Trusted Online Content Moderation

Mike Matthys

mikem@betterinternet.us

Trusted Online Content Moderation

- **Current social media content moderation is not trusted**
  - People on both sides view content moderation as being biased and broken
- **Proposed Solution**
  - Define content moderation guard rails for each platform's independent rules
  - Certification of platforms that transparently publish content moderation rules
  - Outsource appeals of user/platform content disputes to trusted 3rd party
- **Benefits**
  - Reduce legal costs, PR issues, and costs of employees handling appeals of disputes
  - Reduce political will for antitrust break-up or heavy-handed federal/state regulations

IBI
TRUST & SAFETY

Guard Rails of Content Moderation

- **Safety**
  - Protect users from content that is criminal or imminently harmful
- **Neutrality**
  - Don’t favor one side on an issue – except to protect against imminent harm
- **Transparency**
  - Publish clear & detailed content moderation rules as well as enforcement actions
- **Accountability**
  - Appeals outsourced to independent non-govt entity with power to ensure platform compliance

Ensuring Safety

Criminal and imminently harmful content should be blocked

- **Criminal content**
  - Exhorting violence, criminal/terrorist planning, child porn, trafficking, extortion
- **Cyber-security threats & viruses**
- **Imminently harmful content**
  - Hateful/bullying speech targeting specific non-public persons, interference in election voting, doxxing
- **Disinformation from designated terrorist orgs & overseas governments**

IBI
TRUST & SAFETY

Ensuring Neutrality

- **Avoid taking sides on controversial issues**
  - Avoid role of true/false arbiter - except to protect against imminent harm to a person(s)
  - True/false inevitably breaks neutrality and perception of neutrality "Who decides?"

- **Ensure platform algorithms amplify or depress content on neutral basis**
  - Protect against viewpoint amplification/SPAM from bots, troll farms & inauthentic user networks

- **Ensure enforcement is consistent with published rules & applied fairly to all users**
  - Violations of safety rules enforced regardless of user popularity, viewpoint or political affiliation

- **Government officials can offer info, but inherently cannot act as neutral fact-checkers**

IBI
TRUST & SAFETY

Neutrality is not Uncontrolled Free Speech

- **Content moderation innovations within the Guard Rails are encouraged**
  - Content moderation can contribute to Quality of Service & Features
- **Some categories of content moderation allowed, even if not imminently harmful**
  - Adult content
  - Requirements for authenticate users
  - Spam and click-bait for purpose of commercial gain
- **Small platforms and user sub-groups may be publicly non-neutral**
  - Platforms smaller than threshold – must publish their non-neutral standards
- **Key is for any content moderation to be Safe, Neutral & Transparent**

IBI
TRUST & SAFETY

Ensuring Transparency

- **Publish content moderation standards & related enforcement actions**
  - Users & content creators should easily know the specific rules/standards they violated

- **For any enforcement action, provide clear explanation & justification to user**
  - Specific content rules broken, explanation of review/adjudication process, steps to appeal
  - Involvement and communications between platform and 3rd party fact-checkers

- **Identify and publish qualifications of 3rd party fact-checkers**
  - Publish their names, qualifications, affiliations, funding, and history of previous decisions

- **Report all communications to/from government employees/contractors within 24 hours**

IBI
TRUST & SAFETY

Ensuring Accountability

- **Platforms benefit when user appeals outsourced to trusted 3rd party**
  - 3rd party entity can ensure perception of neutrality/fairness between users and platforms

- **FINRA as a model for a non-govt trusted entity with enforcement powers**
  - FINRA, funded by financial industry, outsources appeals of investor/advisor disputes
  - Governance designed for neutrality/fairness; no hire/fire influence from govt or industry
  - Access to 8000 arbitration judges for handling scale volume of online appealed disputes
  - FINRA has power to levy material fines, suspend/ban financial advisors to ensure compliance

- **Proposed OMRA entity would operate similar to FINRA – for online platforms**
  - Platforms would need to accept OMRA judgements for legitimacy
  - Industry-wide buy-in ensures benefits to platform companies (PR, govt anti-trust, legal)
  - FINRA = Financial Industry Regulatory Authority OMRA = Online Media Regulatory Authority

IBI
TRUST & SAFETY

IBI's Proposal – Creation of 3rd Party Dispute Resolution Entity

- **Create a non-govt entity (modeled after FINRA) for 4 functions:**
  1. Outsource appeals of selected user/platform disputes
  2. Certification of each platform for transparency & for content rules within the Guard Rails
  3. Review data reports periodically received from companies
  4. Review/publish 24-hour reports of govt communications with companies

- **Industry-wide buy-in from the largest online media companies**
  - Agreement to adhere to 4 Guard Rails for content moderation
  - Each platform sets their own standards which operate within these principles
  - Realistic financial enforcement accepted by platforms to stay within Guard Rails

IBI
TRUST & SAFETY

FINRA as Non-Govt Model for Appeals

- **FINRA handles investor/advisor disputes for a portion of the financial industry**
  - $800M budget, 3000 employees, 8000 arbitration judges
  - FINRA has handed out fines to companies and suspended 100s of financial advisors/brokers/etc
  - Started from NASDAQ/NYSE with governance to avoid regulatory capture by the industry

- **Proposed OMRA (Online Media Regulatory Authority) will operate like FINRA**
  - As FINRA is loosely overseen by SEC, OMRA would be loosely overseen by the FCC or FTC
  - Balanced governance: No regulatory capture, no political appointees and no govt funding
  - Power to levy fines large enough to motivate the companies

- **OMRA governance carefully designed to ensure all stake-holders fairly represented**
  - Social media platforms (large & small), content creators, news publishers, right and left, user representatives

- **FINRA outsources investor appeals, OMRA outsources online user appeals**
  - Build user trust and off-load the costs and headaches of PR controversies and legal issues

IBI
TRUST & SAFETY

OMRA Will Reduce Litigation Costs

- **User EULAs modified to drive users to online arbitration with OMRA**
  - Users will have simple online mechanism to appeal their disputes to OMRA
- **OMRA structure will include two-tiers of appeals**
  - Accepted user disputes are resolved via online video arbitration with 1 arbitration judge
  - Accepted 2nd tier appeal is escalated to an online video arbitration with a panel of judges
- **OMRA precedents will be organized and searchable online**
  - User dispute categories, resolutions, opinions, enforcement actions, penalties if any
  - Platforms can point users to similar disputes previously resolved by OMRA
- **Clear published content moderation rules will reduce user disputes**
  - When users understand the specific rules violated, they're less likely to dispute the enforcement

IBI
TRUST & SAFETY

Dealing with Scale

- **"Certified" online platforms will have fewer & simpler appeals**
  - Accepted appeals would adjudicate based on whether platform followed its own published rules
  - Transparency and clear enforcement explanations to users will reduce appeals

- **OMRA will receive appeals only after companies provide initial dispute resolutions**
  - OMRA will rely on its precedents - and may or may-not accept each appeal based on its merits
  - Refundable nominal fee (e.g. $100) to discourage frivolous appeals by users

- **OMRA penalties incentivize platforms to improve algorithms & transparency**
  - Improve content moderation algorithms and transparency, neutrality rating of fact-checkers
  - Platform moderation algorithms & published standards will provide the heavy-lifting to handle scale

- **OMRA can be industry-funded at similar scale as FINRA (with far fewer employees)**
  - Cost efficient use of online video rather than F2F meetings for arbitrations

IBI
TRUST & SAFETY

Trusted 3rd Party for Dispute Resolution Provides Massive Political and PR Benefits

- **Reduce the push for anti-trust break-up in Congress**
  - Serious effort at ensuring neutral content moderation will move Republicans off anti-trust
- **Reduce chances of online monopolies being designated as Common Carriers**
  - Equal standards for all users will eliminate political driver for common carrier status
- **Dent the push for multiple state laws such as Texas HB-20 and Florida SB-7270**
  - The 3rd party entity + transparency achieves most goals of these proposed laws
- **Off-load PR & legal challenges of controversial topics and decisions**
  - Industry-wide support of 3rd party entity will shift focus away from companies

IBI
TRUST & SAFETY

Guard Rails & OMRA Also Apply to Online Payments

- **Online payment & site monetization are key components of free speech**

- **Examples of near-monopoly online payment platforms covered include:**
  - PayPal and Venmo
  - Credit Card brands/networks
  - GoFundMe and other donation sites
  - Ad-based hosting platforms
  - Mobile app payment gateways


Institute for a Better Internet
Additional Slides

True/False Standard Inevitably Breaks Trust

- **True/False choice eventually breaks trust from key stakeholders**
  - Whichever choice is made on controversial public topics, a large share of users will claim bias
- **Misinformation/Disinformation is disconnected from online safety**
  - Misinformation should only be restricted to prevent imminent harm/danger
- **Outsourcing to third-party fact-checkers does not solve the problem**
  - Fact-checkers have their own funders, biases and opinions
  - Government authorities deciding true/false is a well-trodden path to authoritarianism
- **Who Decides if it's true/false rather than just a disagreement ?**
  - Who decides the basis ? Who adjudicates ?

IBI
TRUST & SAFETY

A Better Standard: Imminently Harmful or Not

- **Harmful/Not Harmful is a test that supports both Safety and Neutrality**

- **Imminently Harmful/Not Harmful is easier to adjudicate than True/False**
  - Avoid issues of " Who Decides? " Avoid disagreements among subject-matter "experts"
  - Arbitration judges capable of assessing harm without needing subject expertise

- **Scalable across millions of contentious topics, disagreeable opinions, etc**
  - Avoid controversial disagreements between users and company-selected media fact-checkers

- **Should "misinformation" be moderated? Only if it leads to imminent harm**

IBI
TRUST & SAFETY