Accountability Infrastructure for Recommender Systems: Applying Lessons from Public Health to Mitigate Structural Harms

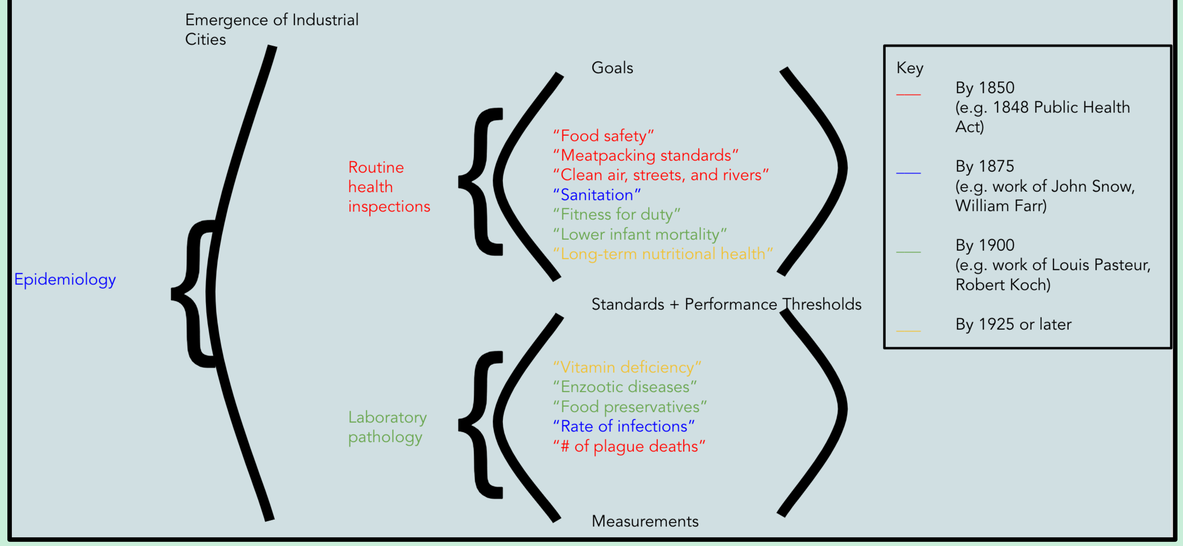
Nathaniel Lubin[1] and Thomas Krendl Gilbert[2]

[1]Digital Life Initiative, Cornell Tech and Berkman Klein Center, Harvard University [2]Digital Life Initiative, Cornell Tech and Center for Human-Compatible AI, UC Berkeley

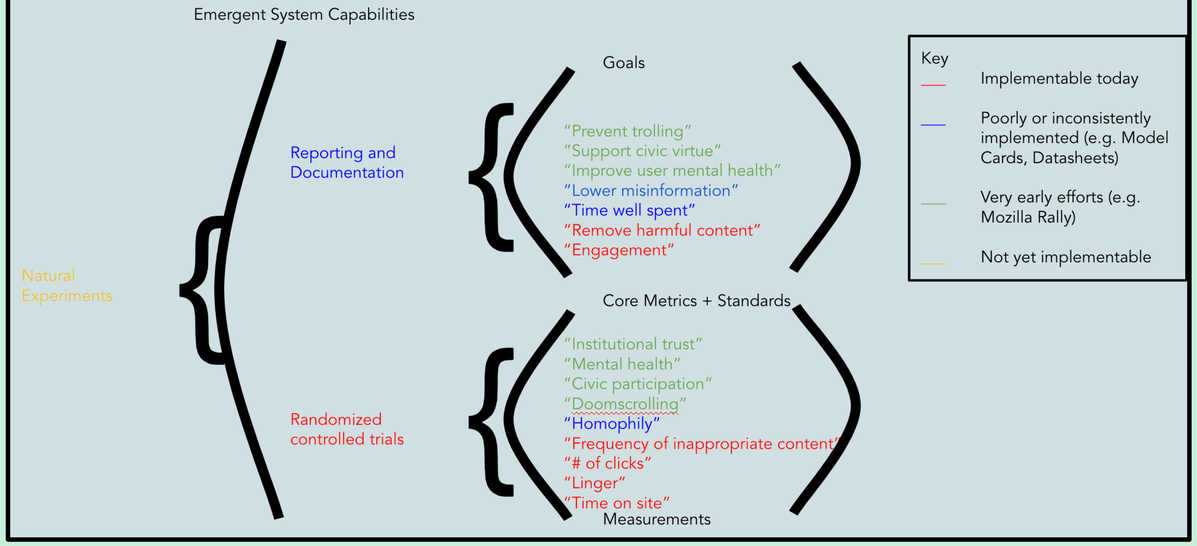
- Efforts to regulate the effects of social media platforms depend on content moderation based on *acute harms*.
- This strategy ignores the *structural harms* generated by the platform itself–e.g. the Facebook emotion contagion study (see table on right).
- These structural harms are analogous to the historical threats and institutional innovations of the public health effort to protect at-risk populations (see below).

Harm Type	Medium	Examples	Regulation Mode	Regulation Regime
Acute	Individual content	Harassment, self-harm, copyright infringement, pornography	Content moderation	Rules-based
Structural	Product architecture	Discrimination, disinformation lifecycle, institutional distrust, mental well-being	Product constraints	Consequentialist

Sequential Developments in 19th Century Public Health



Corresponding Infrastructure for Recommender Systems



Implications for Corporate Governance

- Public health problems had *clear definitions* (mortality) and *structured data collection* (e.g. census statistics). But tech platforms work on a much larger, more opaque scale.
- To acquire both, social media firms need mechanisms for controlled interventions. Tracking the relationship between OKRs, implementation protocols, and development criteria would enable a move from data-driven optimization to *data-driven evaluation*.

Next Steps

- Major social media platforms must implement methods for *randomized controlled trials* so that metrics for structural harms are made available.
- As we argued in the *MIT Tech Review*, a new federal agency or "social media czar" may someday be needed to hold companies accountable–but not until data exists that ties structural harms to specific design decisions.
- We are looking for feedback on design mechanisms for our work-in-progress whitepaper. Collaborators welcome!