BEST PRACTICES OVERVIEW

TRUST & SAFETY

This chart provides an overview of key trust and safety operational safeguards to consider when launching a service. The overview covers (1) content moderation; (2) safety by design; (3) security controls; (4) law enforcement requests; (5) reporting; and (6) AI safety.

Each item is assigned a priority level. This does not purport to cover every potential best practice but rather priority best practices.

Please consult your Ashurst Perkins Coie privacy lawyer for more information.

CONTENT MODERATION

	Best Practice	Description
■	Use policies	Publish publicly available policies that explain prohibited behavior and content.
■	Training materials	Develop and maintain operational guidance on content moderation and enforcement.
■	User reporting tool	Provide a way for users to report content that violates the terms of service or local law.
■	User notice process	Establish a process for notifying users of moderation decisions that affect their content.
■	Appeals process	Set up procedures to review challenges to moderation decisions.
■	Automated moderation	Decide which automated moderation systems to use and for what purpose.
■	Enforcement guidelines	Develop procedures for reviewing reported or detected content, including operational steps for violations and nonviolations.

■ Address immediately ■ Build near-to-medium term ■ Consider longer term

ASHURST PERKINS COIE

CONTENT MODERATION CONTINUED

	Investigations capability	Develop the ability to assess patterns of abuse (*e.g.,* scams) on the service beyond standard content review.
	Geo-blocking	Restrict content in specific markets where content is illegal.
	Third-party/industry collaboration	Work with trusted third parties (*e.g.,* human rights groups), industry peers, and users to share best practices, gather feedback, and refine policies.
	Keyword lists	Maintain lists of key terms that may indicate violations and inform automated tools.
	Audit	Track moderation metrics (*e.g.,* report volume, response time) and regularly sample decisions for consistency and accuracy, including through peer reviews or statistical methods.

SAFETY BY DESIGN

	Best Practice	Description
	Risk assessment	Identify and anticipate potential violations and abuses throughout the development process.
	Risk mitigation	Employ mitigations in design and architecture.
	Evaluation	Evaluate effectiveness of risk management.

SECURITY CONTROLS

	Best Practice	Description
	Access controls	Institute controls that limit access to content moderation tools. Secure content removed as appropriate and consistent with legal obligations, where applicable.
	Moderation logs	Develop a system that enables the moderation team to track enforcement data.

Address immediately Build near-to-medium term Consider longer term

LAW ENFORCEMENT REQUESTS

	Best Practice	Description
■	Data retention/ deletion	Establish data retention and deletion policies in line with relevant legal requirements.
■	Government requests	Create process for reviewing requests for user information or content takedown from law enforcement or governmental authorities.
■	Emergency disclosure	Establish an urgent disclosure process when law enforcement reports an imminent threat of death or serious injury.
■	Preservation	Build process and capability to preserve user data as required by legal requests.
■	Data production	Develop procedures for gathering user data in response to legal requests for information.
■	User notice	Establish policy on whether and how to notify users of legal requests for information.

REPORTING

	Best Practice	Description
■	External referral	Define when and how to report threats to life or safety directly to external bodies, including law enforcement.
■	Transparency reports	Prepare for transparency reporting requirements (*e.g.,* content moderation, law enforcement requests, trust and safety practices).

AI SAFETY

	Best Practice	Description
■	Access controls	Limit access to generative AI models for testing.
■	Model governance	Adopt safety-focused policies and processes to manage risk (*e.g.,* acceptable use, input/output controls, model parameters).
■	Content provenance	Record content creation process, including by watermarking or securing credentials, to help identify fraud and other abuse.
■	Data quality	Ensure datasets used to train models are diverse, relevant to their intended use, and tested for bias.

■ Address immediately ■ Build near-to-medium term ■ Consider longer term

FOR MORE INFORMATION

Contact us at PrivacySolutions@ashurstperkins.com

ASHURST PERKINS COIE