CONCEPT

Content Moderation and Trust and Safety: Policy, Review Queues and Appeals

28 June – 2 July 2027
Dubai - Five-star CBD venue



Content Moderation and Trust and Safety: Policy, Review Queues and Appeals

Ref: 506_16809 **Date:** 28 June – 2 July 2027 **Location:** Dubai - Five-star CBD venue **Fees:** 19500 SAR

Introduction:

User reports, marketplace listings, comments and live streams arrive faster than most review teams can handle consistently, so harmful material stays visible while legitimate posts are removed and appeals pile up. This Core Concept course builds content moderation and trust and safety operations skill: turning community guidelines into a harm taxonomy and enforcement ladder, reviewing text, image, video and live content to one standard, running human and automated review queues, checking decision quality and publishing transparency data. Participants produce a Moderation Policy and Operating Workflow for a case platform.

Course Objectives:

- Map the trust and safety functions, moderation models and harm categories that apply to a platform, marketplace, media site or e-service portal
- Write community guidelines and internal enforcement guidance with definitions, examples and a graduated enforcement ladder
- Review text, image, video and live content against policy and record consistent, reasoned decisions on context-dependent cases
- Design a human-plus-automation review workflow with classifier thresholds, queue priorities, escalation paths and user appeals
- Run quality assurance sampling, measure inter-rater agreement and track prevalence, time to action and appeal overturn rates
- Prepare a transparency report and a moderator wellbeing plan that protect users, reviewers and the organisation

Target Audience:

- Content review and moderation functions that assess reported and flagged user material day to day
- Trust and safety policy functions that draft community guidelines, harm definitions and enforcement guidance
- Community operations and user support functions that handle reports, notices and appeals from users
- Marketplace integrity and listing review functions that screen sellers, products and reviews
- Quality, workforce and operations functions that manage review queues, service levels and reviewer performance
- Digital service and portal management functions responsible for user-submitted content on public e-services

Course Outline:

Day 1: Trust and Safety Functions and Moderation Models

- Trust and Safety Function Map: Policy, Operations, Investigations, Tooling and Legal Liaison
- Moderation Models Compared: Pre-Moderation, Reactive Post-Moderation, Distributed and Automated Review
- Harm Taxonomy Build: Illegal, Harmful, Deceptive, Spam and Low-Quality Content Classes
- Platform Risk Profile Canvas for Social, Marketplace, Media Comment and E-Service Settings
- Current-State Moderation Diagnostic: Report Channels, Backlog, Tools and Decision Records

Day 2: Community Guidelines, Policy Writing and Accountability Principles

- Community Guidelines Drafting Method: Plain-Language Rules, Definitions and Allowed Versus Prohibited Examples
- Internal Enforcement Guidance Template with Decision Trees and Borderline Case Annotations
- Graduated Enforcement Ladder: Labels, Reduced Visibility, Removal, Feature Limits and Account Suspension
- Santa Clara Principles Applied: Numbers, Notice and Appeal Requirements
- Cultural and Language Competence Checklist for Multilingual Policy Enforcement

Day 3: Consistent Review of Text, Image, Video and Live Content

- Text Review Protocol: Slurs, Threats, Coded Language, Harassment Patterns and Counter-Speech
- Image and Video Review Technique: Frame Sampling, Blurring Settings and Graphic Content Handling
- Live Stream Intervention Playbook: Real-Time Signals, Interruption Options and Post-Stream Review
- Context Assessment Matrix: Newsworthiness, Satire, Education, Documentation and Intent
- Decision Rationale Log Linking Each Action to a Policy Clause and Evidence

Day 4: Automation, Escalation, Appeals and Quality Control

- Human-Plus-AI Workflow Design: Classifier Confidence Thresholds, Auto-Action Limits and Human Review Routing
- Queue Prioritisation Model by Severity, Reach and Imminent Risk with Service-Level Targets
- Escalation and Appeals Procedure: Uninvolved Reviewer, User Statement and Reasoned Outcome Notice
- Quality Assurance Sampling Plan and Inter-Rater Agreement Calibration Sessions
- Legal Takedown Request Intake: Notice Validation, Jurisdiction Checks and Response Records in General Terms

Day 5: Metrics, Transparency, Wellbeing and the Case Platform Workflow

- Moderation Metrics Dashboard: Prevalence, Proactive Detection Rate, Time to Action and Appeal Overturn Rate
- Transparency Report Structure: Actioned Content by Violation Type, Appeals, Automation Accuracy and External Requests
- Moderator Wellbeing and Resilience Plan: Exposure Limits, Rotation, Debriefing and Support Access
- Case Platform Brief: A Growing Community Marketplace with Mixed Media and Live Features
- Moderation Policy and Operating Workflow Presentation and Peer Review Panel

Skills You Will Gain:

- Harm Taxonomy Design
- Community Guidelines Writing
- Enforcement Ladder Design
- Multimedia Content Assessment
- Moderation Workflow Automation
- Appeals Case Handling
- Reviewer Calibration and QA
- Transparency Reporting

Why Attend This Course:

- Take back a Moderation Policy and Operating Workflow that a review team can put into use on its own platform or portal
- Reduce inconsistent decisions and overturned appeals by giving reviewers shared definitions, examples and calibration routines
- Explain moderation performance to leadership and users with credible metrics and a structured transparency report
- Compare review practice with peers from social, marketplace, media and public e-service teams

Conclusion:

Moderation quality depends on clear rules, consistent reviewers and systems that route the right case to the right person at the right time. The week moves from trust and safety functions and harm taxonomies, through community guidelines, enforcement ladders and the Santa Clara Principles, to reviewing text, image, video and live content in context, then to automation, queues, appeals, quality sampling and takedown requests. The final day covers metrics, transparency reporting and moderator wellbeing, and closes with each participant presenting a Moderation Policy and Operating Workflow for a case platform.