Social Media as Fragile State

Caroline Haythornthwaite
Syracuse University
chaythor@syr.edu

Philip Mai
Toronto Metropolitan University
philip.mai@torontomu.ca

Anatoliy Gruzd
Toronto Metropolitan University
gruzd@torontomu.ca

Abstract

Social media platforms are grappling with how to respond to hate speech, misinformation, and political manipulation in ways that address human rights, free speech, and equality. As independent 'states', they are enacting their own rules of conduct, deriving their own 'laws', convening their own extrajudicial self-regulatory institutions, and making their own interpretations and enactments of human rights. With the rise of social states such as Facebook, TikTok, X (formerly Twitter) and Reddit, how fragile are they in their ability to achieve outcomes of fair, equitable and consistent application of their own laws? Could an assessment of the fragility of these social states help identify areas of focus for stability in design, use and operation of social media platforms? What indicators would measure such fragility? This paper draws on the Fund For Peace Fragility State Index for parallels in social media to detail, measure and understand issues of platform precariousness, governance, and support of human rights.

Keywords: social media, platform governance, content moderation, human rights, fragile state

1. The social states of the Internet

Social media has evolved into a pervasive means of communication for information dissemination, education, commercial applications, and more. It is no longer a side show for social interaction; use has infiltrated all aspects of contemporary life. Yet, social media platforms, as independent 'states', are enacting their own rules, deriving their own 'laws', making their own interpretations and enactments of human rights according to the whim and design of leadership or ownership. Social media platforms are grappling with how to respond to hate speech, misinformation, and political manipulation in ways that address human rights, free speech, and equality. But they remain commercial enterprises, with mercurial leadership or ownership. With the rise of the social 'states' of Facebook, X (formerly Twitter), and others, whose law is it? How will social media achieve the rule of law, with outcomes of fair and equal application of laws?

The case of social media is not trivial. The various social media platforms each connect millions, and in some instances billions of people, with users and owners together establishing practices for online behavior. But, as the platforms have grown, so have the abuses, pitting users against owners over ethical dilemmas about what constitutes free speech, and how to ensure the greatest benefit to the greatest number – all backed up by a need to remain commercially viable. As they institute practices that address privacy, human rights, and free speech, each platform is acting alone as independent entities, devising and adhering to their own interpretations of user rights and liberties. These platforms are intertwined with daily life, relied on by their users to always be present. Yet, the recent takeover of Twitter (now X) by Elon Musk shows how the operation of a taken-for-granted, open social media platform can suddenly become fragile, susceptible to collapse through the whim of ownership. Like fragile states, social media platforms can be fragile, becoming unable to support the operation of their social state.

The idea of considering a social media platform as a fragile state builds on the work of The Fund For Peace and their Fragile State Index (FFP, 2022). A fragile state fails to support the basics of statehood, with poor or ineffective governance, and the inability to provide services, security, and protection for citizens. It fails to support the rule of law, i.e., the fair, equal and consistent application of well-defined laws. Members of fragile states face ongoing precarious circumstances: uncertain and changing conditions, lack of control and means to amend conditions. Fragile states are particularly vulnerable to changing conditions, such as the major health crises associated with COVID-19, environmental disasters, or outside challenges to authority (Haken, 2022). Confidence and trust in the state to provide and persist is lost, as is trust in fair and equal application of opportunities, laws, and human rights.

Law and policy to regulate social media have been playing catch-up in areas such as human rights,

URI: https://hdl.handle.net/10125/106679
978-0-9981331-7-1

intellectual property, advertising, defamation, right to privacy, cyberbullying, protection of children, data security, and more. The pervasiveness of internet connectivity, and rapid changes in technologies, require law and policy to advance beyond oversight of local, closed systems to transnational and 'a-national' systems, ones with a nominal locale but present in multiple locales via the internet.

The ongoing question is how social media will be able to create well-defined laws and apply the rule of law in a fair, equal, and consistent manner. Policy research suggests there are thin and thick approaches to the rule of law (van Veen, 2017). 'Thin' approaches promote attention to procedures as a means of achieving fair and equitable outcomes, looking to the design and implementation of judicial systems. By contrast, 'thick' approaches pay attention to outcomes and how procedures achieve results relating to rights and duties.

At present, social media platforms lean toward 'thin' approaches, with published platform standards for behavior and content moderation procedures for identifying and deleting egregious posts, banning repeat offenders, and appealing moderation decisions. Criticisms of social media call out these practices for a lack of transparency and consistency in how speech is identified as offensive, lack of timely recourse for deleted posts or banned users, lack of protection from hate speech, and lack of protection for free speech. To manage these criticisms Meta established an oversight board with members called in to act as independent judges of individual cases. However, the board found that its greater duty is to evaluate policy, giving attention to 'thick' issues of the rule of law, of values, transparency, and human rights (Levy, 2022).

While content moderation is a visible response to the fragility of social media, other matters also affect these platforms, including ownership change, leadership models, new legal requirements and/or restrictions, questions of surveillance, data protection and privacy. Given the dependence of so many on the facilities of these platforms, could an assessment of the fragility of these social states help users in choosing platforms to depend on? What would a fragile state index for social media look like?

2. A fragile state

The Fund For Peace (FFP, 2022) fragile nation-state indicators suggest a way to examine the fragility of social media platforms as if they are social states, Framing and seeing social media platforms through the lens of their being independent social states is both a purposeful provocation and a useful construct to help in understanding social media governance and how issues of human rights can be addressed on these platforms. Such examination can open up consideration of how the rule of law in social media can be defined in a way that transcends any single platform, and, like other nation-states, enacts a common human rights approach to social media.

The FFP Fragile State Index assesses the fragility of a state based 12 indicators addressing Cohesion, Economic, Political, Social and Cross-cutting issues. (Table 1; for full text see FFP, 2022).

Table 1: FFP Fragile State Indicators (in brief)

Cohesion
C1: Security Apparatus: Security threats to a state
C2: Factionalized Elites: Fragmentation of state institutions along ethnic, class, clan, racial or religious lines
C3: Group Grievance: Divisions and schisms between different groups in society
Economic
E1: Economic Decline: Factors related to economic decline within a country
E2: Uneven Economic Development: Inequality within the economy, structural inequality based on group or region
E3: Human Flight and Brain Drain: Economic impact of human displacement
Political
P1: State Legitimacy: Representativeness and openness of government and its relationship with its citizenry
P2: Public Services: Basic state functions that serve the people: provision of essential services, ability to protect its citizens.
P3: Human Rights and Rule of Law: Relationship between the state and population in how fundamental human rights are protected, freedoms observed and respected
Social
S1: Demographic Pressures: Pressures deriving from the population (e.g., disease, epidemic, skewed population distributions) or environment
S2: Refugees and IDPs: Pressure caused by the forced displacement of large communities as a result of social, political, environmental, or other cause.
Cross-cutting
X1: External Intervention: Influence and impact of external actors re security, economic engagement, humanitarian intervention

Three data streams provide input for the FFP scoring: automated content analysis of "media articles, research reports, and other qualitative data points from over 10,000 different English-language sources around the world"; quantitative data sets, "from international and multilateral statistical agencies"; and a separate qualitative in which "a team of social science researchers independently reviews each of the 178 countries, providing assessments based on key events from that year, compared to the previous one." (FFP, 2017). The FPP Conflict Assessment System Tool (CAST) framework is applied to assign scores to the indicators (FPP, 2014), with each indicator assessed based on areas to consider and guiding questions.

3. Social media as fragile states

Social media platforms exert such pervasive influence in contemporary society that in many ways they have become state-like entities of their own. They are managing and enacting policies and practices that address civil discourse, diversity representation, anti-social behavior, political speech, and hate speech. They are enacting governance structures equivalent to a state, with services, security, law enforcement, due process, and inclusiveness. The policies and practices they enact play an overwhelming role in influencing how opinion, information, and news are delivered by whom, to whom, and under what circumstances.

Social media states are fragile in part because the rules of online interaction are continuously emerging, changing with each new wave of participants, and each new type of social media. They join the ongoing reconfiguration of rights, laws and practices associated with the Internet and user-generated content. Where they differ from previous online innovations is in the interactivity – the 'social' of the social media – a relation between participants rather than between the participant and the platform. Participants have leeway to act more like citizens than users, engaging with each other under the umbrella government of the medium, forming platform-specific local governments (Facebook groups, Reddit subreddits, Mastadon instances), and local policing of behavior (moderators). They bring to their participation their ideas about free speech, online community, digital public goods, and how the 'commons' should be used. A lack of community rules fits with the pioneering wild west ideal of open, unfettered, 'free' speech ethos of social media; and the enactment of community rules invokes the participatory democracy ethos of social media. Platform level implementation of rules aligns with neither of these, and instead fits with emerging statehood.

Viewing social media as social states suggests these media may be considered from the perspective of state functions: the extent to which they provide effective governance, services, security, and protection for citizens. However, as states with emerging laws and practices, and continuous challenges to the protections they are putting in place, they are fragile states. A view of social media as fragile states can provide input for participants about conditions affecting the platform, an indication of the stability and long-term prospects of the platform, and a focus for design and implementation of rules, policies, and procedures to stabilize the fragile state.

4. Deriving social media fragile state indicators

Observation of ongoing controversies and practices of social media, and how these affected the stability of these platforms, suggested the idea of fragility; and the FFP index suggested a framework for defining indicators of SM fragility. A process of qualitative analysis was used to derive the set of fragility indicators for social media presented here. First, a wide-ranging set of materials relating to use, controversies, and stability of social media were reviewed to identify major sources of SM instability. Second, results of the review were used to suggest SM equivalents to the FFP Fragile State Indicators. The SM equivalents were assigned themes (independent from the FPP indicator category) which were refined into categories that form the basis of the SM Fragility Indicators discussed below.

5. Social media sources of instability

A wide-ranging review of materials relating to social media was undertaken to assess the fragility of social media platforms. The review included: academic literature on social media; literature and reports on 'fragile states'; news reports about social media; information on major social media platforms detailing their standards, policies, procedures, and policing; and social media platform transparency reports providing data on the prevalence of offences and success in application of procedures to address offending content and actors. The review identified a number of sources of instability where the response in policy and practice indicates state-like activities of governance, service, security, and protection. These are discussed here under the general headings of *Content Moderation*, *Ownership and Leadership*, *Economic Climate*, and *Technology Limitations.* The issues, examples and cases from this review became the basis for assigning SM equivalents to FPP fragility indicators.

5.1. Content moderation

Instability arises from anti-social content and behaviors, including the posting of offensive material, and repeated and persistent individual and group behavior. The constellation of activities associated with content moderation demonstrate a platform's approach to participant protection and policing. Responses include establishing platform standards of behavior and policing methods for responding to transgressions of the standards. Content is moderated by algorithmic and human review for removal of offending materials before dissemination, and participant sanctioning or removal for persistent transgression of standards. Instability also arises from workforce stress when human reviewers are repeatedly exposed to offensive content during the moderation process.

Some key aspects of states include transparency and consistency in the application of laws, protection of human rights, equal opportunity, and equal treatment under the law. Some SM platforms address transparency by providing reports about the rate and success of content and account removal, including rates of appeal and reversal or removal. However, criticisms of social media content moderation call out a lack of transparency in how posts are identified as offensive, whether the moderation rules are applied consistently against all similarly situated parties and lack of timely recourse for deleted posts or banned users. Moderation requires managing a balance of protections and rights, such as protecting participants from hate speech while also protecting freedom of expression. Criticisms point out how moderation algorithms can systematically disadvantage one group of posters over another, leading to fragmentation in services and a lack of equity. For example, the discussion of hate speech (often by those affected by it) is difficult to distinguish from intended hate speech, and the presence of offending speakers who can alienate potential participants.

To date, this kind of content moderation has been determined and implemented at the platform level. New ideas of composable moderation provide alternative structures, putting more control at the community level, implementing new governing structures for these social states, e.g., as in this explanation for the social media BlueSky: "Centralized social platforms delegate all moderation to a central set of admins whose policies are set by one company. This is a bit like resolving all disputes at the level of the Supreme Court. Federated networks delegate moderation decisions to server admins. This is more like resolving disputes at a state government level, which is better because you can move to a new state if you don't like your state's decisions." (Garber, 2023). See also Technology Limitations.

5.2. Ownership and leadership

Major social media are companies, owned and led by individuals who provide direction and strategy for the company. The ownership provides stability for the company but is a source of instability when their decisions or behavior incite criticism. Cases include using social media data for experiments (Kramer, Guillory, & Hancock, 2014), participant data provided for use by outside companies, as was the case with Cambridge Analytica (e.g., Cadwalladr & Graham-Harrison, 2018); and ownership decisions to limit distribution of data, as was the case with a New York Post article on Joe Biden (Bond, 2020).

Changes in ownership have the potential to create major instability in a platform, e.g., as the impact the sale of Twitter (now X) and new leadership by Elon Musk has had on the viability of the platform (Connelly, 2023). Such instability can lead to migration away from one platform to another with alternate social media ownership structures (e.g., Mastadon, BlueSky, Steemit).

Ownership also has a role in control of who gets to post what online. Banning accounts or preventing distribution of posts can create challenges about freedom of the press when journalists are banned (Kaltheuner, 2022). SM companies have implemented special procedures for postings on social or political topics. For example, Meta's policy on Ads about Social Issues, Elections or Politics "requires enhanced transparency from elected and appointed officials, candidates for office, and advertisers of content that includes social issues, electoral or political ad content." (Meta, n.d.-a). Special considerations provide verification of identity for politicians, celebrities, and others.

Further instability arises from issues related to the country of ownership, the country of use, and the influence of those countries on information privacy and banning use of the platform. Non-domestic ownership can lead to partial or full bans on use, as is happening for TikTok over its Chinese ownership (Kari, 2022; Maheshwari, & Holpuch, 2023; Anguiano, 2023, Jalonick, 2023). Government oversight of non-domestic platforms can lead to restrictions or bans, as is the case for Facebook in China, X (formerly Twitter) in Nigeria, both in Iran and North Korea, TikTok in India, VK in Ukraine and more (Barry, 2022). See also Economic Climate Indicators.

5.3. Economic climate

Downturns in the economy, as in the 2020 post-Covid recession, can affect business viability where there is a lack of diversity of revenue sources. Heavy reliance on business advertising is a source of instability for SM platforms; resulting workforce reductions lead to a 'brain drain' of experienced workers (Mac, Isaac & Conger, 2023; Rushe et al, 2022). The Covid-19 epidemic affected the economy through the major changes of work from home. An unexpected outcome for social media companies was that affected human reviewers could not work from home due to the offensiveness of the materials they were vetting (Dworkin & Tiku, 2020).

5.4. Technology limitations

Moderation by algorithm is a necessity given the billions of posts to be served to millions of users every minute. Yet there are difficulties in accurately categorizing data as compliant or non-compliant with platform standards. Detection challenges include: rapidly changing language and multi-media use; intentional actions to bypass detection (e.g., disguising text with obscured fonts; Cobbe, 2021); differences in genre and media use (e.g., memes, parody, sarcasm, media effects); and contextual use of the same text (e.g., discussion of hate speech).

The details of the texts used to create screening algorithms, and how they are defined are not shared by social media platforms, making it difficult for evasion, but also for external review of criteria and sharing across platforms. Challenges have led to some open-source initiatives for sharing technology, and some kinds of data, e.g., Meta sharing of photo and video matching technology, and the Hasher-Matcher-Actioner (HMA) for identifying extremist content by matching to hashes created from earlier copies of content (Clegg, 2022).

Critique of algorithms also extends to the choice of language training data used to validate the algorithm accuracy. Bender, Gebru, McMillan-Major & Shmitchell (2021) argue that where algorithms are trained on large language models, bias can be introduced in determining acceptable speech. The authors discuss how artificial intelligence models that use text scrapped from the Internet as language training data overrepresent "white supremacist and misogynistic, ageist, etc. views" and risks "perpetuating dominant viewpoints, increasing power imbalances, and further reifying inequality" (Bender et al, p. 613-614). A further critique addresses shortcomings of anti-bias research due to predominant attention to Western ethics issues, conducted by researchers in Western institutions, thus lacking a global view (Prabhakaran et al, 2022).

6. Identifying social media equivalents for fragile state indicators

The FFP Fragile State Indicators were used as a starting point for identifying indicators of social media fragility. Each indicator was considered for equivalents affecting social media platforms. For example, the FPP Index 'C1: Security Apparatus' "considers the security threats to a state, … serious criminal factors, such as organized crime and homicides, and perceived trust of citizens in domestic security." (FPP, 2022). SM equivalents to 'threats to the state' were identified as technical threats (e.g., such as hacking, data breaches), social threats (e.g., fake accounts, cyberbullies), threats to social engagement (e.g., offensive content, hate speech), use for illicit activities, and trust in platform governance (e.g., content moderation and appeals process).

This step produced a list of multiple SM equivalents for each FFP Indicator, with a number of these appearing against several FPP indicators. Themes were identified and assigned to each SM equivalents, e.g., security, fragmentation, leadership, rights. Themes were then evaluated to derive a set of categories that became the foundation for the SM Fragility Indicators. (A table of SM equivalents to FPP indicators is not presented here, but elements are discussed in the descriptions of each SM indicator.)

7. Social media fragile state indicators

The previous steps resulted in a set of categories of SM fragility based on sources and cases of instability found in the literature and social media site descriptions of their governance. In taking these as input for SM Fragile State Indicators, the focus is on describing the indicators in a way that elicits consideration of that category of fragility. These are presented as indicators (as done for the FFP indicator descriptions) rather than as lists of example cases. Thus, the text refers to what would be considered in assigning a fragility score under each indicator.

In brief, the indicators are: *Security, Threats, Protection and Safety; Fragmentation; Human Rights and the Rule of Law; Demographic Pressures; Economic Indicators;* and *External Indicators*. The order of presentation generally moves through from threat to outcome to rights, and then to outside influences; however, there is no intended information in the order. While the derivation of common themes is intended to reduce redundancy, some indicator conditions are found under more than one heading.

7.1. Security, threats, protection, safety

FFP equivalents C1: Security Apparatus; C3: Group Grievance; P2: Public Services

Security, Threat, Protection and Safety indicators consider both technical and social threats to the platform, and threats to participants' safety and well-being. Technical threats to the platform include challenges to system operation, such as denial of service attacks, data breaches, system outages, and social threats such as fake accounts, scammers, spammers, cyberbullies, anti-social postings and behavior. Technical threats to participant safety include harm by data loss, privacy invasion, and unapproved disclosure of person information, and social threats include harm by exposure to inappropriate content, forms of misinformation, and political or personal manipulation (e.g., by scammers, online predators). The indicator includes threats to social engagement such as disruption of social engagement by trolls, cyberbullies, and attacks on identifiable groups; and threats from participant circumvention of protections and breach of platform standards, including illegal use and use to facilitate illegal activity, misrepresentation of participant information, postings in violation of platform standards, and circumvention of age verification for underage users.

Indicators address the external or ambient social, technical, and cultural indicators that relate to the rate of change affecting technical challenges (e.g., software updates, hacker challenges), accepted social norms and acceptance and conformity to platform standards. External indicators address the production and speed of change of misinformation, disinformation and malinformation indicators relates to the stability of content moderation practices in keeping up with change. (See External Indicators)

Protection indicators consider the platform response to threats, including proactive and reactive responses to technical threats, adjudication methods for responding to social threats, and transparency in efforts. These include protection from harm from content, such as algorithmic and human review of content before posting; warning and/or removal of content not conforming to site standards; removal of fake accounts; warning and/or removal of offending accounts; banning topics (e.g., pro anorexia sites, anti-vaxxers; age verification measures). Protection indicators extend to employee work conditions, particularly for those reviewing content. See also Economic Indicators.

Other indicators address policing of platform standards through content moderation, content appeals and oversight boards, transparency of efforts, and equal application of rules. See Human Rights and the Rule of Law Indicators.

7.2. Fragmentation

FFP equivalents C2: Factionalized Elites; C3: Group Grievance; E2: Uneven Economic Development

The Fragmentation indicators consider divisions, separations, filter bubbles existing in and exacerbated by social media platform design and practices. The indicators consider how historical fragmentation stemming from societal systemic bias is exhibited in the population distribution of social media platform participants, including distribution in digital literacy (digital divide, digital spectrum), socioeconomic indicators, and access by language, ability, location. The indicator considers bias exhibited by algorithms used for social media content evaluation, marketing, etc. exhibit bias, systematic error and exclusion by race, gender, socioeconomic indicators, etc.

Fragmentation indicators include divisions within and across social media resulting from design and participant populations, e.g., in filter bubbles that separate social network components by socioeconomic status, political views, language use, reading habits, etc. Indicators can also include attitudes to social media use, e.g., in attitudes driven by varying definition of 'free speech' and the culture of the medium, and by leadership in their attitudes to social media use.

Economic and political fragmentation are indicated by differential treatment of elite users (e.g., politicians), suspension of users or their posts (e.g., members of the press), limitations on use by country, and pay for service agreements. See also Human Rights and the Rule of Law Indicators.

7.3. Human rights and the rule of law

FFP equivalents: P3: Human Rights and Rule of Law; C3: Group Grievance; P1: State Legitimacy; P2: Public Services

The Human Rights and the Rule of Law indicators consider SM platform awareness, protections, and review procedures for fundamental human rights and freedoms in the SM context, as described in the United Nations universal declaration of human rights. The context includes platform implementations to safeguard human rights, including well-defined rules that are applied consistently and equitably, and review procedures both for the rules and for that application of the rules. The indicators consider the way the platform implements rules for decision making that affirm fundamental human rights and attend to human rights outcomes, and how review of these procedures achieve of human rights on the SM platform. The indicator may consider how effort is put toward each of the human rights (e.g., when outside attention focuses only on some rights, e.g., 'free speech'), and to rights in a global context (Prabhakaran et al, 2022).

The indicators consider the extent to which rules address the content of any submission to the platform and the actor(s) making the submission, and the context of that submission and the actor(s). Evaluation of context may include determining the way a submission continues conversations that violate human rights, supports or is supported by other actors in a challenge to human rights, creates a self-sustaining community that challenges such rights, and gives a platform for extreme views (e.g., sub-communities of violators). Indicators consider the policing of rules, and the fair and equitable application of sanctions for violating rules; platform procedures for obtaining external, impartial oversight for evaluation of aims; and processes for content moderation and sanctioning actor behavior (e.g., oversight boards). The indicators consider SM platform engagement with other entities in the discussion and promotion of human rights protections applicable to all SM. The indicators may also address how the platform frames rights, and/or promotes new considerations of human rights in the SM context (e.g., reframing rights in terms of participant perspectives, e.g., the 'right not to be subject to hate speech').

The indicators consider the extent to which participant behavior, both individually and in groups, aligns with respecting and protecting human rights, and how platform ethos and rules affects participant behavior in a way that contributes to support or violation of human rights. The indicators consider how social media manage the discussion of violation of human rights versus actual violation of rights; and handle grievances about violations of human rights from participants, employees and/or external stakeholders in a fair and equitable manner.

The indicators may also consider the extent to which digital platforms comply with the principles under development by UNESCO (2022): respect human rights in content moderation and curation; be transparent in how they operate; empower users to understand and make informed decisions about use; and be accountable to relevant stakeholders for their terms of service and content policies.

7.4. Demographic pressures

FFP equivalents S1: Demographic Pressures; S2: Refugees and IDPs; X1: External Intervention

Demographic pressures indicators for consider the population using the social media, and environmental pressures on that population. This includes cultural attitudes to social media use that lead to restrictions by age, location, sector, etc. (e.g., bans on using TikTok in government business); skewed user demographics (e.g., by age); environmental influences on social media operation (e.g., changes to 'work from home' during Covid and impact on human reviewer work practice). Demographic pressures may also include increased population due to migration from other social media. Indicators may also consider the support given to different user populations, e.g., abiding by Universal Design Principles (CEUD, 2023) for inclusive access, providing instruction for new participants, and educating participants on social risks with posting personal information publicly.

7.5. Economic indicators

FFP equivalents E1: Economic Decline; E3: Human Flight and Brain Drain; S2: Refugees and IDPs

Economic indicators consider the site's economic viability and stability, including ownership, stock market evaluation and shareholder confidence, revenue source and stability of the source, and domestic economic climate (e.g., Rushe et al, 2022). This indicator also includes human flight and brain drain as dissatisfaction with a platform can lead to a flight to another (e.g., X (formerly Twitter) to Mastadon, BlueSky and other social platforms), a platform may cease to maintain its critical mass and/or be overtaken by a newer SM fad (fading and failing SM platforms, such as Friendster; rising platforms such as TikTok). Brain drain in this case is conversation, interaction drain, with follow on effects on advertising revenue. See also Leadership and Ownership.

7.6. External indicators

FFP equivalents X1: External Intervention

External indicators consider factors outside the social media platform's control that affect operations. This includes government intervention, both domestic and foreign, in development of new laws (e.g., DMCA, GDPR), prohibiting use (of TikTok; or of Facebook), monitoring, regulating, limiting, or censoring use (Chinese government requiring filters for SM content). This can also include consideration of the location of company ownership (e.g., TikTok Chinese ownership) and its impact of social media operation. See also Leadership and Ownership.

8. Futures

The analysis so far has collected the SM indicators under themes that can serve as categories for assessing fragility. A next step is to define more specific indicators under these categories as done for the Funds for Peace indicators. The more specific indicators would help in creating a more consistent assessment of fragility under each of the SM fragile indicator categories on the way to creating a SM fragile state index.

8.1. From indicators to index

The purpose of the index is to assess both the state-of-the-art in SM fragility (or stability) across media, and from year to year. To move from indicators to index requires creating a scoring schema for the six SM fragile state indicators. The FFP methodology can serve as a model. As an example, for FFP, Fractionalize Elites are assessed on areas of concern such as: Fragmentation, National Identity; Extremist Rhetoric, Stereotyping, and Cross Cultural Respect. These and the other areas of concern are further defined with questions, e.g., for stereotyping "Is religious, ethnic, or other stereotyping prevalent and is there scapegoating?". With evaluation of these areas and questions, a score is given from 0, no factions in the political leadership to 10, no political class or national leader exists acceptable to the majority of the population (FPP, 2014, p. 15).

What would be a similar scoring schema for the SM Fragile State indicators? Taking the Fragmentation indicator as an example, areas to consider might address: *Design*. Is bias exhibited by algorithms used for social media content evaluation, marketing, etc.? *Historical inequities*. Do divisions exist that support or advocate for discrimination on the basis of race, gender, sexual orientation, national origin, etc.? *Divisions*. Does the design facilitate divisions within the social medium (e.g., filter bubbles)? *Attitudes*. Do attitudes to free speech (freedom of expression, etc.) create divisions within the medium? *Economics and politics*. Does differential treatment of participants exist? To what extent is the economic or political success of the platform tied to fragmented use that is counter to fair and equitable use? From this assessment a score could be given from 0, no fragmentation, to 10 for fragmentation along increasingly fractionated and antagonistic divides with intentional promotion of historical fragmentations.

While more work is needed to further articulate the questions and scoring procedure, this is the kind of schema to be created for each indicator, followed by testing for reliability and validity.

8.2. From thin to thick approaches

Most social media procedures for managing online behaviors use a 'thin' approach to applying the rule of law, depending on defined procedures. Yet, current challenges suggest a 'thick' approach, moving from procedure, e.g., to identify hate speech and ban offending users, to outcomes, e.g., to provide a safe space for speech. Recently, the Meta oversight board, which is called in to act as independent judges on applying procedure to individual cases, found that its greater duty is to evaluate policy. The board called in to act as judges for a 'thin' procedure for the platforms' rule of law, discovered the need for attention to 'thick' issues of rule of law, of values, transparency, and human rights (Levy, 2022).

In general, social media states have been acting independently, facing challenges and responding with in-house solutions. However, the complexity of the social exchanges on these platforms, the increasing challenges in defining and maintaining platform standards, and the global reach suggest the utility of pooling strategies and resources for pursuing 'thick' policy directions and human rights.

What would a social media United Nations-like council address as human rights across social media platforms? (See also Article 19, 2021; Freeman Spoligi Institute for International Studies, 2019.) At present, social media frame their policy issues in technology user terms, as rules that individuals must adhere to for continued permission to use the space. What has not happened yet is to reframe those rules as rights, e.g., reframe the rule that 'you must not post hate speech' to a right: 'you have the right not to be subject to hate speech'. If the rhetoric is reframed in this way, and in terms of human rights outcomes, it also reframes the orientation to technology design and rule of law in social media contexts. Value-sensitive design (Knobel & Bowker, 2011; Friedman & Hendry, 2019), an approach that considers human ethical and moral values at the technology design stage, reorients the design process from implementation of features to design for outcomes, the 'thick' rule of law. For example, *composable moderation* (BlueSky, Mastadon and others) originates from ideas about participant control: "Anyone should be able to create or subscribe to moderation labels that third parties create." (Garber, 2023, online; Cross, 2023; Oleaga, 2023). Composable moderation changes the orientation from platform-determined protections to participant-selected protection, moving toward the 'right to choose what you see'.

8.3. From single platform action to shared knowledge

Is Meta's oversight board the beginning of a United Nations approach to social media? One Meta board member is quoted as posing the question "Is access to Facebook a human right?" (Levy, 2022). This is not as absurd as it seems. Not long ago no one would have said access to the Internet is a human right. But its penetration into every aspect of commerce, information dissemination and political discussion has made social organization dependent on internet access. In 2016, the United Nations added specific reference to the Internet as necessary for the right to freedom of opinion and expression through any media regardless of frontiers. Is access to social media the next specification? Could a unified, cross-platform approach to management of the rule of law on social media aid in creating stability for social media states?

Joint initiatives are beginning to appear for sharing tools and pooling knowledge. Among these initiatives are the Global Internet Forum to Counter Terrorism (GIFCT), "an NGO that brings together technology companies to tackle terrorist content online through research, technical collaboration and knowledge sharing." (Clegg, 2022, online); and The Global Network Initiative (GNI), a multistakeholder effort to establish principles and implementation guidelines for the ICT sector for "responsible company decision making in support of freedom of expression and privacy rights" (GNI, 2023). There have also been discussions of Social Media Councils to address content moderation and issues of free speech (e.g., Article 19, 2021; Freeman Spoligi Institute for International Studies, 2019), but also warnings from the Electronic Freedom Foundation that such an effort might end up "legitimating a profoundly broken system", with questions such as: Who determines council membership? How will the council remain independent from the SM companies and funders? How are cases for review selected? (McSherry, 2019).

9. Conclusion

As noted, more steps are needed to arrive at a reliable index. As well as further specification within categories, and scoring from indicator to index, some further questions are also outstanding. Among these are: Who are the 'citizens' of these social states? What is the relationship between social states and nation states? Whose responsibility is it to implement and monitor SM platforms for adherence to rule of law and ethical behavior? While there are still questions to consider, viewing SM platforms as fragile helps in bringing forward such questions, and going further in considering how the social states of social media overlap and interaction with other social and regulatory spheres. For now, the collected themes and categories presented here for the SM Fragile State Index begin to show the points of vulnerability across platforms. Attention to fragility provides a way to assess strong and weak points in the viability of platforms that may be addressed by technical, social, or policy developments, by the platforms or external regulation. Where fragility indicators highlight issues affecting all social media, they suggest places where knowledge can be pooled, and group effort applied to support the operation of social media.

Acknowledgement. This research was funded in part by the Social Sciences and Humanities Research Council of Canada (PI: Gruzd).

10. References

Anguiano, D. (2023, May 17). Montana becomes first US state to ban TikTok. *The Guardian.* https://www.theguardian.com/us-news/2023/may/17/tiktok-ban-montana

Article 19. (2021, Oct. 12). Social media councils: One piece in the puzzle of content moderation. *Article 19.* https://www.article19.org/resources/social-media-councils-moderation/

Barry, E. (2022, Jan. 18). These are the countries where Twitter, Facebook and TikTok are banned. *Time.* https://time.com/6139988/countries-where-twitter-facebook-tiktok-banned/

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. [Margaret Mitchell] (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? FAccT '21, March 3–10, 2021. https://doi.org/10.1145/3442188.3445922

Bond, S. (2020, Oct. 14). Facebook And Twitter limit sharing 'New York Post' story about Joe Biden. NPR. https://www.npr.org/2020/10/14/923766097/facebook-and-twitter-limit-sharing-new-york-post-story-about-joe-biden

Cadwalladr, C. Y. & Graham-Harrison, E. (2018, Mar 17). Cambridge Analytica files. *The Guardian*. https://www.theguardian.com/news/series/cambridge-analytica-files

Centre for Excellence in Universal Design (CEUD, 2023). What is universal design. The 7 principles. Retrieved June 5, 2023 from https://universaldesign.ie/what-is-universal-design/the-7-principles/

Clegg, N. (2022, Dec. 13). Meta launches new content moderation tool as it takes chair of counter-terrorism NGO. Retrieved May 24, 2023 from https://about.fb.com/news/2022/12/meta-launches-new-content-moderation-tool/

Connelly, E. A.J. (2023, Mar. 30). Twitter ad revenue plunges 89% since Musk takeover as major brands stay away. *The Wrap.* https://www.thewrap.com/ twitter-ad-revenue-plunges-elon-musk-bloomberg/

Cobbe, J. (2021). Algorithmic censorship by social platforms: Power and resistance. *Philos. Technol.* 34, 739–766. https://doi.org/10.1007/s13347-020-00429-0

Cross, K.A. (2023, May 12). BlueSky Ain't It. The new social platform will never be the "next Twitter". *Wired.* https://www.wired.com/story/bluesky-twitter-social-media.

Dwoskin, E. & Tiku, N. (2020, Mar. 23). Facebook sent home thousands of human moderators due to the coronavirus. *Washington Post.* https://www.washingtonpost.com/technology/2020/03/23/facebook-moderators-coronavirus/

Freeman Spoligi Institute for International Studies (2019, Feb. 1-2). Social media councils: from concept to reality. Retrieved May 29 from https://fsi.stanford.edu/content/social-media-councils-concept-reality-conference-report

Friedman, B. & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination.* Cambridge, MA: MIT Press.

Fund For Peace (2014, Mar. 10). *CAST Conflict Assessment Framework Manual (2014 Reprint).* https://fundforpeace.org/2014/03/10/cast-conflict-assessment-framework-manual-2014-reprint/

Fund For Peace (2017). *Fund For Peace Index Methodology and CAST Framework.* Retrieved May 29, 2023. https://fragilestatesindex.org/methodology/

Fund for Peace (2022). *The Fragile State Index*. Retrieved May 29, 2023. https://fragilestatesindex.org/.

Fund of Peace (n.d.). *P3: Human Rights and Rule of Law*. Retrieved May 29, 2023. https://fragilestatesindex.org/indicators/p3/.

Garber, J. (2023, Apr. 13). Composable moderation. Retrieved May 25, 2023 from https://blueskyweb.xyz/blog/4-13-2023-moderation

Global Internet Forum to Counter Terrorism (n.d.). Preventing terrorists and violent extremists from exploiting digital platforms. Retrieved May 29, 2023 from https://gifct.org/

Global Network Initiatives (2023). Protecting and advancing freedom of expression and privacy in the ICT sector. Retrieved May 29, 2023 from https://globalnetworkinitiative.org/

Guinness, H. (2022, Dec. 19). Meta is open sourcing its automated content moderation tool. The Hasher-Matcher-Actioner, explained. *Popular Science.* https://www.popsci.com/technology/meta-hasher-matcher-actioner-open-source/

Haken, N. (2022, July 8). Coming apart at the seams: Fragility in a time of COVID-19. https://fragilestatesindex.org/2022/07/08/coming-apart-at-the-seams-fragility-in-a-time-of-covid-19/

Jalonick, M.C. (2023, Mar. 29). TikTok ban pushed by Missouri's Hawley blocked in Senate. https://apnews.com/article/hawley-tiktok-ban-senate-bill-blocked-dc504ad560b949bd11e2970afd3669dd

Kaltheuner, F. (2022, Dec. 16). Twitter's suspension of journalists threatens media freedom. *Human Rights Watch.* https://www.hrw.org/news/2022/12/16/twitters-suspension-journalists-threatens-media-freedom

Knobel, C. & Bowker, G.C. (2011, July). Values in design. *Communications of the ACM, 54*(7), 26-28.

Kramer, A. D.I., Guillory, J. E., Hancock, J. T. (2014, June 2). Experimental evidence of massive-scale emotional contagion through social networks, *PNAS. 111* (24) 8788-8790. https://doi.org/10.1073/pnas.1320040111

Levy, Stephen (2022, Nov. 8). Inside Meta's oversight board: 2 years of pushing limits. *Wired Magazine*. https://www.wired.com/story/inside-metas-oversight-board-two-years-of-pushing-limits/

Mac, R., Isaac, M. & Conger, K. (2023, Feb. 28). 'Sometimes things break': Twitter outages are on the rise. *NY Times.* https://www.nytimes.com/2023/02/28/technology/twitter-outages-elon-musk.html

Maheshwari, S. & Holpuch, A. (May 22, 2023). Why countries are trying to ban TikTok. *NY Times.* https://www.nytimes.com/article/tiktok-ban.html

McSherry, C. (2019, May 10). Social media councils: a better way forward, window dressing, or global speech police? Electronic Freedom Foundation. https://www.eff.org/deeplinks/2019/05/social-media-councils-better-way-firward-lipstick-pig-or-global-speech-police

Meta (2022). *Facebook community standards*. Retrieved March 2023 from https://transparency.fb.com/policies/community-standards/.

Meta (n.d.-a) Ads about social issues, elections or politics. Retrieved May 29, 2023 from https://transparency.fb.com/policies/ad-standards/siep-advertising/siep.

Meta (n.d.-b) Reviewing high-impact content accurately via our cross-check system. Retrieved May 29, 2023 from https://transparency.fb.com/enforcement/detecting-violations/reviewing-high-visibility-content-accurately/

Oleaga, K. (2023, Apr. 21). Behind BlueSky's comprehensive approach to algorithmic transparency and content moderation. *Hypemoon.* https://hypemoon.com/2023/4/bluesky-takes-a-comprehensive-approach-to-social-media-moderation-and-algorithmic-transparency

Paul, K. (2022, Dec. 30). US bans China-based TikTok app on all federal government devices. https://www.theguardian.com/technology/2022/dec/30/us-tiktok-ban-government-devices-china

Prabhakaran, V., Mitchell, M., Gebru, T. & Gabriel, I. (2022, Oct. 6). A human rights-based approach to responsible AI. Poster at EAAMO '22. Retrieved May 22, 2023 from https://doi.org/10.48550/arXiv.2210.02667

Rushe, D., Oladipo, G., Bhuiyan, J., Milmo, D., & Middleton, J. (2022, Nov. 4). Twitter slashes nearly half its workforce as Musk admits 'massive drop' in revenue. *The Guardian.* https://www.theguardian.com/technology/2022/nov/04/twitter-layoffs-elon-musk-revenue-drop

TikTok (2022). *TikTok community guidelines.* Mar 1, 2023 from https://www.tiktok.com/community-guidelines

van Veen, Erwin (2017). A shotgun marriage: Political contestation and the rule of law in fragile societies. Clingendael's Conflict Research Unit. https://www.clingendael.org/pub/2017/a_shotgun_marriage/