AI, Extreme Speech, and the Challenges of Online Content Moderation

Elonnai Hickok

AI4 Dignity

Multilingual and Multimodal Hate Speech Workshop

September 6 2021

AI in Content Moderation

- For purpose of our policy brief "AI Extreme Speech and the Challenges of Online Content Moderation", we worked with the following understandings:
  - AI in content moderation involves the deployment of machine learning models in automated actions and decisions around user generated and paid content with text, image, video, and/or audio.
  - Such actions can center around detecting content (flagging and labelling content), removing content (blocking, suspending, and removing content), and curating content (recommending content, making content more or less prominent, and ranking content).
  - In policy, the use of AI in content moderation is also referred to as 'automated content moderation' or 'use of automated tools.'

Challenges in AI and Content Moderation

Challenges in using AI for content moderation can include:

- Inability to fully account for evolving context, practice, linguistic diversity, and forms of content
- Opaque decisions
- Lack of notice
- Proactive moderation and censorship
- Shifting the burden of determining unlawful content
- Authoritarian use including targeting dissenting voices and minority communities
- Amplification of harmful content

Human Rights and AI in Content Moderation

Recognition that:

- The use of AI in content moderation can directly and indirectly impact human rights including privacy, freedom of expression including the freedom to seek, receive and impart information and the freedom to hold opinions without interference, and freedom of assembly.
- The misuse of automated content moderation systems can result in harm to individuals, communities, and societies and can result in the further marginalization of groups and communities as well as threaten democratic institutions and civic spaces.
- Recognition of the importance of a human rights based approach to the use of AI in content moderation including multi-stakeholder geographically representative engagement in the development of relevant norms, rules, and standards for the development, procurement, use, certification, and governance of AI systems.

*(*OHCHR Report on Artificial Intelligence technologies and implications for freedom of expression and the information environment and Freedom Online Coalition Statement on Human Rights and AI)

Emerging Policy Trends

The use of AI in content moderation is being recommended for different purposes including:

- Demoting the ranking of content that exposes users to problematic content.
- Prioritizing relevant, authentic, and accurate and authoritative information where appropriate in search, feeds, or other automatically ranked distribution channels.
- Identifying, reviewing, and reducing illegal content.
- Redirecting users from problematic content.
- Promoting counter narratives.

(Examples from the Australian Code of Practice on Disinformation and Misinformation, European Commission Code of Practice on Disinformation, and the Christchurch Call)

Policy Questions

The use of AI in content moderation raises a number of policy questions:

- For what purposes and in what instances should AI in content moderation be used? (public vs. private channels, flagging illegal content)
- In what configuration should AI be used in content moderation? I.e flagging content for human review? Removing content?
- Should the use of AI be mandated or voluntary?
- Should algorithms be certified before being deployed by a company? Are there standards that algorithms should be measured against to account for bias, diversity, and accuracy? Who/how should these standards be developed?
- How can oversight and accountability be built into the development and implementation of AI in content moderation?
- How can transparency be built into AI system used in content moderation?
- How can users be given control and choice over the use of AI in content moderation?
- How should liability be determined?
- What forms of remedy are needed?

Examples of Legal Frameworks for use of AI in Content Moderation

Legal frameworks for the use of automated content moderation are beginning to emerge:

- *UK Online Harms White Paper*: Vests responsibility in the regulator for determining use and safeguards.
- *Proposed Digital Services Act:* Focuses on notice, transparency, redress, risk assessment, and audits.
- *Indian Guidelines for Intermediaries and Digital Media Ethics Code Rules 2021*: Provides high-level framework for the use.

Examples of Legal Frameworks for use of AI in Content Moderation

These proposals have highlighted several important regulatory elements:

- **Mandatory Nature**: Frameworks are proposing different configurations as to if the use of the technology can be mandated or not.
  - UK Online Harms White Paper proposes that the regulator can require the use of automated tools and encourages companies to proactively use these tools to remove terrorist content and child exploitation and abuse. .
  - Indian Intermediary Guidelines states that companies should endeavour to proactively deploy these technologies to identify content depicting rape, child sexual abuse or conduct, whether explicit or implicit, or any information which is exactly identical in content to information that has previously been removed.
- **Principles to Guide Use:**
  - UK Online Harms White Paper recommends that application is guided by the persistence of illegal content and that no less intrusive means for identifying and removing the content is available.
  - Indian Intermediary Guidelines requires that the use of the tools must be proportionate with respect to privacy and free speech.

Examples of Legal Frameworks for use of AI in Content Moderation

Instances for Use and Human-AI Collaboration: Proposals have focused on using AI for identifying clearly illegal content.

- The UK Online Harms White Paper proposes using automated tools to identify illegal content with a focus on child exploitation and terrorist content on public and private channels. Recommends that automated tools can be used to identify, flag, block, or remove illegal content for human review.
- The Indian Guidelines recommend using automated tools to proactively to identify content depicting rape, child sexual abuse or conduct, and information that has already been identified as illegal. The rules do not clarify if the use should be on public or private channels.
- The DSA does not limit the use of automated tools, but does not require it. The DSA does specify that complaints related to content or accounts that has been suspended cannot be resolved using automated means.

Examples of Legal Frameworks for use of AI in Content Moderation

Notice: Different configurations for communicating the use of automated tools are beginning to emerge.

-UK Online Harms White paper recommends that the Regulator provides notice to the public when a decision is made to require a company to use automated tools.

-DSA requires that intermediaries must communicate if automated tools were used as part of the decision making process when content is removed and must clarify if automated tools were used to flag the content or remove the content and provide, the reasons, and possible redress mechanisms. DSA also requires that intermediaries communicate in their ToS the tools they use in content moderation.

-Indian Guidelines must display a notice to any user attempting to access such information stating that such information has been identified by the intermediary

Transparency and Reporting Requirements:Structures of transparency and reporting requirements accounting for the use of automated tools

-DSA requires that transparency reports include the use of automatic means for content moderation including precise purposes, indicators for accuracy, and any safeguards applied.

Examples of Legal Frameworks for use of AI in Content Moderation

Risk Assessment, Safeguards, and Review: Different structures for risk assessment, safeguards, and review are emerging for both automated content moderation and the algorithmic amplification of content.

-DSA very large online platforms to undertake a risk assessment that includes a review of how their content moderation systems, recommender systems, and systems for selecting and displaying advertisements influence any of the systemic risks identified. Allows for the Digital Coordinator to ask for more information on the accuracy of tools and data used.

-Indian Guidelines require a review of automated tools with regard to accuracy and fairness, propensity of bias and discrimination, and the impact on privacy and security. Intermediaries must also implement mechanisms for appropriate human oversight of such measures.

- UK Online Harms White Paper commits to form legal safeguards for the use of automated tools, requires the authority to take permission from Ministers that sufficiently accurate tools exist, produce a report on the effectiveness of these tools,

Observations

- Policy around the use of AI in content moderation is in very early stages. This is an opportunity to get it “right” and involve communities in the design of policy.
- There is a need to bring community centric human-AI collaboration into the center of policy around AI and content moderation. AI4Dignity is focused on the principle of “human in the loop”by developing a hybrid human machine process model with curated space of coding.
- Potential focal areas for policy and community centric human - AI collaboration going forward can include:
    - Enabling meaningful transparency as an important first step towards community centric human - AI collaboration.
    - Articulating a collaborative graded approach to human–AI collaboration, such as type of content, nature of content, and potential of harm, to determine the extent to which the AI augments a decision vs. takes
    - Defining standards and mechanisms for human review and oversight
    - Community collaboration to define policies that guide AI

Thank you!