



JOB DESCRIPTION TEMPLATE

Trust and Safety Manager

Owns the policies, processes, and enforcement that keep a platform safe, from content moderation to abuse response and regulatory obligations.

Department	Trust and Safety / Policy / Operations
Reports to	Head of Trust and Safety / VP Operations / General Counsel
Location	Remote / Hybrid / On-site
Employment type	Full-time
Experience	6 to 10 years

POSITION OVERVIEW

The Trust and Safety Manager leads the organization's efforts to keep its platform safe, trustworthy, and compliant. This role owns content and conduct policy, moderation operations, platform integrity, and the response to abuse, harmful content, and regulatory obligations.

Working across policy, operations, engineering, legal, and increasingly AI teams, the Trust and Safety Manager sets enforcement standards, oversees moderation, and builds the tooling and metrics that scale safety. As platforms adopt AI, the role also helps ensure automated moderation and recommendation systems are used responsibly.

This is a hands-on leadership role for someone who can balance user safety, free expression, legal duty, and operational reality under real-world pressure.

KEY RESPONSIBILITIES

POLICY DEVELOPMENT

- Develop and maintain content, conduct, and community policies.
- Define enforcement standards and escalation thresholds.
- Balance user safety, expression, and legal obligations.
- Keep policies current with emerging harms and regulation.

CONTENT MODERATION AND ENFORCEMENT

Oversee moderation operations across human review and automated systems. Ensure decisions are consistent, well documented, and appealable.

- Manage moderation workflows and quality assurance.
- Set severity tiers and response timelines.
- Oversee appeals and enforcement review.
- Support moderator wellbeing and training.

PLATFORM INTEGRITY AND ABUSE

Detect and disrupt fraud, spam, coordinated abuse, and manipulation. Partner with engineering and data teams to build detection and response at scale.

REGULATORY AND REPORTING OBLIGATIONS

- Support compliance with platform regulation such as the Digital Services Act.
- Coordinate transparency and regulatory reporting.
- Manage law enforcement and legal escalations per procedure.
- Track obligations across the jurisdictions the platform serves.

RESPONSIBLE AI IN MODERATION

Help ensure automated moderation, classification, and recommendation systems are governed responsibly, with attention to accuracy, bias, human oversight, and transparency.

METRICS AND CONTINUOUS IMPROVEMENT

Define safety metrics, monitor enforcement quality and prevalence of harm, and drive improvements to policy, tooling, and process.

REQUIRED QUALIFICATIONS

- Bachelor's degree in a relevant discipline, or equivalent experience.
- 6 to 10 years of experience in trust and safety, content policy, platform operations, or a related field.

- Strong understanding of content moderation, platform integrity, and online harms.
- Familiarity with platform regulation such as the Digital Services Act.
- Experience leading operations, policy, or enforcement at scale.
- Sound judgment and resilience handling sensitive content.

PREFERRED CERTIFICATIONS

Relevant experience matters most. Backgrounds in policy, law, operations, or responsible AI are valued, and AI governance credentials such as AIGP are a plus for AI-heavy platforms.

TECHNICAL KNOWLEDGE

Content and conduct policy, content moderation operations, platform integrity and abuse detection, appeals and enforcement review, regulatory and transparency reporting, responsible AI in moderation, safety metrics, and moderation tooling.

ESSENTIAL COMPETENCIES

Sound judgment under ambiguity, clear communication, cross-functional leadership, operational rigor, resilience, and the ability to balance safety, expression, and legal duty.

SUCCESS MEASURES: FIRST 12 MONTHS

- Refresh core content and conduct policies.
- Strengthen moderation workflows and quality assurance.
- Establish severity tiers and response timelines.
- Stand up or improve the appeals process.
- Build platform integrity detection with engineering.
- Prepare transparency and regulatory reporting.
- Define safety metrics and a prevalence baseline.
- Set responsible AI guardrails for automated moderation.

This is a free, editable template. Replace bracketed fields with your specifics before posting.

Post this role, or find more templates, at ai-governance-jobs.com/templates/trust-and-safety-manager-job-description · A resource from **GRC Careers**