Regulating & Auditing Social Media Algorithms

A Perspective Compatible with Section 230 and Free Speech Protections

Sarah H. Cen | MIT EECS | shcen@mit.edu

April 10, 2024

Challenges of regulating disinformation

Tackling disinformation is difficult for **legal** and **regulatory** reasons:

Legal challenges

Section 230: Platform can't be treated as "speaker" or "publisher"

We can still regulate the algorithm!

1st Amendment: Why not require that algorithm removes disinfo?

It's unconstitutional! (cf. "chilling effect")

Regulatory challenges

Measurement: It's often unclear what the law means in practice

How do we measure compliance?

Auditing: There must be ways to scalably audit for compliance

How do we monitor complex algorithms?

Our proposal

Legal

TL;DR Flexible standard that still allows for personalization

What we don't want:

- Strict global standard of "good"
- Too restrictive (e.g., removes main revenue or prevents personalization)

Our approach: Require that curated content is "similar" to flexible standard of "good." Allows personalization!

not what we propose

Contrasting example: Ask that all election-related content only come from whitelisted sources. This may be too restrictive and hurt content ecosystem!

Our approach: Require that personalization (i) with only <u>whitelisted</u> sources and (ii) with all sources are informationally "similar"

Our proposal

Regulatory

TL;DR Easy-to-run auditing procedure

Characteristics:

- Black-box audit (many benefits!)
- Privacy protecting
- Interpretable
    - Easy for regulators to tune
    - Easy to interpret results
- Strong theoretical guarantees
- Allows personalization & revenue

Why black-box audits?

- Minimal access (don't need access to proprietary algorithms and data)
- Model agnostic (don't need to design new audit for each model)
- Prospective (not limited to past data, can test unseen scenarios)

Summary

We propose a way to regulate & audit content algorithms that is:

- Compatible with Section 230 and 1st Amendment
- Practical, privacy-protecting, interpretable, and low-cost

Why? Without the ability to monitor & audit content algorithms, it's impossible to hold platforms accountable for content curation

Can it be used to counter disinformation? Yes, the audit can be used to curb disinformation and test for the relative impact of untrusted sources

Status of work: Conference paper is online (we're working on extended version of paper). We've also run a live audit (ask me about it!)