Digital Trust
& Safety Partnership

The Safe Framework

Tailoring a Proportionate Approach to Assessing Digital Trust & Safety

December 2021

Table of Contents

Digital Trust
& Safety Partnership

Message From The Executive Director

The Digital Trust & Safety Partnership ("DTSP") launched in February 2021 to mature and professionalize Trust & Safety, the field of industry professionals dedicated to a safer and more trustworthy internet.

Earlier this year, DTSP released the first set of common Trust & Safety commitments by leading technology companies, articulating 35 best practices our members are using to keep users of digital services safe from abuse. Since then, our members have been working diligently to develop a framework for assessing and evaluating these practices that we are showcasing in this publication. We have also co-hosted events with peer organizations and engaged extensively with experts from industry, governments, and civil society.

This publication provides the public with an update on our primary task: constructing a robust assessment process that will increase trust in digital services by enabling independent third-party assessments of how companies are using our best practices. We are presenting our framework and seeking input from stakeholders, which we will continue through a series of conversations in the coming months.

The progress we have achieved has taken extensive discussion, deliberation, and research, and it will take time for us to reach our next steps, because it is imperative that we get this right. We recognize the urgency of this work, but there is no one-size-fits-all approach to Trust & Safety. In a world where digital services face unique and constantly changing risks, we are working to not only respond to the issues in the headlines today but build a process that can handle the safety challenges over the horizon that we cannot yet foresee.

The past two years have made it more clear than ever that the internet powers so much of the best in our world: economic opportunities, connections with friends and family, education, and access to information, and so much more. Although nearly everyone agrees we need a safer and more trustworthy internet, few agree on what that means and the steps to get there. Our contribution to the global debate about the future of digital services is industry insight into what Trust & Safety excellence entails, backed by a durable third-party assessment framework that provides assurance to users. Having witnessed the collaboration among DTSP members, I'm enthused about the progress we've achieved and I look forward to sharing more of our learnings in the near future.

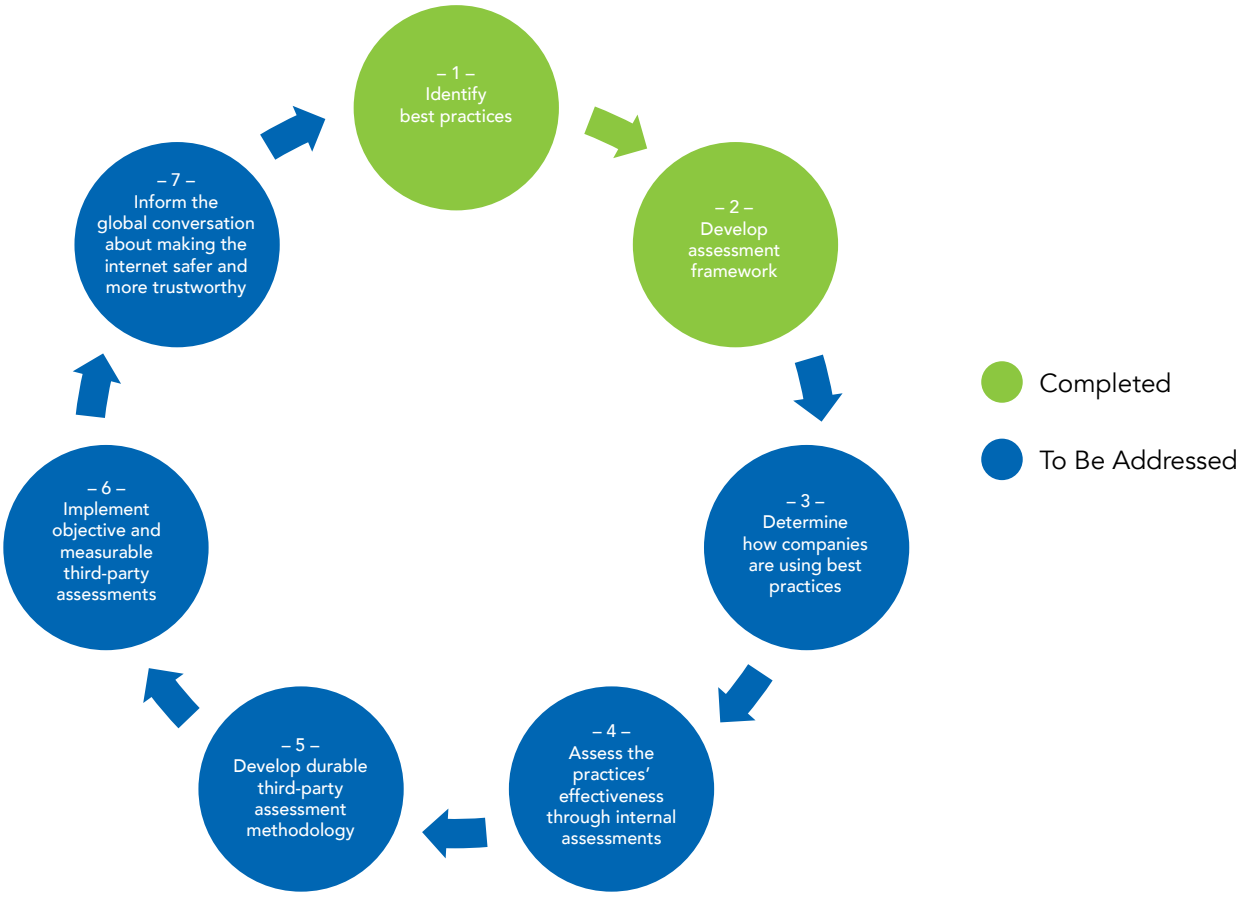
As DTSP's first executive director, I would like to thank all the advocates, policymakers, and experts that have engaged with us so far, and will encourage and welcome further dialogue, constructive criticism, and active participation. We look forward to continuing our work with you in the shared conversation about how to make the internet safer and more trustworthy.

David M. Sullivan
Executive Director
DTSP

Executive Summary

For the internet to continue to be an enabler of innovation, connection, and expression, we need digital products and services that are safe and trustworthy. "Trust & Safety" is the industry term for the internal teams taking on the most difficult areas of the internet to reduce and prevent harm to people on and offline. There is no one-size-fits-all approach to this work, as each digital service faces unique risks, which are constantly changing.

A diverse range of innovative technology companies of different sizes and business models created the DTSP to articulate Trust & Safety best practices and establish a rigorous and flexible approach to their evaluation. We are publishing our assessment approach, the Safe Framework, the next step in our process of evaluation, learning, and accountability:



This is the first effort of its kind to present a common, risk-based approach to evaluating the adoption of best practices for Trust & Safety.

We believe that the Safe Framework will bring rigor and consistency to assessment processes, while also providing flexibility across the diverse products and services our members provide. As our initial internal assessments get underway, and as we develop a process of third-party verification for future assessments, we present this framework to be transparent about our work and seek input from key stakeholders.

The 11 current DTSP partner companies are Discord, Google, LinkedIn, Meta, Microsoft, Patreon, Pinterest, Reddit, Shopify, Twitter, and Vimeo. DTSP members have committed to five fundamental areas of best practices ("the Commitments") that a digital service must consider to promote a safer and more trustworthy internet. These 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 Digital Trust & Safety Partnership Best Practices Framework, which provide concrete examples of the kinds of activities and processes that organizations will have in place to mitigate risks from harmful content and conduct.

DTSP Inventory of 35 Best Practices

Product Development	Product Governance	Product Enforcement	Product Improvement	Product Transparency
Abuse Pattern Analysis	Policies & Standards	Policy Enforcement	Effectiveness Testing	Transparency Reports
Trust & Safety Consultation	User Focused Product Management	Violation & Escalation Management	Resource Allocation	User Notice
Accountability	Community Guidelines/Rules	Training & Awareness	External Communications	Complaint Intakes
Risk Assessment	User Input	Advanced Detection	Policy Alignment	Research & Academic Support
Mitigation & Control	External Consultation	Wellness Programming	Remedy Mechanisms	Complaint Response
Monitoring & Evaluation	Community Self Regulation	Rules & Responsibilities		In-Product Indicators
Ongoing Improvement		Internal Reporting		Abuse Reporting
User Control		Response Management		
		Risk Communication		

DTSP partners have diverse missions, business models, and ways that they communicate with their users. But each partner shares the common goal of Trust & Safety excellence. Each organization selects the combination of the best practices that mitigates the most relevant content- and conduct-related risk exposures for their products and services. Our next publication will provide more information about the use of these practices by our members, but based on initial baseline data all of our members are using 80 percent or more of the practices.

Initially, each organization will conduct an internal assessment, which evaluates adherence to the Commitments, documents existing practices, and identifies opportunities to improve them or develop new ones. An objective and measurable third-party assessment will follow. The goal of the assessment is to facilitate a baseline understanding of each organization's posture as it relates to digital Trust & Safety practices. More importantly, the assessments will help inform a common understanding of how the DTSP practices are being used to manage content- and conduct-related risks across the industry. This understanding will in turn help advance the Trust & Safety discipline, increase meaningful transparency, demonstrate accountability, and ultimately inform an eventual third-party assessment process.

Scoping and tailoring assessments to account for the diversity of digital services

The Safe Framework examines the people, processes, and technology that contribute to managing content- and conduct-related risks for member companies. In these inaugural assessments, each company will evaluate their practices in a flagship product, a bundle of features or products, or through a central Trust & Safety function. We anticipate that the scoping of future assessments will evolve based on learnings from initial assessments and future third-party assessments.

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. Impact is based on user volume and the presence of product features or complexity that introduce potential risks.

Executing assessments to deepen understanding and develop capacity

The assessment is designed to help organizations understand how DTSP practices will help them manage content- and conduct-related risks. The outcome of the assessment will help organizations better understand the current state of their capabilities and their dependencies with respect to people, processes, and technologies. The resulting understanding can inform internal investment decisions and external engagements with policymakers, users, and civil society.

Consulting with stakeholders to inform our future work

DTSP recognizes that there are important discussions occurring in homes, schools, businesses, and at various levels of government all around the world, on what digital Trust & Safety should look like. We continue to learn from these discussions, and intend to contribute to them, sharing our own insights and experiences from implementing the Safe Framework. DTSP will synthesize results across companies into overall industry analysis that we will share with the public. Concurrently, DTSP is also moving forward with plans for independent third-party verification of our best practices framework.

DTSP is also seeking input from other experts in the field on our own approach. As a first step toward stakeholder engagement, we invite comments on this paper and are posing specific questions that have arisen in our work. We seek feedback from consumer and user advocates, policymakers, law enforcement, relevant NGOs and various industry-wide experts as we consider our approach to the following issues:

- **How to weight commitments and practices:** Should each of the five DTSP Commitments be weighted equally? Are some of the best practices that support those commitments of greater importance than others?

- **How to provide meaningful transparency while mitigating safety risks:** How should DTSP and its member companies increase meaningful transparency? How can industry efforts complement company reporting? Can disclosure of assessment processes and results inform stakeholders without potentially providing malicious actors with information that could be used to avoid or subvert company policies and practices in ways that could cause harm?

- **How industry best practices relate to regulation around the world:** How can industry best practices be informed by current legislative and regulatory initiatives, and at the same time inform those processes to potentially help avoid conflicting requirements that could burden smaller organizations while posing risks to human rights, universal access to information, and innovation and economic opportunity?

Introduction

What is 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. 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 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.

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.

DTSP participants are committed to leading by example to develop a common industry approach that can inform the broader global conversation about digital Trust & Safety. By using and promoting industry best practices, which will then be reviewed through internal and third-party assessments, we will provide a practical means to ensure consumer Trust & Safety across a wide array of digital services and products. DTSP currently has 11 member companies: Discord, Google, LinkedIn, Meta, Microsoft, Patreon, Pinterest, Reddit, Shopify, Twitter, and Vimeo.

The DTSP Best Practices Framework

All participating companies in the DTSP have agreed on five fundamental Commitments that a digital service makes to promote a safer and more trustworthy internet.

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. These sample practices are summarized in the following graphic.

DTSP Inventory of 35 Best Practices

Product Development	Product Governance	Product Enforcement	Product Improvement	Product Transparency
Abuse Pattern Analysis	Policies & Standards	Policy Enforcement	Effectiveness Testing	Transparency Reports
Trust & Safety Consultation	User Focused Product Management	Violation & Escalation Management	Resource Allocation	User Notice
Accountability	Community Guidelines/Rules	Training & Awareness	External Communications	Complaint Intakes
Risk Assessment	User Input	Advanced Detection	Policy Alignment	Research & Academic Support
Mitigation & Control	External Consultation	Wellness Programming	Remedy Mechanisms	Complaint Response
Monitoring & Evaluation	Community Self Regulation	Rules & Responsibilities		In-Product Indicators
Ongoing Improvement		Internal Reporting		Abuse Reporting
User Control		Response Management		
		Risk Communication		

All DTSP partners embrace the Commitments, but 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.

Accounting for and keeping pace with technological change

The five Commitments and the supporting best practices are designed to tackle many challenging content- and conduct-related risks and are agnostic to particular technologies, so that they can evolve over time.

The DTSP Best Practices do, however, provide a means for companies to manage content- and conduct-related risks that arise from the use of technology, such as algorithmically driven products and services. By applying the DTSP Best Practices, companies can address concerns that have arisen around fairness, user controls, and transparency. These issues are particularly germane to the Product Development pillar.

The best practices also provide a means for companies to assess the use of technologies, such as artificial intelligence (AI) and machine learning, as part of Trust & Safety operations. For example, there is an increasing focus on the use of automated tools to manage the scale of policy enforcement. Automated tools use AI or machine learning to execute Trust & Safety operations. The algorithms that power them can be considered and managed across the five Commitments and through some supporting practices.

The DTSP assessment roadmap

DTSP participants have worked together to define Trust & Safety as a key business function, identified common Trust & Safety commitments, and outlined the best practices they are using to uphold these commitments. Going forward, participants are conducting internal assessments to better document these practices and develop a common understanding of how to evaluate and assess these practices.

Findings from these assessments will inform "state-of-the-industry" reports that document Trust & Safety practices across the membership, providing more information to the public about this discipline than has been made available previously. At the same time, DTSP will collaborate with industry assurance experts to develop an assessment framework that members can use to evaluate Trust & Safety practices and ultimately have them reviewed by independent third parties.

The Safe Framework

This section summarizes the DTSP assessment methodology, the Safe Framework.

Why this approach?

There is no one-size-fits-all approach to Trust & Safety operations. Companies provide very different digital products and services, which cater to unique communities with different expectations about online content and conduct. Each service faces unique risks relative to the various products or features they provide – different threats, different vulnerabilities, and different consequences.

Similarly, the approach that DTSP companies take to assessing themselves against best practices should be tailored to the nature of the companies themselves, as well as the digital products and services they provide. Since DTSP launched in February 2021, Trust & Safety practitioners from DTSP companies have convened on a regular basis to develop an initial methodology, the Safe Framework, that aims to provide a flexible approach that incorporates a company's size and scale, and the impact of its products in terms of user volume and features that may introduce risks. The Safe Framework will initially be used for internal assessments, and then provide the basis for independent third-party assessments.

As this is the first attempt by any organization to develop a Trust & Safety assessment methodology, we are publishing this overview of the methodology to be transparent about our work to-date, and to receive feedback and evolve our approach in the future.

Scoping assessments appropriately

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. To this end, DTSP companies will apply the Safe Framework to relevant people, processes, and technology that contribute to managing content- and conduct-related risks and that reflect existing practices.

For the inaugural assessments, organizations will evaluate existing practices used to identify and mitigate risks otherwise present in:

- A flagship product;
- A bundle of features or products; and/or
- A central Trust & Safety function.

We anticipate that the scoping of future assessments may change based on what is learned in this initial internal assessment as well as the third-party assessments that will follow thereafter.

Tailoring assessments proportionately

DTSP companies are varied in organizational size, scale, and resource capacity. Due to the diverse range of products and services they provide, they face a broad spectrum of content- and conduct-related risks, with varying levels of systematic impact on the digital ecosystem. Moreover, there are different degrees of maturity for Trust & Safety teams and practices across the membership's products and services.

To account for these differences, while embracing a common commitment to Trust & Safety excellence backed by internal and external evaluation, DTSP takes a tiered approach to assessment. The Safe Framework provides a common approach that can be applied by companies with very different resource levels without imposing the same requirements on products with dissimilar digital footprints. Moreover, it provides a means for companies at different stages of growth to assess their Trust & Safety practices consistently over time, as they evolve.

DTSP has proposed 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.

The tailoring framework defines common criteria that each company will use to determine an assessment level of detail that is proportionate to the distinct nuances and risks for each organization or product. It provides flexibility for each company to conduct an assessment tailored to the capabilities and maturity of the company or product being assessed, while defining common standards, terms, and goals.

The tailoring framework comprises the following components:

- **Organizational size and scale:** at a company level, consider the availability of resources, and financial capacity to address or mitigate Trust & Safety risks;
- **Product or digital service impact:** at a product or individual digital service level, consider systemic impacts on the digital ecosystem, as well as the content- or conduct-related risks associated with the product features or services offered; and
- **Business landscape considerations:** consider additional risk-based factors associated with the business landscape or environment in which the company and specific product/service are operating.

Applying this framework, each company can determine whether a Level 1, Level 2, or Level 3 assessment should be performed for a particular product or service.

i. **Evaluate the organization's size and scale**

It is important to establish a set of objective criteria for determining the size and scale of an organization. This component defines inputs for consideration that are indicative of an organization's size and scale:

- Previous year's revenue; and
- Total number of employees, or number of employees for the products/services in scope of assessment.

Together, these inputs are measured to categorize each company into a "low", "medium", or "high" classification.

Organizational Size/Scale Inputs	Resulting Categorization
Both inputs "Low"	Low
At least one "Medium" (and neither is "High")	Medium
At least one "High"	High

ii. **Evaluate the impact of the product or digital service**

Once organizations have applied the size and scale criteria, they evaluate their product or digital service risk drivers to inform their approach to the assessment. The following inputs measure a product or digital service's impact and associated risks:

User volume: measured as the average monthly active registered users over the past twelve months.[1] The broader the audience consuming the content or services of the product, the greater the impact of content- and conduct-related risks.

[1] Monthly active registered users is defined as the number of users with a registered account who logged in or otherwise authenticated to visit the product website, mobile website, desktop or mobile application, within the last 30 days, from the date of measurement.

Product feature risk and complexity: measured as the number of product or service features that may implicate content- or conduct-related risks. Certain features, such as live streaming or video sharing or hosting of user-generated content, can expand the risk landscape. In general, the more of these features that a product makes available to users, the more complex and broader the set of risks. Similarly, some business models or topic areas may lend themselves to particular risks — a health-oriented site might have a higher risk of medical misinformation, for example, but a lower risk of violent extremism or hate speech.

DTSP has developed a list of features that implicate content- or conduct-related risks to evaluate product feature risk and complexity. The number of features present is used to quantify risk as low, medium, or high.

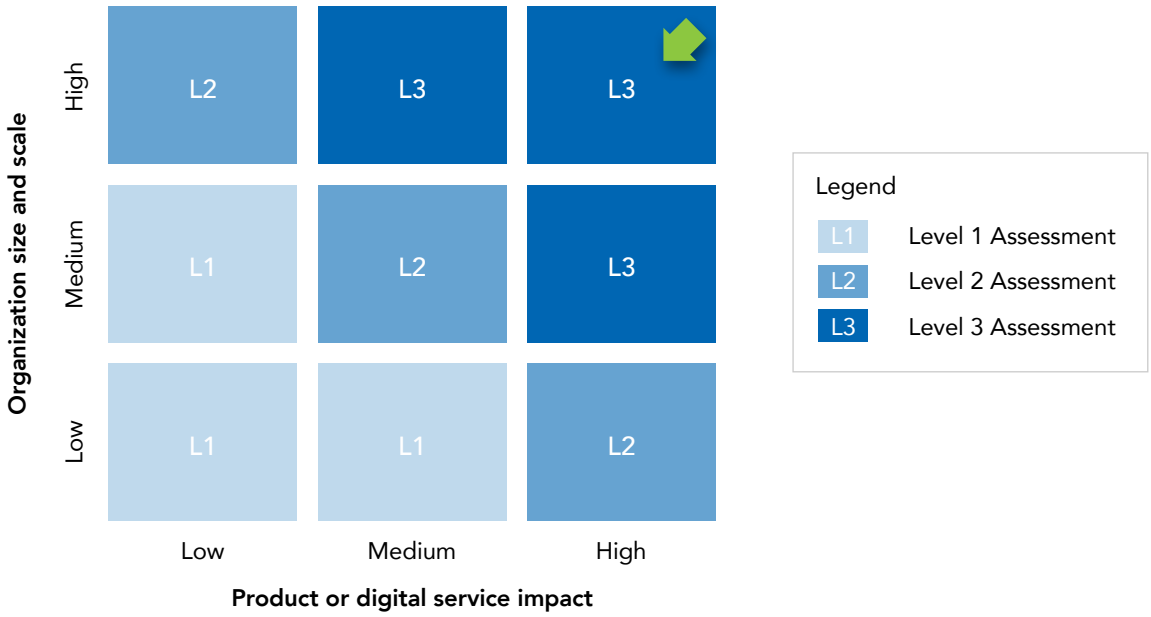
The measurements for each of the inputs are aggregated to determine an overall categorization of the product/service's impact.

Product/Service Impact Inputs	Resulting Categorization
Both inputs "Low"	Low
At least one "Medium" (and neither is "High")	Medium
At least one "High"	High

iii. **Determine the initial recommended assessment level**

The evaluations of organizational size and scale and product or service impact are combined to determine the initial recommended level of assessment, as depicted in the below matrix. The idea is that both the organizational size and scale and product impact should be factored in when contemplating a proportionate level of assessment.

For example, if a company is determined under organizational size and scale to be "high" and its product as "high impact", it is placed in the top-right box of the matrix, where a Level 3 assessment is recommended:



iv. **Additional business landscape considerations**

The final step in applying the tailoring framework involves integrating considerations related to the business landscape in which an organization or product is operating. This is an optional internal measure, relating to factors that are nonpublic or proprietary, where businesses may be aware of a factor that could justify a different level of assessment. It is anticipated that these business landscape considerations would generally be used to increase the recommended level of assessment, rather than decrease it.

The level of assessment should be informed by any unique circumstances or events that may impact the risks that a particular product or digital service must navigate. For example, a company may be aware of factors that may increase risk and merit a higher level of assessment (e.g., if the product

was due to expand into new markets). In addition, a company may have information that does not become apparent in the initial determination of the assessment level, which could impact the appropriate assessment level.

There are several factors that could impact the level of assessment chosen for a product or service. Examples include if a product provides a new service or services a new geographic region for the company, a recent merger/acquisition or joint venture/partnership impacts the product, or other factors that might increase the likelihood, scope, or severity of content- and conduct-related risks.

The specific impact and magnitude of these events can vary widely from company to company, and from product to product. If one or more of these circumstances or events apply, the individual company makes a risk-based determination as to whether an adjustment in the level of assessment is warranted.

Executing assessments effectively

After applying the tailoring framework to determine the appropriate assessment approach (L1, L2, or L3), the assessment itself is executed. The assessment that the company undertakes is in accordance with the scoping described above and is assessed for adherence to the five overarching DTSP Commitments with specific focus on the relevant company practices that underpin those commitment areas. The practices are then objectively evaluated across three key dimensions: people, process, and technology.

The assessment is designed to help organizations develop a deeper understanding of the implementation of selected DTSP practices to mitigate content- and conduct-related risks. The outcome of the assessment will help organizations better understand the current state of their capabilities and their dependencies with respect to people, processes, and technologies.

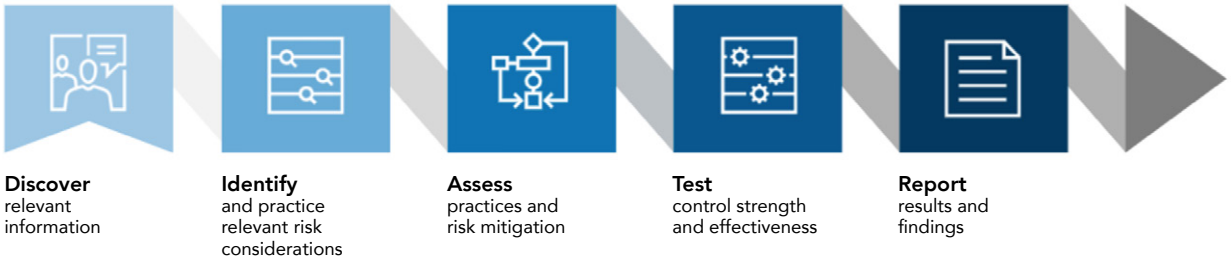
In instances where organizations are assessing their abilities to meet the five DTSP Commitments and have chosen a scope of assessment to evaluate commitments across multiple products or services, they will be able to gain additional insights into the relative maturity and effectiveness of these practices across the organization. The resulting understanding may help inform internal investment decisions and external engagements with policymakers, users, and the broader assurance community.

The Safe Framework is a tool that is designed to help organizations understand how the selected Trust & Safety practices are working and how they support their adherence to the five DTSP Commitments. As the Commitments are both technologically and content agnostic, not all practices will apply to all products. This provides flexibility for companies to select practices to manage their distinct content- and conduct-related challenges.

v. **Five step methodology**

There are five proposed stages or steps that make up the assessment process, from initial information gathering or discovery, to reporting of findings and results. The corresponding activities or procedures performed within each step will differ based on the selected level of depth for the assessment (L1, L2, or L3). For example, a Level 3 assessment may include detailed testing of the effectiveness of specific process controls (e.g., are target turnaround times for user complaint reviews being met?), while a Level 1 assessment may involve a higher-level review and understanding of processes.

DTSP assessment 5 step methodology



DTSP ASSESSMENT STEP DESCRIPTIONS

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

To provide more information about how this process applies across the DTSP Best Practices, we have included as an appendix the Question Bank that companies will use as a resource when performing initial information discovery.

Next Steps and Key Questions for Stakeholder Consultation

DTSP recognizes the urgency of the discussions occurring in homes, schools, and businesses around the world and at various levels of government on what digital Trust & Safety should look like. DTSP intends to share insights from implementing the Safe Framework, so that those discussions are informed by industry experience.

DTSP will synthesize results across companies into overall industry analysis that we will share with the public. We will cover the range of industry practices, providing observations and insights without identifying the individual companies. This will encourage companies to be transparent and continue to share best practices with the goal of evolving their own Trust & Safety practices over time. The state-of-the-industry report will not be an assessment of individual companies, but rather an assessment of how the industry is using various practices to mitigate and manage content- and conduct-related risks in a dynamic, ever evolving threat landscape.

Concurrently, DTSP will collaborate with industry assurance experts to develop an approach to support objective and measurable third-party assessments of our best practices framework.

Stakeholder consultation

DTSP is seeking input from outside parties on our own approach. As a first step toward stakeholder engagement, we invite comments on this paper and pose a series of questions based on issues that have arisen in our work.

This is a public consultation, but we particularly welcome comments from the following audiences:

- Civil society, including human rights and safety non-government organizations and consumer and user advocates;
- Governments, including policymakers from legislative and executive branches, as well as law enforcement; and
- Experts from academia and other sectors, including technical experts as well as those with experience in compliance and assessment and assurance.

Feedback should be sent to consultation@dtspartnership.org by February 15, 2022.

Contributions will be used to inform our approach as we iterate on the Safe Framework and implement internal and third-party assessments.

→
Key Questions

Key questions:

1. **Weighting commitments and practices:** DTSP is considering whether some Commitments or best practices should be given greater consideration than others when conducting assessments.

    1.1. Should each of the five DTSP Commitments be weighted equally?

    1.2. Are some of the best practices that support those commitments of greater importance than others?

    1.3. How should DTSP approach these differences as we refine and evolve the Safe Framework and work toward third-party assessments?

2. **How to provide meaningful transparency while mitigating safety risks:** Increasing transparency on Trust & Safety is a core objective for DTSP. We seek input from stakeholders on specific ways to increase meaningful transparency, beyond simply increasing the volume of information disclosed. In addition, complete transparency would provide malicious actors with information that could be used to avoid or subvert company policies and processes and potentially cause harm.

    2.1. How would the disclosure of assessment processes and results inform the work of external stakeholders working on Trust & Safety issues?

    2.2. How should DTSP complement the information already disclosed by member companies and other transparency initiatives in this space?

    2.3. How should DTSP manage the tension between transparency and safety?

    2.4. What forms of transparency and types of information should be provided to which audiences to maximize transparency while minimizing risks?

3. **How industry best practices relate to regulation around the world:** Some legislative initiatives have proposed audits for digital services, others envision industry-developed codes of conduct to guide company approaches.

    3.1. How can industry standards inform these efforts and potentially help avoid conflicting requirements that could pose risks to human rights and innovation and economic opportunity?

Appendix: Question Bank

Applicable Commitment	People, Process, Technology	Topic Area	Question
Commitment 1: Product Development	Process	Risk Identification and Assessment	How does your team evaluate and consider Trust & Safety risks during the product development lifecycle?
Commitment 1: Product Development	Process	User Experience	How do you balance product useability with security when considering the design of product features?
Commitment 1: Product Development	People	Roles and Responsibilities	Do you have a Trust & Safety team or individual involved in the product development process?
Commitment 1: Product Development	People, Process	Risk Identification and Assessment	How are Trust & Safety risks evaluated pre- and post-product launch? Is there a team accountable for this?
Commitment 1: Product Development	Technology	User Experience	How does your product allow users to control their own product experience as it relates to content? What sorts of technical measures (e.g., blocking or muting) are in place?
Commitment 1: Product Development	Process	Risk Identification and Assessment	How does your team perform or participate in risk assessments to better understand potential risks?
Commitment 1: Product Development	Process, Technology	Risk Identification and Assessment	What capabilities do you leverage to understand patterns of abuse prevalent on the platform, product, or service?
Commitment 1: Product Development	Process	User Feedback	How do you seek and incorporate user feedback related to Trust & Safety in the product design process?
Commitment 2: Product Governance	Process	Policies and Terms of Service and Guidelines	Are terms of service, policies, or applicable community guidelines made easily accessible to users?
Commitment 2: Product Governance	Process	Policies and Terms of Service and Guidelines	What is the frequency at which terms of service, policy updates, and community guidelines are communicated to users? By what means are these communicated to users?

Applicable Commitment	People, Process, Technology	Topic Area	Question
Commitment 2: Product Governance	Process	User Feedback	Do you have processes for taking user considerations into account when drafting and updating relevant Product Governance, such as policies, terms of service, or community guidelines?
Commitment 2: Product Governance	Process	Policies and Terms of Service and Guidelines	How do you document the interpretation and practical application of policy rules based on precedent, or other forms of research and analysis?
Commitment 2: Product Governance	People, Process	User Feedback	Do you have any forms of community-led self-regulation (e.g. forums for governance or tools for community moderation)?
Commitment 2: Product Governance	People	Policies and Terms of Service and Guidelines	Which team(s) is/are involved in updating or writing the product's content, conduct, and/or acceptable use policies?
Commitment 2: Product Governance	Process	Feedback and External Collaboration	Do you work with industry groups, third-party civil society groups, and/or external experts to solicit input on product policies?
Commitment 3: Product Enforcement	Process, Technology	Detection by Users	Are users able to report/flag content, conduct, or a user account as potentially violating policy? If so, please describe the process.
Commitment 3: Product Enforcement	Process	Review Processes and Procedures	What is the process for reviewing content that has been identified or flagged as potentially violating policy?
Commitment 3: Product Enforcement	Process	Review Processes and Procedures	How are content reviews prioritized, and what factors are taken into consideration?
Commitment 3: Product Enforcement	Technology	Review Processes and Procedures	What types of tools/systems are used to review content or manage the review process?
Commitment 3: Product Enforcement	Process	Enforcement Actions	What types of actions may be taken against a piece of content or user in relation to policy violations?

Applicable Commitment	People, Process, Technology	Topic Area	Question
Commitment 3: Product Enforcement	Technology	Enforcement Actions	What tools/systems are used to enforce content policies or manage the enforcement process?
Commitment 3: Product Enforcement	Process, Technology	User Notifications	How are users notified of enforcement actions taken relating to their content or activity on the product/service (e.g., broad public notices, icons)?
Commitment 3: Product Enforcement	Process, Technology	Detection Mechanisms and Processes	What types of processes or mechanisms are in place to proactively detect potentially violating content or conduct (automated or manual)?
Commitment 3: Product Enforcement	Technology	Detection Mechanisms and Processes	How do you make decisions about provisioning technology to conduct enforcement operations? How do you determine whether to build, buy, adapt, or collaborate when assessing available tools or technologies?
Commitment 3: Product Enforcement	Process	User Recourse	Is there a mechanism available for users to appeal decisions or actions taken on the product or service? If so, please describe the process.
Commitment 3: Product Enforcement	Process, Technology	Data Management and Retention	How are data related to enforcement actions (such as data relevant for investigations or key contextual data) retained and managed?
Commitment 3: Product Enforcement	People	Training and Awareness	How do you invest in reviewer wellness and awareness? What types of training programs and benefits are available to team members?
Commitment 3: Product Enforcement	Process	Detection by Third Party Partners	If applicable, how do you collaborate or partner with third parties to identify and flag potentially violating content or conduct?
Commitment 3: Product Enforcement	Process, Technology	Detection Mechanisms and Processes	How does your team protect against coordinated dissemination of illegal or violating content (e.g. public health misinformation, content harmful to minors, electoral processes) through automated or manual means?

Applicable Commitment	People, Process, Technology	Topic Area	Question
Commitment 3: Product Enforcement	Process, Technology	Detection Mechanisms and Processes	How does your team protect against the amplification of harmful content or conduct? What processes and systems are in place to deter bad actors and behaviors that violate product policies?
Commitment 3: Product Enforcement	Process	Feedback and External Collaboration	How and when are notifications and/or appropriate reporting sent outside the company, such as to law enforcement, in cases of credible and imminent threat to life?
Commitment 4: Product Improvement	Process, Technology	Process Quality and Continuous Improvement	Please describe current assessment methods for evaluating accuracy and effectiveness of content-related policies and/or operations.
Commitment 4: Product Improvement	Process	Risk Identification and Assessment	How often do you conduct risk assessments and how are emerging threats or risks taken into account?
Commitment 4: Product Improvement	Process	Risk Identification and Assessment	What are some of the key risk areas or focus areas that are top-of-mind as it relates to user Trust & Safety?
Commitment 4: Product Improvement	Process, People	Risk Identification and Assessment	Please describe if and how you use risk assessments to determine allocation of resources for emerging content- and conduct-related risks.
Commitment 4: Product Improvement	Process, People	Process Quality and Continuous Improvement	Please describe any existing methods for internal product feedback and evaluation, as it relates to mitigating content- and conduct-related risks.
Commitment 4: Product Improvement	Process	User Feedback	How do you seek and incorporate user feedback in the company's approach and processes to protect users?
Commitment 4: Product Improvement	Process	Feedback and External Collaboration	Please describe how you work with recognized third party civil society groups and experts (e.g. qualified fact checkers or human rights groups) to help evolve efforts to mitigate content- and conduct-related risks.
Commitment 4: Product Improvement	Process	User Recourse	Please describe any remedy mechanisms in place for users that have been directly affected by moderation decisions. (i.e. content removal, account suspension or termination).

Applicable Commitment	People, Process, Technology	Topic Area	Question
Commitment 5: Product Transparency	People	Transparency Reporting	If applicable, how is your team involved in developing or providing input into company transparency reporting or content risk reporting?
Commitment 5: Product Transparency	Technology	Transparency Reporting	How frequently, and via what means (e.g., publicly available website), are transparency reports made available to the public and other external stakeholders?
Commitment 5: Product Transparency	Process	Transparency Reporting	Please describe at a high level metrics or data retained for the purposes of regular transparency reporting (e.g. abuses reported, processed, data requests processed and fulfilled).
Commitment 5: Product Transparency	Process, Technology	Data Management and Retention	Do you have a process in place to log user complaints, decisions, and enforcement actions in accordance with relevant data policies?
Commitment 5: Product Transparency	Process, Technology	User Notifications	How and when are notices provided 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)?
Commitment 5: Product Transparency	Process	Feedback and External Collaboration	How do you collaborate with academic and other researchers working on relevant Trust & Safety subject matter (to the extent permitted by law, security and privacy standards, and other business considerations)? Do you share data and/or insights on a regular basis?