Knowledge

Online harms impacts on users ⟷ Legal and regulatory frameworks

Research

Define harms
Classify content
Corporate policies

Considerations

Severity | Impact | Context

INDIVIDUAL | AT-RISK GROUPS | SOCIETY

Balance

Thresholds for action

Legality | Legitimacy | Necessity proportionality

Intelligence gathering

Detection and evidence

- Classifiers, classification, recognition
- Detection monitoring remediation tools
- Signals
- Legal or Regulatory notices
- External inputs or advice
- Human moderation and user-based moderation
- AI modelling systems, Software-based chips
- Complaints, reports, flags

Thresholds for error

Evaluation | Monitoring and updating

Execution

Types of action

- Limit access to users
- Counter messaging
- Restrict access
- Report to official bodies
- Provide users with controls
- Removal of content or account
- Provide warnings to users
- Reward positive behaviour/interactions

Transparency

Audits | Reasons for decisions | Public reports | Feedback loops

Oversight | Accountability