Trusted Online Content Moderation

Mike Matthys

mikem@4betterinternet.com

Trusted Online Content Moderation

- **Search & social media platforms are not widely trusted for safety & neutrality**
  - Government officials doubt efforts for safe & neutral content moderation; Surveys show users lack trust in platforms
  - Opaque enforcement actions with unclear explanations to affected users will continue to drive user distrust in platform

- **Proposed Solution: Mandate transparency on all aspects of content moderation**
  - Report on each enforcement action: user, specific rule broken, specific content category, use of named fact-checker, govt request
  - Published enforcement action reports can be sorted, tabulated, analyzed to measure online safety and viewpoint neutrality
  - Any communications with govt or entities/contractors funded by govt (except specific law enforcement /natl security actions)

- **Glare of publicity will help ensure consistent application of content moderation enforcement actions for both online safety and viewpoint neutrality**

IBI
TRUST & SAFETY

Why Focus on Transparency?

- **Peer-to-peer platform comparisons will improve content moderation performance**
  - Published reports will enable peer-to-peer comparison of platforms – like reports to SEC on financial performance
  - Observers, critics, media and academics can study and measure content moderation performance & results

- **Ensure content moderation requests by govt & govt-funded entities are appropriate / visible**
  - Specific national security and law enforcement actions would be excepted

- **Expanded transparency is workable for companies to implement**
  - Large companies already gather, report and publish subsets of relevant generalized information and statistics

- **Broader support for transparency than for Section 230 changes in divided Congress**
  - Sec230 change requires difficult agreement on standards for online safety, viewpoint neutrality, role of govt, liabilities

IBI
TRUST & SAFETY

Ensuring Full Transparency

- **Publish reports on all enforcement actions taken**
  - Specific content categories affected, type of action (including de-amplify), content rules broken, repeat offenders
  - Disputes and appeals of dispute resolutions, role of specific 3rd-party fact checkers in enforcement actions
  - Demonetization or non-promotion of third-party content sites
- **For any enforcement action, provide clear explanation & justification to user**
  - Specific content and specific rules broken, explanation of review/adjudication process, steps to appeal
- **Publish criteria and ranking of websites as authoritative sources for search results**
- **Report relevant company communications to/from govt & govt-funded entities**
  - Whistleblowers protected and rewarded if companies are purposely evading public transparency
  - Actual national security and law enforcement actions are excepted
- **Identify and publish affiliations, funding & history of fact-checkers**

IBI
TRUST & SAFETY

Transparency ≠ Divulging Tech Trade Secrets & IP

- **Majority of enterprise value is their proprietary algorithms and software**
  - Output/results are regulated in other industries rather than “how it works” trade secrets
  - Academic researchers cannot be prevented from eventually joining/assisting competitor companies

- **Monitoring and regulating content moderation does not require knowledge of proprietary algorithms and software**
  - Online safety and consistency in enforcement actions can be tested and monitored

- **Echo chambers & features of their proprietary algorithms become less important if online safety & consistency of enforcement is transparently monitored**

IBI's Proposal:

Mandate transparency, with financial penalties to ensure accountability

- **Require all platforms to publish mandated transparency data and reports quarterly**
  - All content rules, enforcement actions, user dispute resolutions, whistleblower protections, fact-checkers
  - All website ranking categories for use as authoritative sources for search results
  - All communications with govt or govt-funded entities (except natl security & law enforcement actions)

- **For each enforcement action, provide clear explanation/justification to user**
  - Specific content and specific rule(s) broken, explanation of appeals process

- **Define criminal/financial penalties to ensure accountability for transparency**
  - Federal criminal actions can be brought by federal DOJ or appropriate regulatory agency (FCC or FTC)
  - Audits/discovery triggered by civil or regulatory enforcement actions
  - Protection and reward for whistleblowers (employees or contractors of the online companies)
  - Financial penalties TBD based on intent, repeat of offenses, and harm caused by missing or incorrect enforcement actions

- **A transparency-only solution that relies on the glare of publicity to improve trust in content moderation for online safety and viewpoint neutrality**

Other Industries Provide Safety & Performance Statistics

- Automobiles are measured for safety, mileage, performance, resale value, etc
- Airlines are measured for on-time arrivals
- Banks are measured for innumerable financial and fairness measures
- Heavy industry measured for employee safety and environmental protections

When statistics are measured and reports are published – safety and performance improves