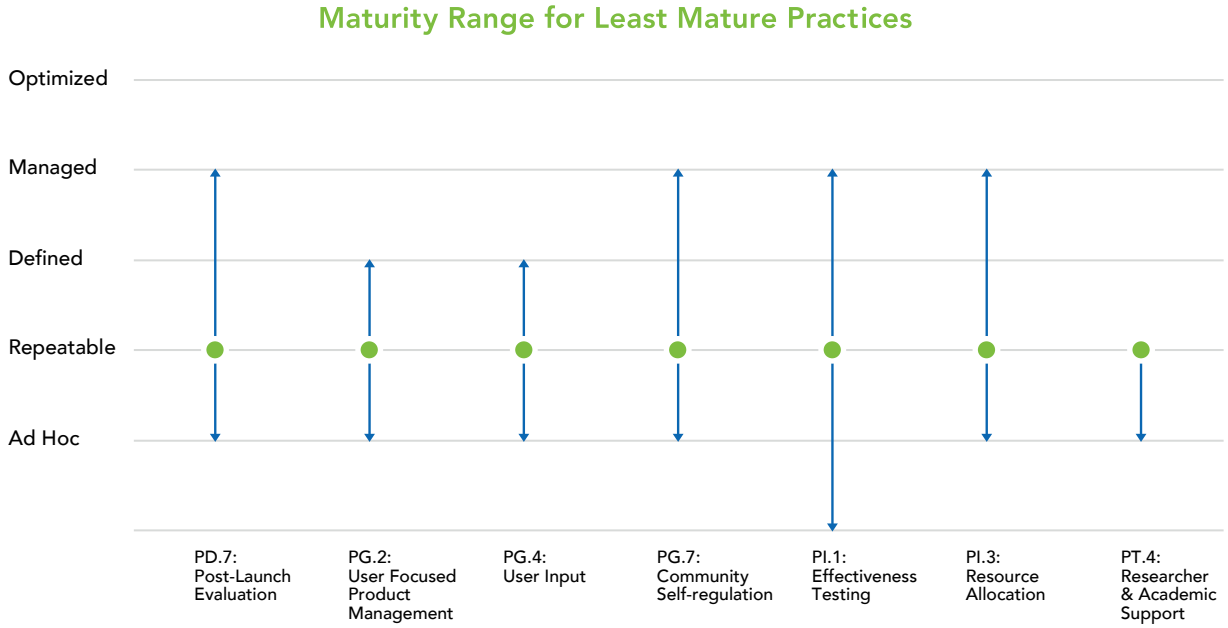
- Constituting teams responsible for content policies;
- Making those policies available and explainable;
- Developing enforcement infrastructures that span people, processes, and technology, and
- Notifying users whose content is at issue.

[6] 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.

The relatively narrow maturity range for investing in the wellness and resilience of teams involved in content review shows that while all companies are making significant efforts on this practice, almost all companies recognize there is more work to be done.

The least mature practices relate to user feedback and external collaboration

Seven best practices were assessed at an overall level of Repeatable.[7]



Three of the practices that were least mature according to the self-assessments related to incorporating user and third-party perspectives into Trust & Safety policy and practices. This reflects the state of maturity of Trust & Safety as a discipline that has developed with less external engagement outside of organizations until recently.

Other less mature practices, such as post-launch evaluation of products from a Trust & Safety perspective, and using risk assessments to allocate resources toward specific risks, suggest the industry is still in the process of creating formal structures to integrate approaches to managing content- and conduct-related risks across the product lifecycle.

Assessments indicated that the least mature practice is creating processes to support academic and other researchers working on relevant subject matter. This is a topic of great interest covered by the EU's Digital Services Act and in legislation proposed in the United States. The lack of maturity found in our assessment

[7] A rating of Repeatable is assigned when execution of best practices occurs without standardized processes. Organizations aim to document more formalized practices.

attests to the challenges instituting such partnerships under current legal frameworks where companies face significant impediments to sharing data due to security and privacy concerns as well as other business considerations. Although there is an opportunity for thoughtful legislation to improve the enabling environment for this practice, it should be informed by industry expertise to help navigate legal complexity and other sensitivities around private user data and complex data sets. The European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access is one example of an ongoing multi-stakeholder effort to bridge this gap.

Practices related to incorporating user feedback into Trust & Safety also lagged behind other practices. This suggests an opportunity for companies to evolve formal channels or processes for input.

External collaboration is also in need of improvement, whether it is engaging with civil society groups and experts for input on policies, or working with fact checkers and human rights groups on meaningful enforcement responses.

Product Development is where the most improvement is expected

The majority of assessments indicated companies were in the process of formalizing the relationship between Trust & Safety and product teams to better integrate these perspectives into product development, anticipating improvement across the Product Development commitment within the coming year.

Integrating Trust & Safety into Product Development is complex, with multiple approaches to ensuring product adherence to Trust & Safety principles. This can include interoperable integration across product lines through automated harms detection and enforcement systems, customer reporting systems, and other tooling provided by third-party vendors or custom built by and for the organization. It can also entail custom Trust & Safety features specific to the product based on its unique risks factors. Examples include in-product contextual prompts, user controls and settings, and adjustments to algorithmic choices, among others. Product specific solutions may provide unique opportunities to mitigate particular content- and conduct-related risks, but may require substantially more time and resources than less customized approaches.

Companies are seeking to improve their practices across varied cultural and linguistic settings

Companies use a combination of techniques and methods to undertake the best practices in varied cultural and linguistic settings, but more improvement is needed throughout the practices in this area.

The assessments indicated that companies currently focus on:

- Providing their policies in the most commonly used languages used on their platforms and keeping them updated;
- Continuously improving product and operational evaluations that enable scalable support across languages and locations;
- Using external experts that are from a variety of cultural and professional backgrounds that reflect the diversity of their community;

- Conducting geo-specific interviews with the users and asking for users' feedback in different languages about policy development;
- Training and providing support to local reviewers;
- Working with local experts especially in conflict zones; and
- Using language and regional priorities to inform content review.

The centralization of Trust & Safety functions brings notable tradeoffs

Each participating company takes a unique approach to structuring their Trust & Safety function, but broadly these structures fall into three categories, with distinct advantages and disadvantages to each:

- Companies with a centralized Trust & Safety function tend to benefit from more consistent application of standards across one or more products. These teams may operate at a distance from product teams which can be a barrier to collaboration, but may also have dedicated resources, particularly engineering capabilities specific to Trust & Safety, lacking elsewhere.

- Hybrid models that use a "hub and spoke" approach, where functions such as risk assessment, mitigation planning and scoping, and mitigation technical architecture are undertaken centrally at the hub, and where integration of Trust & Safety functionality is the responsibility of "spoke" product teams with support from the "hub." For some companies, this helps enable centralized consistent application of Trust & Safety policies at the product level, while also balancing workloads and reinforcing the importance of a culture of Trust & Safety throughout the organization.

- Decentralized approaches entail some Trust & Safety functions at the corporate level, but most responsibility for Content and Conduct-Related Risks occurs within product teams. While this can encourage ownership of Trust & Safety concerns by product teams, it also presents challenges for consistency, and for ensuring that all products have sufficient resources to mature their practices.

Regardless of which approach each participating company employs, the assessments revealed the multi-disciplinary nature of Trust & Safety. Different teams and functions, including engineering and marketing, bring unique perspectives and may use different terms to describe common challenges. In some cases the assessments also generated insights for other functions, such as privacy or security.

The diversity of approaches to assessment provide learning opportunities

Companies took diverse approaches to implementing their Safe Framework assessments. While several assessments examined a flagship product or a centralized Trust & Safety function, others focused more narrowly on specific aspects of Trust & Safety. As an industry initiative, it is not the role of DTSP to compare company performance against one another, but to develop a common framework against which company progress can be measured and provide opportunities for shared learning. DTSP views this diversity of approaches as a strength during this first cycle of assessments, which will enable a greater degree of learning to inform future industry efforts. DTSP will convene our members with other stakeholders to develop insights and resources to support this work.

3. Insights By Commitment

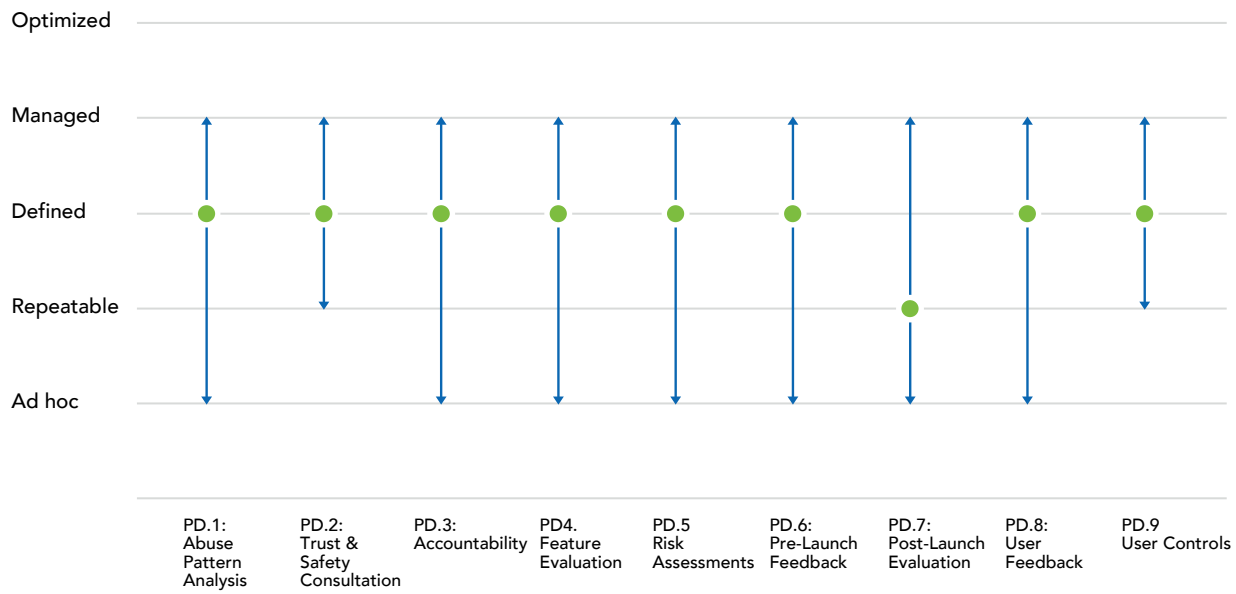
This section analyzes the maturity of assessed practices at the level of each commitment. We provide the range of maturity for each practice, followed by insights regarding "focus practices" based on responses to specific questions in the Assessment Results Survey completed by all participating companies.[8]

C1 Product Development: Identify, evaluate, and adjust for content- and conduct-related risks in product development.

Minimum Maturity	Overall Maturity	Maximum Maturity
Ad Hoc	**Defined**	Managed

Anticipating and reducing risk during product development is key to preventing the potential misuse of digital products and services. Trust & Safety is an emerging discipline, however, that is not supported by international standards. Efforts to develop threat models and technologies to prevent or mitigate abuse are still maturing. This first Commitment appears among the less mature on aggregate, with most assessments identifying some ad hoc or repeatable practices. However, Product Development also presents one of the greatest areas for improvement, with many assessments describing plans for enhancements to these practices that will increase maturity in the next year.

Product Development Maturity Range



[8] These focus practices are neither more nor less important than other practices in the DTSP Best Practices Framework; in future rounds of assessments DTSP may pose questions that will explore implementation of other practices.

Focus Practices for Product Development

Develop insight and analysis capabilities to understand patterns of abuse and identify preventive mitigations that can be integrated into products

Less mature practices had capabilities for certain content categories, but lacked visibility into the full range of abuse patterns on the service. Defined practices included the use of dashboards and alerts, complemented by using third-party vendors to proactively search for abuse trends to be referred to in-house teams for investigation.

Companies with more mature capabilities pointed to review processes that combined human review and automated means to identify patterns of abuse, with automation identifying examples at scale and enabling manual deep dives on content and account samples to gather context and root causes behind steady or emerging patterns of abuse. These exercises are in some cases conducted monthly or quarterly on a fixed cadence. They are also combined with more frequent monitoring of proactive defenses employed to detect fraudulent behavior and policy-violating content, to address any gaps or challenges in how these systems operate.

Include Trust & Safety team or equivalent stakeholder in the product development process at an early stage, including through communication and meetings, soliciting and incorporating feedback as appropriate

Multiple assessments indicated efforts are underway to formalize processes to involve Trust & Safety stakeholders in product development pre- and post-launch.

The structure of Trust & Safety teams led to different levels of maturity for this practice. Some less mature practices identified in the assessments showed that while product teams might engage with corporate-level Trust & Safety teams on discrete issues, they were not engaging more broadly, especially on the implications of features and functionality for content- and conduct-related risks.

More mature practices across this commitment entailed not just having a means for Trust & Safety teams to provide input during product development, but defined policies and processes that ensure such input is incorporated prior to product launch. This can occur via cross-functional reviews in which diverse teams from product, policy, and operations consider risks and mitigations.

Documenting specific factors and questions that Trust & Safety teams can pose during cross-functional product development meetings may help systematize these functions. Examples of questions used by one or more companies include the following:

- Does the product facilitate new user generated content, or make use of existing UGC?
- Are you introducing a new type of media?
- Is your feature adding a new attack surface?

- Are you introducing new collection of data?
- How can a user resolve issues they encounter with this feature? Can they control what they see and with whom they interact?
- In what ways could your feature be abused?
- How can you monitor, detect, mitigate, and remediate the abuse?
- Does the feature include mitigation and remediation considerations?
- Does the feature provide an opportunity for appeal, so that mistakes can be corrected if warranted?

These examples of questions also demonstrate the multidisciplinary nature of Trust & Safety, which touches upon other issues including privacy and security. Some assessments noted that the experience of conducting a Safe Framework assessment valuably surfaced such issues and fostered positive collaboration between teams responsible for these issues and those working on Trust & Safety.

Adopt appropriate technical measures that help users to control their own product experience where appropriate (such as blocking or muting)

Some assessments reported high levels of maturity for technical measures that help users control their own product experience. They demonstrated a wide range of controls that reflect the diversity of ways that users interact via digital services:

- Reporting content or users for violations;
- Blocking users;
- Muting users or content;
- Unfollowing accounts or content;
- Hiding unwanted content;
- Express interest/disinterest in content through "show more/less" features;
- Toggle on/off safe browsing;
- Toggle on/off adult content;
- Granular controls for who can send private message requests; and
- Granular controls for notifications.

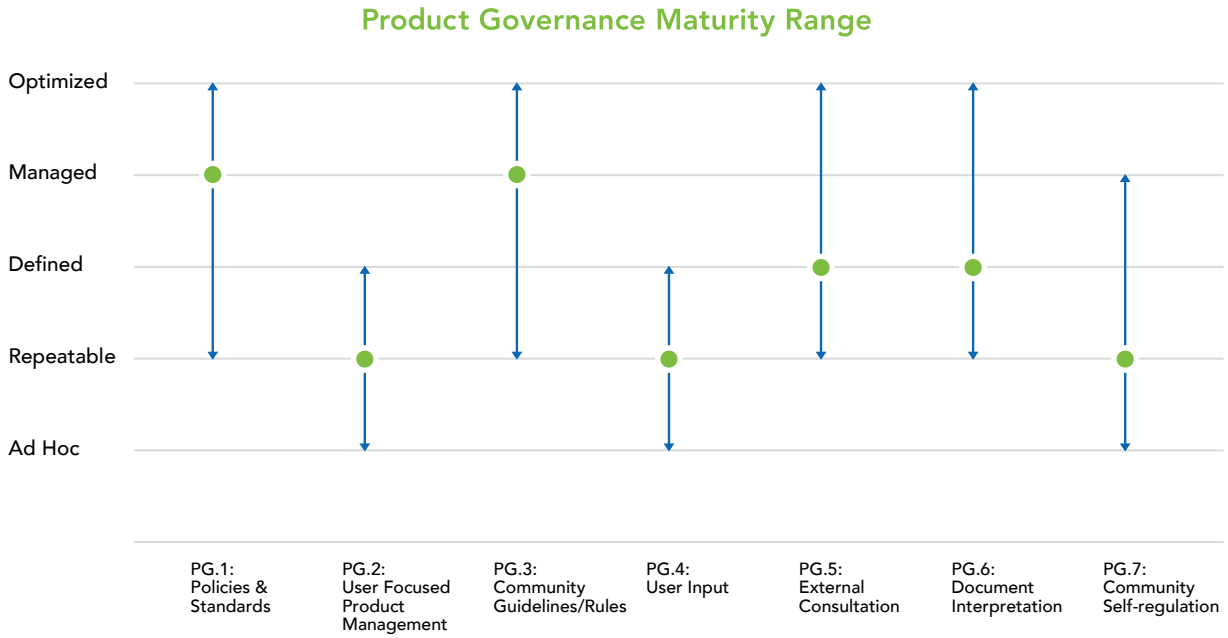
More mature assessments described the use of adoption metrics, which are generated and assessed, to help ensure newly applied controls are being successfully applied across the product or service.

Digital services that are not social in nature, such as file storage and sharing systems, may not have technical features related to blocking and muting, aside from the ability for a file owner to control access by adding or removing others from having access to stored files. As product features evolve over time, it is important for companies to consider whether additional control measures may be appropriate.

C2 Product Governance: Adopt explainable processes for product governance, including which team is responsible for creating rules, and how rules are evolved.

Minimum Maturity	Overall Maturity	Maximum Maturity
Ad Hoc	**Defined**	Managed

The Product Governance commitment includes practices with the most uneven levels of maturity found by the assessments, including two of the most mature practices in the framework and three of the least mature. Practices related to establishing a team responsible for governance and developing user-facing policy descriptions and explanations were among the most mature practices across the entire framework, while practices related to incorporating user feedback into Product Governance lagged behind.



Uneven performance in Product Governance reflects the extent to which Trust & Safety has been a one-way function in many companies, where rules are developed internally and conveyed to users. The need to create two-way processes that use multiple means to incorporate user perspectives into product governance decision making is more recent. Although several assessments described using a combination of internal processes and signals to gather and analyze user feedback to inform policy development and updates, such processes were more ad hoc than other Product Governance practices and need to be formalized. In other cases, assessments acknowledged that updates to corporate terms of service are more often driven by new features and functionality, or by legal need, rather than by user input.

At the same time, the wide range of maturity for facilitating self-regulation by the user or community reflects the diversity of approaches in this space. Products and services where community-based moderation is core report high degrees of maturity relative to others. In some cases, these companies have dedicated "community teams" responsible for engaging users, which are used to preview rule changes and gather direct feedback in writing or in real-time via calls. These teams complement technology and processes that enable self-governance, such as moderator guidelines, configurable automated moderation bots and other community tools and resources. In other cases, products and services that do not focus on social or community user bases reported this practice was not applicable to their service.

Focus Practices for Product Governance

Establish a team or function that develops, maintains, and updates the company's corpus of content, conduct, and/or acceptable use policies

One of the most mature practices across all assessments is among the most fundamental for providers of digital services that involve user generated content: designating a team to develop, maintain, and update the rules that govern its use. The maturity of this practice reflected several companies implementing formal frameworks for content governance, including innovative practices for the creation and evolution of content policies. These include policy lifecycles, maturity measurement frameworks, and evaluation of policies on a fixed cadence with a standardized policy audit review process.

Mature aspects of this practice also included major concerted efforts with internal and external stakeholders to announce policy changes and significant efforts to make sure users are informed.

Create mechanisms to incorporate user community input and user research into policy rules

While many assessments described methods to incorporate user research into policy rules, this usually took place via methods where users would not be aware of these channels. For example, if large amounts of a certain type of content is reported by users for violating policy, this could be brought to the attention of teams responsible for content policy updates by human reviewers. Other assessments reported user surveys as a key tool for incorporating user perspectives into content policies.

Planned improvements expected to result in more mature practices across Product Governance include collaboration between Trust & Safety and public policy teams within companies to create user and third-party input channels for policy input and feedback. In some cases this will include subjecting new and existing policies to cross-functional processes that incorporate, by design, user feedback loops, external research, and user experience metrics.

C3 Product Enforcement: Conduct enforcement operations to implement product governance.

Minimum Maturity	Overall Maturity	Maximum Maturity
Ad Hoc	**Defined**	Managed

Product Enforcement featured the highest levels of aggregate maturity across the commitments, including both the highest floor and highest ceiling in terms of performance.

Five practices within this commitment were reported as Optimized in one or more assessments. More importantly, one practice had the highest minimum maturity across the assessments. Investing in the wellness and resilience of content review teams was uniformly at the Defined level or above, suggesting a common level of focus upon this practice across the industry.

Focus Practices for Product Enforcement

Formalize training and awareness programs to keep pace with dynamic online content and related issues, to inform the design of associated solutions

Assessments reported relatively high levels of maturity with regard to training programs. Practices that were assessed as Optimized featured centralized and standardized training provided to relevant individuals, including employees, vendors, and third-party labor, at regular intervals. Onboarding and other training courses detail specific procedures and protocols for topics such as manual content review and flagging, as well as defined processes to generate guidance regarding new or updated content policies, with different types of training materials based on whether the substance of the policy update and its impact is likely to be small or large.

Invest in wellness and resilience of teams dealing with sensitive materials, such as tools and processes to reduce exposure, employee training, rotations on/off content review, and benefits like counseling

Supporting the wellness and resilience of teams that deal with sensitive materials, whether full-time employees or external workers, is a top priority recognized across all the assessments. Several assessments noted working with academic experts and psychological professionals to evolve wellness programs.

Methods for providing wellness and resilience resources identified across the assessments focused on resources for employees as well as approaches to working with vendor partners who provide contracted content moderation services.

For employees, the types of resources and approaches described across the assessments include the following:

- Integration into training and onboarding;
- Special training for managers on how to identify wellness risks and signs of burnout resulting from engagement with graphic content;
- General wellbeing benefits;
- Provision of mental health services, including on-demand support to content reviewers;
- Regional support for content reviewers including group education and one-to-one wellness coaching sessions with a trauma-informed lens geared towards the most at-risk employees; and
- Monitoring of how these benefits are used and their impact, via employee surveys.

For vendor partners, assessments noted the importance of implementing wellness standards, safeguards, and controls in contract terms. Examples of standards and controls include:

- Limiting the number of hours reviewing sensitive materials;
- Diversifying content types to ensure that a particular vendor is not over-exposed to sensitive materials;

- Providing extended break times for reviewers dealing with sensitive materials;
- Reporting requirements for vendor partners back to the company on the implementation of their wellbeing activities; and
- External audit of vendor partners, including testing psychological health and wellbeing controls.

Assessments also noted the use of technology and tooling features geared toward wellbeing and resilience, including:

- Automated rotations and breaks;
- De-duplication of similar content; and
- Limiting audio and visual exposure to harmful content.

Work with recognized third parties (such as qualified fact checkers or human rights groups) to identify meaningful enforcement responses

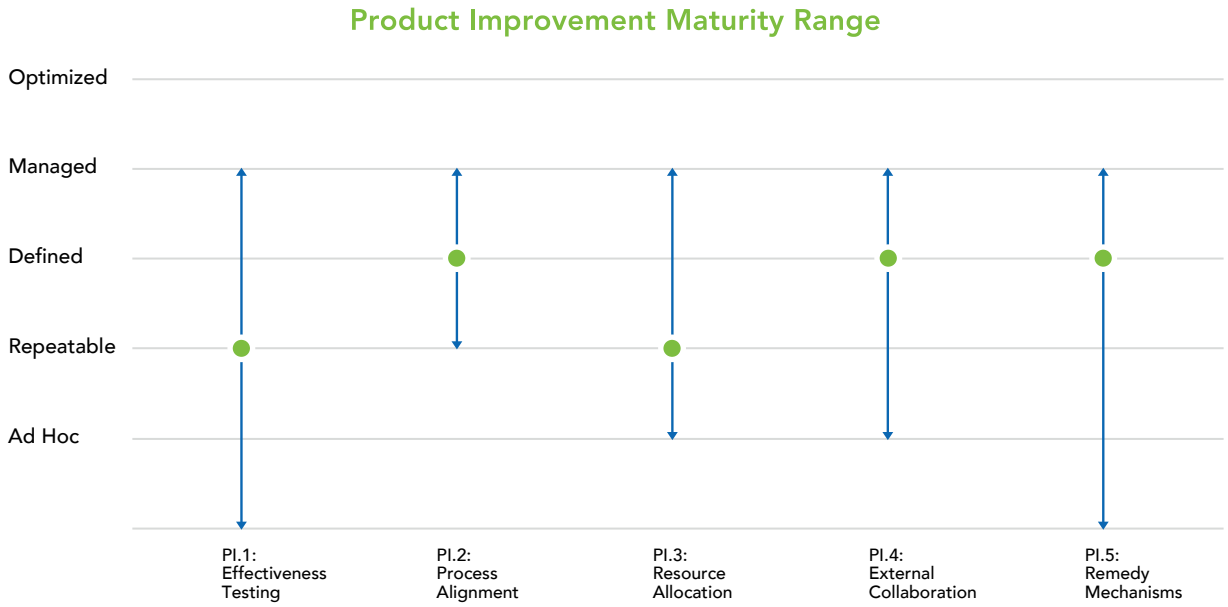
This practice concerns the involvement of third parties in enforcement, which is distinct from the involvement of external experts in policy development. For example, it includes programs where external fact checkers are involved in enforcement actions on types of content. This collaboration with third parties on enforcement lagged behind several other enforcement practices across the assessments. Notably, although some companies described the relationships they have developed with key third parties, such as civil society organizations, these engagements are often conducted at the corporate level or by public policy teams, rather than via the teams that conduct enforcement responses. Assessments more commonly pointed to providers of technical support and tools, whether vendors or nonprofit partners, with regard to this practice.

Creating channels for teams directly responsible for Trust & Safety to engage with recognized third parties may engender greater trust between stakeholders with benefits for both the company and external stakeholders.

C4 Product Improvement: Assess and improve processes associated with content- and conduct-related risks.

Minimum Maturity	Overall Maturity	Maximum Maturity
Ad Hoc	**Repeatable**	Managed

Product Improvement is the least mature Commitment across the DTSP Best Practices Framework. This may reflect that industry views of Trust & Safety have relatively recently shifted from a narrow focus on content moderation to a more holistic approach that encompasses the product development lifecycle, with a recognized focus on continuous improvement and quality assurance.



Focus Practices for Product Improvement

Use risk assessments to determine allocation of resources for emerging Content- and Conduct- Related Risks

Several DTSP Best Practices pertain to risk identification and assessment. Notably, the practice of using risk assessments to drive resource allocation across emerging risks as part of the Product Improvement Commitment lagged behind the use of these assessments in Product Development. Some assessments described collaboration between Trust & Safety, policy, and product teams to develop a methodology for ad hoc risk assessments based on product launches and other key events, but noted the need for more mature capabilities. These include: the performance of annual risk assessments to identify and report on top risks areas, the development of systemic risk strategies, enabling systemic risk assessments that will be required under regulation, and reporting to regulators and the public.

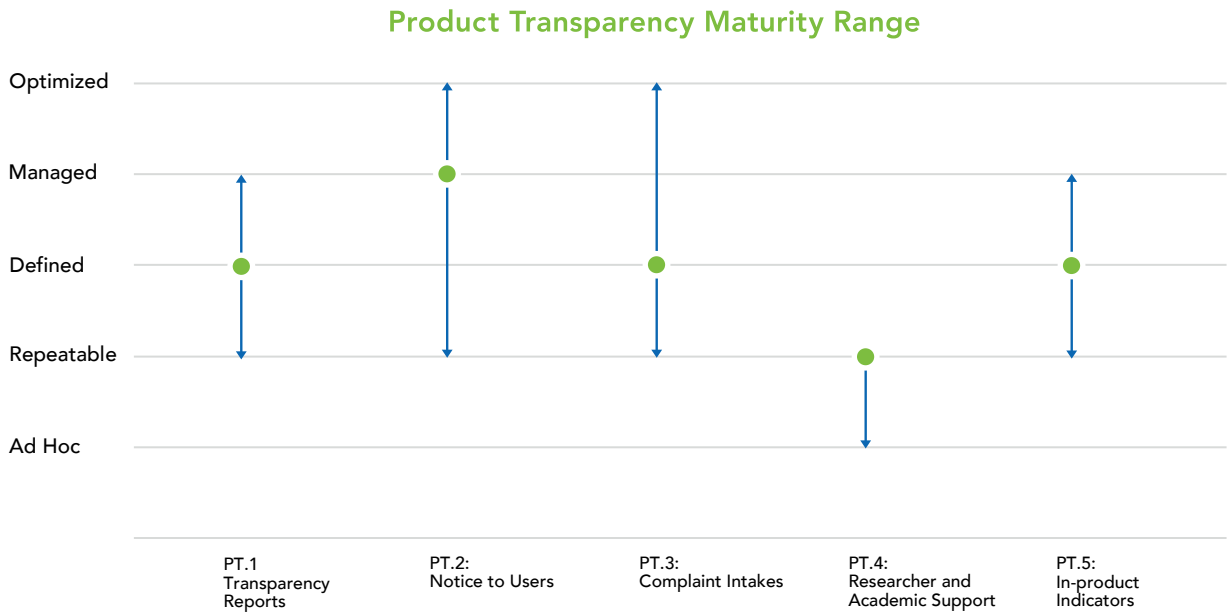
Mature aspects of this practice included proactive and regularly managed reviews of policy enforcement and the use of impact probability matrices to prioritize high-probability/high-impact risk areas to be managed.

Another noteworthy practice is developing methods of risk assessment to allocate resources for human review. This includes the development of measurements of the reduction of harmful experience per dollar spent on human review. Using Machine Learning models, estimates of prevalence can be developed that inform how to efficiently make the most use of human reviewers.

C5 Product Transparency: Ensure that relevant trust & safety policies are published to the public, and report periodically to the public and other stakeholders regarding actions taken.

Minimum Maturity	Overall Maturity	Maximum Maturity
Defined	**Defined**	Managed

Product Transparency is an area where company practices have advanced considerably in the past decade through the advent of transparency reporting, as well as other key practices including providing notice to users subject to enforcement actions and creating in-product indicators. Although the assessments showed considerable similarity in levels of maturity around these practices, all assessments showed opportunities to increase maturity.



Focus Practices for Product Transparency

Publish periodic transparency reports including data on salient risks and relevant enforcement practices, which may cover areas including abuses reported, processed, and acted on, and data requests processed and fulfilled

In a relatively short period of time, transparency reporting has emerged as a key industry best practice across digital services, with the majority of assessment of this practice indicating a Managed maturity level.

Notable metrics included in transparency reports by one or more companies include the following:

- 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 specific policies (including harassment or abusive, misinformation, hateful or derogatory, violent or graphic, adult, and child exploitation);
- Copyright removals (including number of requests, total infringements reported, reported infringements removed/rejected); and
- Number and types of government requests received and actioned (including requests for user data and for content removal).

Enhancements to transparency reporting that one or more companies indicated they would be planning in the future include the following:

- Moving from semi-annually to quarterly reporting;
- Broadening transparency statistics and including more metrics and details about enforcement;
- Production of increased jurisdiction-specific transparency reporting as required by regulation;
- Sharing more about the use of automated tools in content moderation, including on how they impact enforcement actions and how human review is involved; and
- Further streamlining data collection and enhancing reporting through collaboration between Trust & Safety and product, engineering, and data analytics teams.

4. Areas of Future Opportunity and Development

These Safe Framework assessments, a cross-industry assessment of Trust & Safety practices, are the culmination of more than a year of work. However, this is just the beginning of a process of evaluation and improvement.

Each assessment identified areas of opportunity and development for the partner companies:

- Several assessments described the rollout of new Trust & Safety frameworks and formal processes expected to have a substantial impact across all Commitments, with a particular focus on Product Development and/or Product Governance. It is expected that these frameworks will help to clarify who is accountable, outline structures, and encourage oversight.

- Assessments indicated that certain best practices share common characteristics with elements in notable forthcoming regulations. For example, practices related to risk assessment may overlap with aspects of the European Union's Digital Services Act and the U.K. Online Safety Bill. Some companies indicated they will place specific emphasis on these provisions.

The process of conducting Safe Framework assessments generated useful feedback and recommendations for DTSP. These include the following:

- The opportunity to review, refine, and in some cases potentially clarify best practices where there may be duplication or overlap;

- Opportunities for companies to identify and share innovative practices to potentially be added to the DTSP Best Practices framework;

- Clarifying the maturity scale used in the assessment to improve objectivity, especially in preparation for third-party assessments; and

- The development of shared tools and resources that can help new companies to undertake future assessments.

Areas of opportunity for external stakeholders:

- Legislators and regulators that are exploring content regulations should look to the DTSP Best Practices Framework to understand how industry is addressing content- and conduct-related risks and the diversity of products and services related to hosting and publishing content; and

- As governments encourage or mandate companies to share data with researchers, they should also consider ensuring that clear legal frameworks support this practice, assuring security and privacy while fostering collaboration. We expect improved legal frameworks in this space to enable more mature practices.

5. Looking Forward

This report is an inaugural effort to implement a model methodology for benchmarking the Trust & Safety discipline. It is the culmination of more than a year of work by our partner companies, who have dedicated significant resources to collaboratively building and deploying our framework of best practices and assessments. Still, much work remains to be done.

Evolving the DTSP Best Practices Framework

The execution of the Safe Framework, as well as the consultations DTSP held with approximately 120 participants from 27 countries (see Appendix II), have generated substantial feedback and recommendations, which we will incorporate into a review of the DTSP Best Practices Framework to take place later this year. We will evolve and improve as technology and expectations change. We anticipate that the scoping of future assessments may change based on what we continue to learn from future third-party assessments. In particular, we expect that questions raised during initial assessments will inform DTSP of areas that will be further explored in the future.

Moving to Third-Party Assessment

There is no substitute for independent, objective and measurable assessments. DTSP partners are taking the lead in our process because Trust & Safety practitioners inside companies have unique understanding and visibility into how Trust & Safety functions, but a company-only approach will not suffice. DTSP is working with experts to develop our approach to third-party assessment, and we will share more information on concrete next steps in the coming months.

Engaging with Stakeholders

We are accelerating our efforts to raise awareness of the DTSP Best Practices Framework and engage with diverse stakeholders globally. Going forward, we will be putting in place specific mechanisms for stakeholder input and engagement and providing opportunities for dialogue. DTSP is also engaging with international public-private partnerships and multistakeholder initiatives as part of this work, including through the World Economic Forum's Global Coalition for Digital Safety.

Appendices

Appendix I: Links to Key Documents

The DTSP Best Practices Framework and the methodology used to complete industry assessments have been published, and are available below. We continue to accept third-party feedback and welcome a chance to collaborate.

- DTSP Best Practices Framework
- The Safe Framework
- Assessment Results Survey

Appendix II: Summary of Stakeholder Consultations

When DTSP released the Safe Framework in December 2021, we initiated a stakeholder consultation with specific questions posed for feedback. To engage external stakeholders globally, we organized three virtual meetings, held to accommodate stakeholders located in Asia-Pacific, the Americas, and Europe, Middle East, and Africa. Approximately 120 participants joined the meetings from 27 countries. Participants included representatives from government agencies and regulators, intergovernmental organizations, a wide array of academic experts, and NGOs including child safety organizations, digital rights groups, and think-tanks and multi-stakeholder initiatives. DTSP also received written submissions from several think tanks and academic organizations.

The stakeholder consultations discussed the following topics:

1. General discussion on the DTSP Best Practices Framework and the Safe Framework assessment methodology;
2. Weighting commitments and practices;
3. How to provide meaningful transparency while mitigating safety risks; and
4. How industry best practices relate to regulation around the world.

DTSP general discussion

- Publication of the assessment results (at least in part) will help the public to monitor and track performance over time;
- The cost of assessment: consideration for small businesses;
- Involvement of the Trust & Safety teams in assessment to be able to raise concerns and provide feedback;
- Incorporating users/stakeholders' input in companies' practices;
- Not for profit tech corporations and access to research and expertise through DTSP;
- Be mindful of subjective terms such as "harm" and "trust";
- The framework should allow for incorporating external stakeholders' feedback and change accordingly;
- Product enforcement should include actions other than content removal;
- Third-party collaboration should be clarified and collaboration with governments specifically mentioned; and
- The framework and the assessments can help find gaps in companies' practices and experts can help them overcome those issues.

How to weight commitments and practices

- Flexibility is important but there have to be guidelines in place so that companies do not choose the easiest issues to assess themselves;
- All the best practices should be measured equally;
- Long term issues and future products and services should also be born in mind during assessment;
- Assess commitments holistically and not in isolation;
- In product governance, it is important for the best practices to benefit from various nonbinding civil society and other multi-stakeholder initiatives recommendations and principles that take place in other forums;
- Priorities and goals have a substantial role in weighing the commitments, however some stakeholders believe that preferential treatment should not apply to weighing commitments and all of them should be equally weighed; and
- Internal risks (technology and algorithms that platforms use for example) and external risks should be separated.

How to provide meaningful transparency while mitigating safety risks

- Transparency should be incorporated at every stage but also it has to be considered not to reveal secrets that can make safety practices ineffective;
- Establish transparency for whom and for what purposes; and
- Looking at other sectors' transparency approaches might be useful to establish particular actions or processes or to engage stakeholders such as researchers and regulators.

How industry best practices relate to regulation around the world

- DTSP can transform its learnings to provide feedback for policymakers;
- Some best practices can relate to the Digital Services Act;
- As well as regulatory framework, it is important to see how these best practices relate to Human Rights Impact Assessment frameworks;
- Use a modular approach to relate each of the commitments/best practices to a piece of local legislation;
- Look at a menu of levers to align with practices;
- Do comparative analysis with other audit and standard setting organizations such as in the aviation industry;
- DTSP should inform the development of public policy on content moderation:
  - Through creating a duty of care framework; and
  - Tracking which best practices are similar to regulatory frameworks around the world.

Public Comments

The public comments generally addressed the need for DTSP to do the following:

- To publish data to enable researchers and policymakers to evaluate the impact of content moderation practices;
- To develop product-specific moderation norms that are applicable to companies across the tech stack; and
- To inform the development of public policy on content moderation.

Some public comments also provided feedback on assessment questions:

- More clarification on the question of user control on products;
- Adding a separate question on processing government content flag request;
- Adding more context to what types of actions may be taken in case of policy violation; and
- Clarifying what collaboration with academic and other researchers mean.

Assessment results and informing the work of external stakeholders:

Generally, the publicly shared information on assessment results can allow external stakeholders to understand key Trust & Safety trends for each of the commitments. The public comments also mentioned that there is interest from the external stakeholders on what efforts companies are taking to address the gaps, opportunities to collaborate in a multi-stakeholder fashion around some issues, centralize member companies transparency reports and include DTSP assessment reports in that central location and create a repository of its materials and where relevant link to member company reports, documents, etc. The public comments also mentioned that DTSP's multi-stakeholder approaches (if any) should stay focused on illegal content that is harmful to users and the society. Some stakeholders expressed their concern with going beyond the legal definition of harmful content at DTSP.

Appendix III: Links to Publicly Available Company Resources

Discord
Community Guidelines
Transparency Report

Google
Community Guidelines
Transparency Report

LinkedIn
Professional Community Policies
Transparency Report

Meta Platforms, Inc.
Community Standards
Transparency Center

Microsoft
Microsoft Services Agreement
Digital Trust Reports

Pinterest
Community Guidelines
Transparency Report

Reddit
Content Policy
Transparency Report

Shopify
Acceptable Use Policy
Transparency Report

Twitter
Rules and Policies
Transparency Report

Vimeo
Acceptable Use Policy
Transparency Report

Appendix IV: DTSP Assessment Results Survey

This survey was distributed to each DTSP Participating Company to facilitate reporting Safe Framework assessment results to DTSP.

Respondent Contact Information

Respondent Name

Respondent Email

Company Name

Role, Title, and/or Function

Other Points of Contact (if applicable)

Section A: Scoping Questions

The following questions are used to capture information related to the scope of your completed DTSP Assessment.

A.1	Please indicate the subject of your DTSP Assessment.	
	One product or digital service	☐
	Multiple products or digital services	☐
	A single function (e.g., the central Trust & Safety function of the company)	☐
	Other:	

A.2	Optional: Please specify the name(s) of the product(s), digital service(s), or function name(s) represented in the previous section.

A.3	Please indicate what level of assessment (as defined in the Safe Framework) was conducted.	
	Level 1 assessment	☐
	Level 2 assessment	☐
	Level 3 assessment	☐

Section B: General Questions

The following questions relate to the overall assessment, across all five Commitments rather than any one particular Commitment. Subsequent sections of this survey will include specific questions regarding each DTSP Commitment.

B.1	Please share how your company chose which best practices to assess during its DTSP Assessment.

(E.g. We chose the best practices that we expected were practiced at a maturity level of "1" (Ad Hoc) or higher)

B.2	Please share the DTSP Commitments or best practices that your company is focused on developing or seeking to enhance going forward. Please summarize how you expect to strengthen or develop these DTSP Commitments or best practices.

(E.g. We are focused on increasing maturity related to DTSP Commitment 5 and will be created automated transparency tools to streamline publication of relevant data metrics)

B.3 Please share how your company may have prioritized which DTSP Commitments or best practices it will develop or enhance.

(E.g. Our company prioritized best practices that were either implicated by regulation or a high concern for users)

Section 1: Commitment 1 | Product Development

Commitment 1: Identify, evaluate, and adjust for content- and conduct-related risks in product development

Maturity Scoring

Below is the scorecard for DTSP Commitment 1. Please provide the score that you assessed for each Best Practice, based on your completed DTSP Assessment. If a DTSP Best Practice was not assessed, select "N/A".

Below is a list of DTSP Best Practices related to DTSP Commitment 1.
Please provide maturity assessment scoring from your previously completed DTSP Assessment:

PD.1: Develop insight and analysis capabilities to understand patterns of abuse and identify preventive mitigations that can be integrated into products

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PD.2: Include Trust & Safety team or equivalent stakeholder in the product development process at an early stage, including through communication and meetings, soliciting and incorporating feedback as appropriate

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PD.3: Designate a team or manager as accountable for integrating Trust & Safety feedback

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PD.4: Evaluate Trust & Safety considerations of product features balancing usability and the ability to resist abuse

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PD.5: Use in-house or third-party teams to conduct risk assessments to better understand potential Risks

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PD.6: Provide for ongoing pre-launch feedback related to Trust & Safety considerations

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PD.7: Provide for post-launch evaluation by the team accountable for managing risks and those responsible for managing the product or in response to specific incidents

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PD.8: Iterate product in light of Trust & Safety considerations including based on user feedback or other observed effects, including ensuring that the perspectives of minority and underrepresented communities are represented

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PD.9: Adopt appropriate technical measures that help users to control their own product experience where appropriate (such as blocking or muting)

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Additional Questions Related to DTSP Commitment 1

The following questions provide additional insight into the scoring of Commitment 1. Please answer each of the following questions to the best of your knowledge.

1.1 Does your company expect that any of the DTSP Best Practices under this commitment will change maturity ratings in the near future? If so, how and why?

(E.g. Yes, we expect that Best Practice PD.1 will increase in approximately six months after we finish initial training a new machine learning tool that will identify preventive mitigations that can be integrated into products)

1.2 Have there been any recent tooling or technology adoptions that may have recently changed the maturity rating of any DTSP Best Practices within this DTSP Commitment?

(E.g. Yes, we've adopted an automated development workflow that connects T&S to development teams)

1.3 If applicable, please provide any additional detail or commentary related to your insight and analysis capabilities to understand abuse patterns and identify mitigating measures.

(E.g. A dedicated analytics team has developed KPIs to track abuse trends)

1.4 How integrated or embedded is your Trust & Safety Team (or similar function) into the product development lifecycle? Are there particular areas of strength, or areas of opportunity for further development?

(E.g. The Trust & Safety Team coordinates with development to incorporate priorities based on compliance, industry standards, and user feedback)

1.5 What are examples of features, such as blocking or muting, that are incorporated into your product that allows users to control their product experience?

(E.g. Users are allowed to block or mute unwanted content or notifications)

1.6 Please provide any additional comments or details that you believe may be relevant and worth noting for the scores provided in this section.

Section 2: Commitment 2 | Product Governance

Commitment 2: Adopt explainable processes for product governance, including which team is responsible for creating rules, and how rules are evolved

Maturity Scoring

Below is the scorecard for DTSP Commitment 2. Please provide the score that you assessed for each Best Practice, based on your completed DTSP Assessment. If a DTSP Best Practice was not assessed, select "N/A".

Below is a list of the DTSP Best Practices related to DTSP Commitment 2.
Please provide maturity assessment scoring from your previously completed DTSP Assessment:

PG.1: Establish a team or function that develops, maintains, and updates the company's corpus of content, conduct, and/or acceptable use policies

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PG.2: Institute processes for taking user considerations into account when drafting and updating relevant Product Governance

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PG.3: Develop user-facing policy descriptions and explanations in easy-to-understand language

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PG.4: Create mechanisms to incorporate user community input and user research into policy rules

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PG.5: Work with recognized third-party civil society groups and experts for input on policies

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PG.6: Document for internal use the interpretation of policy rules and their application based on precedent or other forms of investigation, research, and analysis

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PG.7: Facilitate self-regulation by the user or community to occur where appropriate, for example by providing forums for community-led governance or tools for community moderation and find opportunities to educate users on policies, for example, when they violate the rules

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Additional Questions Related to Commitment 2

The following questions provide additional insight into the scoring of Commitment 2. Please answer each of the following questions to the best of your knowledge.

2.1 Does your company expect that any of the DTSP Best Practices under this commitment will change maturity ratings in the near future? If so, how and why?

(E.g. Yes, we expect that Best Practice PG.1 will increase in approximately six months after we dedicate and train a team to maintain the company's corpus of content, conduct, and acceptable use policies)

2.2 Have there been any recent tooling or technology adoptions that may have recently changed the maturity rating of any DTSP Best Practices within this DTSP Commitment?

(E.g. Yes, we've created an automated system to manage/update policies)

2.3 How are terms of service, policies, or guidelines related to the product maintained and updated? For example, at what frequency are they reviewed? When are users notified of updates?

(E.g. Terms, policies, and guidelines are presented upon registration or onboarding. Community/ guidelines specific to any one community are communicated upon entering the community)

2.4 Please describe any collaboration and engagement with users and user communities as it relates to terms of service, policies, or guidelines. For example, how is user input or research taken into account?

(E.g. Yes, user input is reviewed and incorporated into changes during an semi-annual review of rules and guidelines)

2.5 Are different teams responsible for creating and updating policies compared to those responsible for updating and executing procedures? If yes, do the teams collaborate?

(E.g. Yes, policies are written by compliance and procedures are written by the affected team, with guidance and approval from compliance)

2.6 Please provide any additional comments or details that you believe may be relevant and worth noting for the scores provided in this section.

Section 3: Commitment 3 | Product Enforcement

Commitment 3: Conduct enforcement operations to implement product governance

Maturity Scoring

Below is the scorecard for DTSP Commitment 3. Please provide the score that you assessed for each Best Practice, based on your completed DTSP Assessment. If a DTSP Best Practice was not assessed, select "N/A".

Below is a list of the DTSP Best Practices related to DTSP Commitment 3.
Please provide maturity assessment scoring from your previously completed DTSP Assessment:

PE.1.1: Constitute roles and/or teams within the company accountable for policy creation, evaluation, implementation, and operations

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PE.1.2: Develop and review operational infrastructure facilitating the sorting of reports of violations and escalation paths for more complex issues

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PE.1.3: Determine how technology tools related to Trust & Safety will be provisioned (i.e., build, buy, adapt, collaborate)

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PE.2: Formalize training and awareness programs to keep pace with dynamic online content and related issues, to inform the design of associated solutions

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PE.3: Invest in wellness and resilience of teams dealing with sensitive materials, such as tools and processes to reduce exposure, employee training, rotations on/off content review, and benefits like counseling

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PE.4: Where feasible and appropriate, identify areas where advance detection, and potentially intervention, is warranted

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PE.5: Implement method(s) by which content, conduct, or a user account can be easily reported as potentially violating policy (such as in-product reporting flow, easily findable forms, or designated email address)

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PE.6.1: Operationalize enforcement actions at scale where standards are set for timely response and prioritization based on factors including the context of the product, the nature, urgency, and scope of potential harm, likely efficacy of intervention, and source of report

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PE.6.2: Operationalize enforcement actions at scale where appeals of decisions or other appropriate access to remedy are available

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PE.6.3: Operationalize enforcement actions at scale where appropriate reporting is done outside the company, such as to law enforcement, in cases of credible and imminent threat to life

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PE.7: Ensure relevant processes exist that enable users or others to "flag" or report content, conduct, or a user account as potentially violating policy, and enforcement options on that basis

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PE.8: Work with recognized third parties (such as qualified fact checkers or human rights groups) to identify meaningful enforcement responses

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PE.9: Work with industry partners and others to share useful information about Risks, where consistent with legal obligations and security best practices

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Additional Questions Related to DTSP Commitment 3

The following questions provide additional insight into the scoring of Commitment 3. Please answer each of the following questions to the best of your knowledge.

3.1 Does your company expect that any of the DTSP Best Practices under this commitment will change maturity ratings in the near future? If so, how and why?

(E.g. Yes, we expect that Best Practice PE.3 will increase in approximately three months after the company implements free wellness resources and mandatory rotations)

3.2 Have there been any recent tooling or technology adoptions that may have recently changed the maturity rating of any DTSP Best Practices within this DTSP Commitment?

(E.g. Yes, we recently adopted machine learning that will automatically process user reports/flags that contain a high level of confidence of accuracy)

3.3 Please share examples, if applicable, of training or awareness programs for enforcement operations.

(E.g. Enforcement operations teams participate in comprehensive onboarding training and semi-annual trainings)

3.4 If applicable, what wellness or resilience methods are being used for those people within enforcement operations, particularly for teams dealing with sensitive materials?

(E.g. The company requires mandatory rotation and breaks every three hours)

3.5 Please share examples, if applicable, of tools, technologies, or methods used to proactively detect or mitigate potentially violative content or conduct.

(E.g., Machine learning algorithms review public content for explicit imagery and flag suspicious content for review)

3.6 Please share a brief summary of any work or collaboration with third parties related to enforcement operations. For example, how do you work with trusted flaggers, government reporters, fact checkers, users, or other stakeholders? Are there separate processes?

(E.g. The company engages with trusted third parties to establish methods for effective enforcement and as a result has implemented additional options for government or other trusted reporter/flaggers)

3.7 Please provide any additional comments or details that you believe may be relevant and worth noting for the scores provided in this section.

Section 4: Commitment 4 | Product Improvement

Commitment 4: Assess and improve processes associated with content- and conduct-related risks

Maturity Scoring

Below is the scorecard for DTSP Commitment 4. Please provide the score that you assessed for each Best Practice, based on your completed DTSP Assessment. If a DTSP Best Practice was not assessed, select "N/A".

Below is a list of the DTSP Best Practices related to DTSP Commitment 4.
Please provide maturity assessment scoring from your previously completed DTSP Assessment:

PI.1: Develop assessment methods to evaluate policies and operations for accuracy, changing user practices, emerging harms, effectiveness and process improvement

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PI.2: Establish processes to ensure policies and operations align with these Commitments

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PI.3: Use risk assessments to determine allocation of resources for emerging content- and conduct-related risks

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PI.4: Foster communication pathways between the Practicing Company on the one hand, and users and other stakeholders (such as civil society and human rights groups) to update on developments, and gather feedback about the social impact of product and areas to improve

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PI.5: Establish appropriate remedy mechanisms for users that have been directly affected by moderation decisions such as content removal, account suspension or termination

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Additional Questions Related to Commitment 4

The following questions provide additional insight into the scoring of Commitment 4. Please answer each of the following questions to the best of your knowledge.

4.1 Does your company expect that any of the DTSP Best Practices under this commitment will change maturity ratings in the near future? If so, how and why?

(E.g. Yes, we expect that Best Practice PI.5 will increase in approximately a year after the company implements an appeals process for those that have been affected by an enforcement action)

4.2 Have there been any recent tooling or technology adoptions that may have recently changed the maturity rating of any DTSP Best Practices within this DTSP Commitment?

(E.g. Yes, a tool has been deployed that will allow users that are subject to an enforcement action to appeal the decision)

4.3 If applicable, could you please provide a brief description of how risk assessments related to content- and conduct-related risks are performed? For example, how often are they conducted? Which individual(s) or team(s) are involved?

(E.g. The Trust & Safety Team performs risk assessments across all products and services)

4.4 Please share a few of the processes, tools, or technologies that may be used to assess the policies and operations that mitigate content- and conduct-related risk. For example, how are opportunities for future development or improvement identified?

(E.g. A proprietary workflow schedules and facilitates periodic review by at least one person from the Trust & Safety Team)

4.5 Please provide any additional comments or details that you believe may be relevant and worth noting for the scores provided in this section.

Section 5: Commitment 5 | Product Transparency

Commitment 5: Ensure that relevant trust & safety policies are published to the public, and report periodically to the public and other stakeholders regarding actions taken

Maturity Scoring

Below is the scorecard for DTSP Commitment 5. Please provide the score that you assessed for each Best Practice, based on your completed DTSP Assessment. If a DTSP Best Practice was not assessed, select "N/A".

Below is a list of the DTSP Best Practices related to DTSP Commitment 5.
Please provide maturity assessment scoring from your previously completed DTSP Assessment:

PT.1: Publish periodic transparency reports including data on salient risks and relevant enforcement practices, which may cover areas including abuses reported, processed, and acted on, and data requests processed and fulfilled

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PT.2: Provide notice to users whose content or conduct is at issue in an enforcement action (with relevant exceptions, such as legal prohibition or prevention of further harm)

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PT.3: Log incoming complaints, decisions, and enforcement actions in accordance with relevant data policies

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PT.4: Create processes for supporting academic and other researchers working on relevant subject matter (to the extent permitted by relevant law and consistent with relevant security and privacy standards, as well as business considerations, such as trade secrets)

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

PT.5 Where appropriate, create in-product indicators of enforcement actions taken, including broad public notice (e.g., icon noting removed content providing certain details), and updates to users who reported violating content and access to remedies

☐ N/A ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Additional Questions Related to DTSP Commitment 5

The following questions provide additional insight into the scoring of Commitment 5. Please answer each of the following questions to the best of your knowledge.

5.1 Does your company expect that any of the DTSP Best Practices under this commitment will change maturity ratings in the near future? If so, how and why?

(E.g. Yes, we expect that Best Practice PT.4 will increase in approximately a year after the company launches a research incubator that will provide grants and API access to enforcement action metadata)

5.2 Have there been any recent tooling or technology adoptions that may have recently changed the maturity rating of any DTSP Best Practices within this DTSP Commitment?

(E.g. Yes, we recently implemented autonomous workflows for notifying users about relevant enforcement actions)

5.3 Could you please provide a brief description of any process(es) in place to log user complaints, decisions, enforcement and remedy actions?

(E.g. Metadata about complaints, decisions, enforcement, and remedy are retained for six months)

5.4 Could you please share any examples of processes or mechanisms by which transparency is provided to the user around enforcement decisions or actions taken on content or conduct?

(E.g. Notices are provided to users after necessary administrative actions have been taken, which generally occurs within 48 hours of an enforcement action is taken)

5.5 Please share how frequently and by what methods transparency reports are published. Are there plans to further develop or expand these efforts in the future?

(E.g. Transparency reports are published once a year within two months of year-end. Cadence is expected to increase to twice a year)

5.6 Please provide any additional comments or details that you believe may be relevant and worth noting for the scores provided in this section.