Digital Trust
& Safety Partnership

Trust & Safety Best Practices Framework

Introduction

Digital services are increasingly central to our daily lives, facilitating social discourse, economic activity, and much more. These services provide powerful tools for users across the globe to engage in a wide range of valuable online activity. But like any tool, they can also be misused to facilitate harmful behavior and content. Awareness of and action against this misuse has grown in recent years, which has led to increasing urgency in understanding, supporting, and evaluating effective ways to reduce harms associated with online content and behavior, while also protecting people's ability to express themselves, carry out business, access information, associate, work, study, and participate in their communities through digital services.

Striking this balance presents a considerable challenge. To begin, there is no one-size-fits-all approach to handling online content and associated behavioral risks or, more generally, to companies' Trust & Safety operations. Depending on the nature of the digital service, each may face unique risks relative to the various products or features they provide — different threats, different vulnerabilities, and different consequences. Products or features may engage with end users directly or indirectly, as well as with other services or businesses. What is an effective practice for one digital service may not suit another, and highly prescriptive or rigid approaches to defining Trust & Safety practices are likely to be too broad, too narrow or have negative unintended consequences. Further, risks change over time and so approaches to mitigating them must also have room to evolve.

Given the diversity of digital services, it is important to define an overall framework and set of aims for what constitutes a responsible approach, to which digital services can then map their specific practices. This flexible approach has been deployed in other domains, such as cybersecurity, yet existing frameworks are not sufficiently concrete to be applied when it comes to addressing harmful behavior and content online.

The Digital Trust & Safety Partnership (DTSP) aims to fill this need by documenting and facilitating the adoption of widely deployed, overarching Commitments set out in this document to foster greater transparency and better understanding of Trust & Safety both within and outside the industry. To that end, the DTSP Best Practices Framework offers a common framework to how companies address Content- and Conduct-Related Risks. While the overarching Commitments are uniform, the method by which they are fulfilled — whether by application of the illustrative practices in this document or alternatives — will vary by digital product or feature and evolve with both the challenges faced and advances made in the field of Trust & Safety.

DTSP regards the five overarching Commitments as representing the necessary steps taken by Practicing Companies to identify and address harmful content and conduct while preserving free expression and other rights, including internationally recognized human rights standards, as well as the social and economic value of digital services.

The DTSP Best Practices Framework marks the first ever attempt to articulate current industry efforts to address online Content- and Conduct-Related Risks. The Practices Framework focuses on considerations in the development, governance, enforcement, improvement, and clear documentation of digital products and services. Over time this Framework will evolve as they grow in maturity and can be assessed in more standard ways, such as in other disciplines like security and privacy.

General Commitment: Account for content- and conduct-related risk in the domains of product development, governance, enforcement, and improvement, and assign responsibilities and resources in each domain.

Practicing Companies take matters of Trust & Safety seriously, shown through investment in and development of relevant personnel and technology; adoption of rights-respecting Trust & Safety principles and considerations in the development, governance, enforcement, and improvement of products; and the clear documentation of digital products and services. The following Commitments characterize Practicing Companies' approach in each of these areas.

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

Aim: Ensure that companies engage in adequate forethought related to Content- and Conduct-Related Risks, and incorporate insights into product features accordingly.

Commentary: Anticipating and reducing Risk as part of product development is an important part of the Trust & Safety function. As product teams conceive, iterate, and refine products or features for launch, products are evaluated from the perspective of addressing potential Content- and Conduct-Related Risks. To this end, a range of mechanisms exist through which a given product or feature may be shaped to ensure it addresses Trust & Safety considerations. Practicing Companies (a) create processes to evaluate Content- and Conduct-Related Risks when developing products for release to the public, (b) seek to prevent or mitigate those Risks at the development stage, and (c) continue to evolve the product post-launch appropriately in light of observed Risks.

Examples of practices embodying a commitment to evaluate and adjust for Content- and Conduct-Related Risks in product development may include:

- Develop insight and analysis capabilities to understand patterns of abuse and identify preventive mitigations that can be integrated into products
- 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
- Designate a team or manager as accountable for integrating Trust & Safety feedback
- Evaluate Trust & Safety considerations of product features balancing useability and the ability to resist abuse
- Use in-house or third-party teams to conduct risk assessments to better understand potential Risks
- Provide for ongoing pre-launch feedback related to Trust & Safety considerations
- 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

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

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

Aim: Ensure the rules and principles governing user content and conduct are clear, rigorous, and consistent.

Commentary: Product Governance includes external and internal rules and processes by which a company fosters certain activities and discourages others in relation to its product(s). This function exists in addition to compliance with and mitigation of risk related to applicable laws. One embodiment of Product Governance is a company's terms of service (and for multi-product companies, sometimes multiple terms) — the contract between users and the company that sets forth underlying expectations and boundaries. Additionally, some companies may maintain additional rules that more directly address acceptable conduct, often in more plain and concrete language. These may be called rules, community guidelines, acceptable use policies, or content policies. Their drafting and evolution may draw on user communities, or a combination of stakeholders with varied relationships to the company.

Examples of practices embodying a commitment to adopt explainable processes for Product Governance may include:

- Establish a team or function that develops, maintains, and updates the company's corpus of content, conduct, and/or acceptable use policies
- Institute processes for taking user considerations into account when drafting and updating relevant Product Governance
- Develop user-facing policy descriptions and explanations in easy-to-understand language
- Create mechanisms to incorporate user community input and user research into policy rules
- Work with recognized third-party civil society groups and experts for input on policies
- Document for internal use the interpretation of policy rules and their application based on precedent or other forms of investigation, research, and analysis
- 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

Commitment 3: Conduct enforcement operations to implement product governance.

Aim: Ensure operations exist to implement the aims set forth in Product Governance to address Content- and Conduct-Related Risks.

Commentary: Companies take a variety of approaches to Product Governance enforcement, because each instance depends on the nature of the digital services provided and reflects the interactions among a particular user community. Nevertheless, some high-level commonalities exist. Companies must define the role of the enforcement function within the company (in relation to functions such as product, legal, communications, business, executive, and public policy) and within the team (with functional roles such as operations, policy, and reviewers). Companies frequently invest in a range of technologies and personnel to do tasks including: proactively detect content that violates rules, allow people to report violative content or conduct, develop systems for managing information from incoming reports, establish queues and processes for workers to make decisions and implement them, and systems of enforcement to deter bad actors or further violating behavior.

Examples of practices embodying a commitment to conduct enforcement operations to implement Product Governance may include:

- Ensure the company has personnel and technological infrastructure to manage Content- and Conduct-Related Risks, to which end the company may:
  - Constitute roles and/or teams within the company accountable for policy creation, evaluation, implementation, and operations
  - Develop and review operational infrastructure facilitating the sorting of reports of violations and escalation paths for more complex issues
  - Determine how technology tools related to Trust & Safety will be provisioned (i.e., build, buy, adapt, collaborate)
- Formalize training and awareness programs to keep pace with dynamic online content and related issues, to inform the design of associated solutions
- 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
- Where feasible and appropriate, identify areas where advance detection, and potentially intervention, is warranted
- 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)
- 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

  - Appeals of decisions or other appropriate access to remedy are available
  - Appropriate reporting is done outside the company, such as to law enforcement, in cases of credible and imminent threat to life
- 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
- Work with recognized third parties (such as qualified fact checkers or human rights groups) to identify meaningful enforcement responses
- Work with industry partners and others to share useful information about Risks, where consistent with legal obligations and security best practices

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

Aim: Ensure that mechanisms exist within the company to keep up with and respond to evolving Content- and Conduct-Related Risks and approaches.

Commentary: Practicing Companies will assess the effectiveness of their work preventing and mitigating Risk, apply an explainable framework to analyze information yielded from that assessment, and improve their processes to keep up with the evolution of these Risks and other relevant developments in the field.

Practicing Companies adopt methods to gather feedback on the effectiveness of their approach to mitigating Content- and Conduct-Related Risks and then evolve their approach to keep up with lessons from experience, developments in the product, and trends in the field.

Examples of practices embodying a commitment to regularly assess and improve processes associated with Content- and Conduct-Related Risks may include:

- Develop assessment methods to evaluate policies and operations for accuracy, changing user practices, emerging harms, effectiveness and process improvement
- Establish processes to ensure policies and operations align with these Commitments
- Use risk assessments to determine allocation of resources for emerging Content- and Conduct-Related Risks
- 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
- Establish appropriate remedy mechanisms for users that have been directly affected by moderation decisions such as content removal, account suspension or termination

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.

Aim: Ensure the public and other stakeholders have insight into the company's Trust & Safety goals, challenges and activities.

Commentary: Transparency serves a key function in informing the public and educating various stakeholders about a Practicing Company's Trust & Safety practices, while also building trust over time in the sufficiency of an industry's standard of care.

Examples of practices embodying a commitment to publishing and reporting on relevant Trust & Safety policies may include:

- 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
- 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)
- Log incoming complaints, decisions, and enforcement actions in accordance with relevant data policies
- 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)
- 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

Definitions

For purposes of the DTSP Best Practices Framework and the accompanying Commentary, the following definitions apply:

Best Practices Framework: refers to this document.

Commitment: For purposes of this Best Practices Framework, the actions committed to by Practicing Companies to identify and address Content- and Conduct-Related Risk.

Practicing Companies: Providers of products or services that have adopted the Commitments described herein.

Product Governance: refers to the set of agreements, rules, and guidelines mediating user interaction with the digital service and structuring conduct related to the product (examples include terms of service, privacy policy, community guidelines, content policy, acceptable use policy, codes of conduct, and any company processes by which these governing statements are created, adopted, or iterated).

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, and 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. (References to "Risks" shall be understood to refer to Content- and Conduct-Related Risks.)