Moderating Moderation: Distributing Power in Content Governance Regimes

Presentation of Deirdre K. Mulligan

Workshop on Section 230 Protections:

Can Legal Revisions or Novel Technologies Limit Online Misinformation and Abuse?

National Academies of Sciences, Engineering and Medicine's Committee on Science, Technology and Law

April 22, 2021

UC Berkeley School of Information

BERKELEY CENTER FOR LAW & TECHNOLOGY

CONTENT MODERATION

The Question

how to *moderate* the process of content moderation

Identifying Answers

Focus on the *function*

Deconstruct it

Examine configurations in existing content moderation regimes*

- who does what?
- what constraints?
- what concerns?
    - democratic deficits, competition

Content Moderation Frameworks*

- 230
- DMCA
- Right To Be Forgotten (GDPR)
- PROTECT Act
- Global Internet Forum to Counter Terrorism (GIFCT)

All explicitly and/or implicitly allocate functions and place constraints on them that shape platforms' "scripts" (Akrich 1992)

**Functional* description

Decomposing Content Moderation

Handoff Model is designed to explore the **Values** implications of different ways (configurations of humans and technical artifacts) of performing a function, particularly salient during automation.

Key premise:

different **Configurations** of **Content Moderation Function** (different Actors in different arrangements executing a function) *explicitly* and *implicitly* redistribute things in addition to **Function** with implications for **Values.**

Functions of Content Moderation

- Defining
- Identifying
- Locating
- Moderating

Platforms

Under those initial *allocations* and *constraints*, platforms further refine *How* to execute the functions

They *inscribe* responsibilities for content moderation on the heterogenous network of humans and nonhumans (sociotechnical system)

Constraints on Platforms' Scripts

Allocations of *Functions*

- Defining
- Identifying
- Locating
- Moderating

Constraints on *How* it is performed

- Processes
- Actors

Allocation of Functions?

Defining

- 230 Companies
- DMCA federal law but rights holders great influence in practice
- RTBF EU law but requestors great influence
- PROTECT federal law but companies/law enf./NCMEC
- GIFCT collaboration among companies

What the Functional Analysis helps assess

- What sub-functions might be most important for particular values
- Which constraints should attach to the delegation of sub-functions
    - What sub-functions might be more appropriate for delegating to particular human v. technical actors
- Where investments in shared content moderation infrastructure could support particular values

Shared Identification Infrastructures

Matching

- data bases containing hashes of content determined to be illegal or impermissible
  - can be cryptographic hashes (1-1 match) or perceptual hashing which identifies similarities between two pieces of content (homologies)

- Individual platforms can use this shared identification infrastructure to screen for illegal or otherwise objectionable content
  - Today frequently used to block uploads (critiqued as prior restraint) but can be used in other ways
- today such shared identification infrastructure is used for
  - child pornography, National Center for Missing and Exploited Children's hash database
  - extremist content, Global Internet Forum to Counter Terrorism Shared Industry Hash Database

Different allocations and constraints

National Center for Missing and Exploited Children’s hash database

- More stable public agreement on definition (and application)
    - no exceptions for sharing
- identification collaborative but non-profits run and curate database w/ law enf input
    - National Center for Missing and Exploited Children (NCMEC) & Internet Watch Foundation
- locating companies
- moderating companies
- Constraints?

Global Internet Forum to Counter Terrorism Shared Industry Hash Database

- Definitions murky (shared def, but company def vary, and unclear how this is dealt with)
    - Recognized need for exceptions
- identification, unclear who can upload hashes
- locating companies
- moderating companies
- Constraints
    - Initial statements emphasized that matches would not automatically result in removal; however recent statements indicate this is no longer true
    - Transparency reports
    - procedural rules

Constructive turn

Shared Identification Infrastructure for non-consensual pornography?

- Voluntary
- Store hashes only

Constructive Turn

Potential benefits of formalization

- provide site for values to be contested
  - “re-politicize governance & decision-making” (Katzenbach & Ulbricht, 2019)
  - reckoning around definitions & identification
  - reveal previously hidden assumptions & biases
- support standardization
- ease auditing

Constructive turn

Constraints to address systemic rather than Individual harms, examples:

- benchmarking against test sets,
- particular forms of reporting (audits of false positives and false negatives),
- Broader transparency reports as envisioned by the Santa Clara Principles
- particular forms of researcher/auditor access

Constructive turn

Constraints to empower users with greater power to tailor informational environments:

- Art 29 of the EU DAS
  - Provide parameters of recommender systems
  - Provide options for modifying
    - At least one not based on profiling
- Crowd sourced distributed controls (blocktogether)

Ad targeting algorithms and 230

Contributing to the development of what makes the content illegal

or

Conduct not content

Thanks

A Functional Approach to Content Moderation, Co-author Prof. Kenneth Bamberger, School of Law UCB, forthcoming Berkeley Technology Law Journal 2021

Related works

- Deirdre K. Mulligan & Helen Nissenbaum, *The Concept of Handoff as a Model for Ethical Analysis and Design*, Oxford Handbook of Ethics of Artificial Intelligence (Markus D. Dubber, Frank Pasquale & Sunit Das, eds., forthcoming Oxford University Press 2020)
- Datta, Amit, Anupam Datta, Jael Makagon, Deirdre K. Mulligan, and Michael Carl Tschantz. "Discrimination in online advertising: A multidisciplinary inquiry." In *Conference on Fairness, Accountability and Transparency*, pp. 20-34. PMLR, 2018.

Support

- NSF INSPIRE: Value-Function Handoffs in Human-Machine Compositions SES 1650589