CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

WORKING PAPER

MAY 2023

Operational Reporting By Online Services: A Proposed Framework

Samantha Lai, Naomi Shiffman, and Alicia Wanless

Operational Reporting By Online Services: A Proposed Framework

Samantha Lai, Naomi Shiffman, and Alicia Wanless

Carnegie does not take institutional positions on public policy issues; the views represented herein are those of the author(s) and do not necessarily reflect the views of Carnegie, its staff, or its trustees.

Carnegie Endowment for International Peace
Publications Department
1779 Massachusetts Avenue NW
Washington, DC 20036
P: + 1 202 483 7600
F: + 1 202 483 1840
CarnegieEndowment.org

This publication can be downloaded at no cost at CarnegieEndowment.org.

Contents

Introduction

The often opaque operations of major tech companies obscure the role social media companies have played in shaping the information environment, and the actions they have taken to mitigate harmful behaviors. Calls abound for greater transparency and for platforms to conduct reporting that can provide insight on how they conduct their operations.[1] Yet these calls only highlight specific items rather than suggest broad change. Many calls are focused on topics like political advertising, disinformation, child abuse and exploitation material, or extremist content on social media platforms.[2]

Adding to this complexity, governments worldwide are increasingly exploring regulation as a tool to introduce and enforce transparency reporting.[3] Consequently, a handful of international companies could face new rules from dozens of countries. This could present serious operational challenges for companies. It would also lead to differing and piecemeal reporting practices across issue areas, rather than facilitating a comprehensive picture of data, policies, and actions in the context of the information environment, which is the space where humans and machines process information to make sense of the world, and of which online service providers control an important aspect.

Researchers, policymakers, and civil society groups need to come together to clarify among themselves and for platforms what type of information would be most helpful to protect the public interest and what framework could ensure this information is feasible for platforms to provide. This policy proposal suggests eight potential categories for reporting, in an effort to enable researchers and policymakers to better understand how online service providers operate. It draws on a review of ninety-eight proposals, eleven regulations or government policies, existing transparency practices, and a yearlong working group of experts.[4]

Broadly speaking, those eight categories of operational reporting are:

1. User level: aggregated information about different types of users.

2. Platform level: platform architecture and how platforms work.

3. Policy development and enforcement: internal policies that govern activity and use of the platform (e.g., content moderation policies).

4. Internal research: the types and findings of research conducted inside the company to understand impacts on users and of interventions.

5. External requests for intervention: requests to platforms from a third party to act on a user, an account, or a piece of content (e.g., user-flagged content, content removal requests, account suspension requests).

6. Data access requests and tooling: requests from third parties for access to the personal information of a user or group of users (e.g., law enforcement requests, court orders) and the tools created to facilitate access to data.

7. Terms of service and privacy policies: terms of service refer to agreements between companies and users regulating the use of the service. A privacy policy is a statement that discloses how and when a company collects, uses, and shares user information. This information is typically shared with third-party services.

8. Third-party relationships: arrangements between companies and third-party organizations regularly accessing user data.[5]

There are a few key caveats to the proposed reporting framework. First, the level of detail in reporting should differ across audiences, as certain categories of reporting include sensitive content that can be misused if made available to the general public. Second, the proposed framework outlines various possible categories of reporting but recognizes that further prioritization is needed to establish a realistic, feasible path for platforms to gradually expand the information they report on. And third, operational reporting is only part of the process. To ensure trust and legitimacy, an independent auditing body must be established to review the accuracy of what platforms are reporting on.

The level of detail in reporting, as it is proposed here, is more than what many other industries provide. Most other industries do not have such differential treatment of users on an individual level, with the exception of the financial industry, which is a heavily audited sector. Moreover, given the role of the information environment in human decisionmaking, the need for citizens within democracies to make free and informed decisions for the very legitimacy of the political system, and the increasingly important position of online services to all of this, it is not unreasonable to require a greater degree of transparency from this sector. Indeed, some of the reporting outlined here is becoming normalized across the tech sector. For example, the Meta Oversight Board, where one of the co-authors works, helps Meta disclose in medium detail its risk assessment processes, organizational charts, decisionmaking structures, and how its policies have changed over time.[6] Meta's Community Standards now include historical versions of those Community Standards so users can see how they have evolved over time.[7] To that end, this paper lays out an ideal for operational reporting, with the understanding that the suggestions are not likely to be adopted in their entirety all at once, but even the act of articulating a framework can help shape norms moving forward.

Guiding Lights

For the purposes of this paper, operational reporting is "the aggregated public, semi-public and private reporting of quantitative and qualitative data (not individual user data) by online services about aspects of their operations, published on a regular and consistent basis."[8] While it can aim to increase transparency, operational reporting should not be confused with transparency reporting by governments aimed at increased accountability to citizens in democracies (for example, government disclosures on how platforms are used for citizen surveillance).[9] Operational reporting is also distinct from other concepts such as data access, which makes raw data available for research purposes through various means.

Existing operational reporting by the tech industry is predominantly undertaken as an elective self-regulatory measure. Companies determine the method, scope, and content of reports, and this results in disclosures that are *ad hoc*, unstandardized, and may contain biases or omissions such as data from certain geographies.[10]

The most common types of transparency reports include those on requests for access to information,[11] content and account removals,[12] and enforcement of community standards.[13] Other industry reports include analysis of coordinated inauthentic behavior on a specific platform,[14] internet disruptions,[15] application removals,[16] and the types of content viewed by users.[17] Piecing together the totality of transparency efforts by companies can be time consuming, requiring researchers to look through series of data releases, press statements, differently structured transparency centers, and synthesize reporting conducted over a range of timeframes and formats. Data between reports don't clearly relate to each other, making it difficult to get a comprehensive picture even for a single platform. Reports are often broken into different topics without a clear explanation of the relationship between topics, such as content removal or information requests, and published in disparate blog posts, although companies have begun aggregating reports into transparency centers.[18]

To help correct these deficiencies in reporting, more structured approaches have been suggested. Reporting regimes such as the European Commission's Code of Practice on Disinformation and the Organisation for Economic Co-operation and Development's (OECD's) Voluntary Transparency Reporting Framework have been implemented.[19] The Digital Services Act (DSA) establishes new transparency obligations in the European Union (EU).[20] The UK Online Safety Bill, which continues to work its way through Parliament, mandates the British regulator OfCom to request annual transparency reports from companies.[21] Pending U.S. legislation such as the Digital Services Oversight and Safety Act of 2022, the Platform Accountability and Consumer Transparency Act, the Platform Accountability and Transparency Act, and the Algorithmic Justice and Online Platform Transparency Act also reflect popular demands for greater platform transparency.[22]

These demands for improved transparency reflect recognition of important public interests. However, each new voluntary reporting mechanism and regulation can introduce new and detailed forms of reporting that, if not harmonized across jurisdictions and topics, could lead to implementation challenges for those organizations that are obliged to comply.

Having researchers, policymakers and civil society groups clarify the purpose of operational reporting—beyond simply informing stakeholders about an organization's operations—could improve the preparation and eventual use of reports. To help organize thinking about reporting requirements, we describe eight potential categories of information and why they are important. The categories are listed below with a short description outlining what each category entails, why it should be reported on, and associated challenges, along with specific questions that would guide such reporting.

Into the Light: Operational Reporting Categories

Category 1: User Level

Guiding Questions for User-Level Reporting

Definition: User-level reporting refers to aggregated information about different types of users.

1. Segment reports by user category:
   a. Individual accounts
   b. Collections of individual accounts
   c. Organizations/businesses
2. Segment reports on user information by user category. Include information on:
   a. Demographics (e.g., age, gender, location)
   b. Psychographics (e.g., interests)
3. Segment reports on user activity by user category. Include information on:
   a. Types of content of public posts, comments, and engagement
   b. Posting patterns, such as where users post and share
   c. Networks users form
   d. Users that purchase ads:
      i. The types of ads purchased by which types of users
      ii. Whom they target
      iii. In what languages and geographies, and when
4. What services are available (e.g., making public/private posts, purchasing ads, setting up public/private groups)?
   a. What can users do with those services?
   b. What languages are those services offered in?
   c. As platforms grow their user base, what is their pathway of progression in terms of expanding available languages?

Operational reporting at the user level sheds light on who is accessing a particular online service. Reporting at the user level entails a mix of quantitative and qualitative information, providing insights into user bases, the kinds of services they can access, and how they engage with those services.

Standardized and consistent reporting about users can help provide insights into the nature of specific information ecosystems. For example, understanding high-level demographics such as age groups, gender, and location of users in a country across online services paints a picture of that information ecosystem. Tracked over time, such information could provide helpful baseline measurements for online behavior to identify patterns and changes. User-level reporting can help researchers put phenomena like disinformation into context by, for example, being able to cross-reference disclosures about the total number of people exposed to a campaign disrupted by an online service with details about overall audience sizes and compositions. However, user-level reporting must be kept broad enough to not divulge specific data about specific users. Personally identifiable data is not included in this category.

For reporting purposes, users can be broken into three broad categories: individual accounts, collections of individual accounts (such as in the context of influence operations, in which a group of accounts may share similar characteristics such as belonging to the same Facebook Group, for example, and engaged in similar behavior, including collectively reporting someone else's account for abuse to have it suspended), and organizations or businesses, each being reported separately. The distinction is important for addressing issues of scale and responsibility. For example, both a collection of individual accounts operating in tandem and a business operating as one entity are more likely to use the service to influence audiences compared to individual users, who are more likely to use the service to connect with friends or for entertainment. Reporting on user information includes broad demographic statistics, psychographic information on common interests of groups of users, services available to users, significant network formations, and activities such as types of posted content.

The user-level category also includes reporting on advertising, since advertisers often must create a user account in order to advertise on a platform. Reporting on a user's advertising activity would include details on the audiences targeted as well as the content of the advertisements themselves, to better understand whom advertisers are looking to influence, in terms of broad groups of users as audiences.

Category 2: Platform Level

Guiding Questions for Platform-Level Reporting

Definition: Platform-level reporting refers to platform architecture and how platforms work.

1. What surfaces within the platform or online service recommend content to users?
2. What is the organizing principle of these algorithms? What does each algorithm in an interlocking set optimize for, and what is the system as a whole optimizing for (e.g., engagement, time spent on a surface, content quality, clicks, something else)?
3. What internal research does the platform have about the impact and unintended consequences of the algorithm(s)? What trade-offs is the platform making in order to optimize for the goals as outlined in point 2?
    a. How is the platform mitigating the risks and externalities that were surfaced in the previously described internal research?
4. What metrics are used to measure the accuracy and impact of user-facing algorithms?
    a. In general, how do changes to user-facing recommender algorithms impact the proliferation of disinformation and harmful content, and/or incentivize bad behavior by users?
    b. What types of removed content has the algorithm optimized for?
5. What is the workflow for a user-facing recommender algorithm's development, implementation, and validation?
    a. Who is involved in the decisionmaking process on a user-facing recommender algorithm's development, implementation, and validation?
    b. What guides decisions to develop user-facing algorithms?
        i. Which teams are responsible for these decisions?
    c. How are ads treated vis-à-vis other types of posts in terms of the parameters by which they are pushed to users? What mitigations are in place to address potential adverse effects of ads being displayed in such a way that violates laws protecting marginalized groups?

Online services depend on many algorithms, and there is a limited public understanding of how these algorithms are used and developed. In turn, social media and search platforms have long faced criticism for their recommendation algorithms driving radicalization of users and promotion of harmful content.[23] To address these concerns, operational reporting in this category can help explain the kinds of algorithms used by online services, identify key user-facing algorithms (including recommendation algorithms and others), and elaborate on their intended purposes. Such reporting gives researchers the context they need to suggest new ways for testing algorithmic impact or to formulate further lines of inquiry to detect and mitigate potential biases in algorithmic design. Given the sheer number and interlocking structure of algorithms virtually all online services use to operate platforms, finding some way to narrow the scope of reporting will be key. Focusing on user-facing algorithms is one viable first step—this section examines content recommender algorithms, an even narrower category within user-facing algorithms.

In the long run, operational reporting on the platform level helps researchers, regulators, and the public better analyze how algorithms influence user behavior, and how user action, user-facing algorithms, or other factors may shape the proliferation of certain narratives. Systematized reporting can also facilitate research into aspects of platform architecture that shape the spread of false information, and how different interventions and changes in algorithmic design could effectively mitigate such effects.

For reporting purposes, online services can provide qualitative descriptions narrowing down what user-facing algorithms they have, how they are used, and the ways in which they change user behavior. In addition, this category of reporting may include details on the development process of a user-facing algorithm and pinpoint teams involved in making key decisions during the process. These details can shed light on larger, systemic issues in the ways online services deploy algorithms.

For example, whether a product development team undertakes certain types of bias auditing before releasing user-facing algorithms could affect the risks of biased outcomes. Recent research has laid out a number of different auditing mechanisms that platforms could be required to undertake and that could be leveraged in transparency reporting, including code audits, crowdsourced audits, document audits, architecture audits, automated audits, and user surveys.[24] Additionally, reporting could require platforms to explain the ways that automated systems using algorithms and human review intersect to reach final decisions about content.[25]

This category of reporting also includes questions on how ads are curated and how recommendation systems select content for different users—given past instances of platforms violating antidiscrimination laws due to algorithmic function, this is critical for truly transparent reporting.[26]

Given the centrality of algorithms in creating competitive advantages for online services, platform-level reporting will likely require careful consideration as to the degree of detail required and what audiences would have access to such information. The guiding questions section above focuses on algorithms that recommend content to users and the particular importance of transparency about the metrics for which a platform optimizes,[27] but this can be expanded to other user-facing algorithms. In particular, algorithms that underlie automated enforcement on users, such as content and account removals, should be subject to similar scrutiny, with particular attention paid to adverse effects on marginalized groups, and mitigation strategies that platforms and online services take to address those impacts.

Category 3: Policy Development and Enforcement

Guiding Questions for Policy Development and Enforcement

Definition: Policy development and enforcement refers to internal policies that govern activity and use of a platform (e.g., content moderation policies).

1. What policies are in place that define the permitted activity and use of the service by a user?
2. What is the decisionmaking process behind policy development?
    a. When and what changes are made to these policies?
    b. What triggers changes to policies?
    c. To whom are these policy changes communicated (e.g., public vs. user base vs. individual user)?
    d. Include an organizational map that outlines who makes policy decisions at what time during policy creation and implementation.
3. What audits and risk assessments have been made on which policies?
4. How frequently are audits/risk assessments carried out?
    a. Are they internal or undertaken by an external third party?
5. What languages are policies and enforcement decisions posted in?
6. How are policy development and enforcement mechanisms applicable to different types of users?
7. How is freedom of expression safeguarded?
    a. How do company policies define, enumerate, and adjudicate freedom of expression?

General Policy Enforcement

1. Which team(s) is responsible for enforcing policies?
2. What types of expertise are applied in enforcement decisions?
    a. Are third parties involved in enforcement?
3. What is the internal workflow for enforcement?
    a. How is automation involved in policy enforcement?
4. What tools are used to enforce policies (e.g., takedowns, specific adjustments to a product)?
5. On what basis is enforcement initiated (e.g., internal mechanism)?
6. What are the rates of enforcement against all policies?
7. What are the rates of false positives and negatives in enforcement decisions?
8. What metrics and analyses are used to measure policy and enforcement impacts?
    a. What were the results of analyses of policy and enforcement impacts?
9. Are there any complaint and appeal mechanisms associated with specific policies?

Content Moderation

1. What is the process for content moderation?
    a. Is content moderation executed in house or contracted to external service providers?
    b. If external contractors moderate content, where are the contractors located?
    c. How much content moderation is human, how much is controlled by automation, and how much is hybrid?
2. What languages are content moderators fluent in?
3. How is data secured when shared with content moderators?
4. What policies are in place to protect content moderators exposed to graphic or violent content?
5. What metrics and analytics are used to monitor content moderation operations?
6. What is the total amount spent on content governance?

Operational reporting on policy development and enforcement provides insight into how online services formulate, communicate, and enforce their policies on content moderation, community guidelines, content monetization, and others. Reporting in this category includes qualitative explanations of the decisionmaking processes behind policy development and enforcement processes, including on content moderation, with a mix of quantitative and qualitative data on those processes and enforcement rates.

Many online services already share publicly available information on their policies. However, this information is not always presented in a formal report—rather, it is often released in different locations (such as blog posts, press releases, or updates to public-facing company policy pages), and addresses different questions. Often these posts do not include timestamps or revision histories when companies change their policies, making it difficult for researchers to track how changes occur over time.

Standardized reporting in this category allows researchers to easily compare company policies across online services and get information on aspects of policy enforcement that are not currently reported on. Most online services have different policies and enforcement mechanisms across types of users—take for example Meta's controversial "cross check" program, which exempts certain high-profile users from the type of platform enforcement action regular users receive.[28] Having further insight on what these policies are and how they are enforced may illuminate when and how key political figures are protected from standard enforcement when they violate policies, and the extent to which online services take adequate action to address escalating political violence incited online. Tracked over time, this provides detailed recordkeeping of different approaches taken by online services in policy development and enforcement. This can encourage the creation of best practices for policy development and inform ways to measure the impact of those interventions.

In terms of reporting, this category includes questions on existing policies, as well as on the decisionmaking process behind policy development and the audits and risk assessments that have been conducted on those policies. Reporting on this category could include organizational charts outlining the teams within the company that create policies and implement them. This provides insight into how change occurs within each company and where conflicts of interest may lie. This category also includes reporting on general policy enforcement and includes questions on teams responsible for enforcing policies, the tools used to enforce said policies, rates of enforcement, and more. Content moderation processes also fall under this category, looking at existing processes, the language fluency of moderators, and the analytics used to monitor content moderation operations.

Reporting on this category must recognize limits on the release of sensitive information. For example, while information on the language fluency of content moderators can provide insight into whether platforms have dedicated resources proportionately to their user bases across languages and geographies, certain forms of reporting could potentially deanonymize the information of moderators in authoritarian countries. Conducting reporting on a regional level, as opposed to on a country-by-country basis, could help mitigate these concerns.

Category 4: Internal Research

Guiding Questions for Internal Research

Definition: Internal research refers to the types and findings of research conducted inside the company to understand impacts on users and of interventions.

1. What internal studies have assessed the platform's impact, policies, or interventions?
2. What teams conduct internal research?
   a. Who is leading the research?
   b. How is the internal research process organized within the company?
   c. How does the company define research (e.g., does the company refer to A/B testing by a name other than "research")?
3. What is the decisionmaking process behind undertaking and designing internal research?
   a. What ethical framework is applied in the design and execution of internal research?
4. How are users informed about internal research?
5. What are the findings of internal research?
   a. How are the findings of internal research used?
   b. How are the findings of internal research shared?
      i. Are there any studies that can be published for peer review?
   c. Exclude internal research used to increase market competitiveness.

Operational reporting on internal research sheds light on the types and findings of analysis conducted inside an online service. Reporting in this category consists of qualitative reports on what internal research is being conducted, which teams are involved in the process, and how findings are used to inform aspects of operations and disseminated internally and externally.

Many online services conduct a wide breadth of internal research focused on user behavior and user experience. Operational reporting in this category primarily focuses on the research conducted internally that measures how platforms impact users. For example, this can include internal research on the impacts of interventions or research on whether certain incentives to boost engagement have increased the proliferation of misinformation. Using this reporting will help inform researchers on what types of measurement research can be conducted based on the data platforms are able to or choose to collect, and conversely, reveal research gaps. Product testing research would be excluded as part of this reporting due to valid concerns about trade secrets. Internal research might be an operational reporting category that requires careful consideration of what types of audiences are able to access such information given trade secrets. At the same time, reporting on internal research is also key for rebuilding trust of online service providers with users in the wake of past disclosures emerging from employee leaks.[29]

For this category of reporting, online services can provide descriptive context on questions such as the specific teams that conduct research, how internal research processes are organized within a company, and how companies themselves define different kinds of research. For example, some companies may not categorize A/B testing as research, and so those outside the company would not know that certain types of research are housed under other branches within the company. This information helps researchers understand which teams oversee various research areas, thereby providing the means for industry to bridge gaps with academia and civil society. In the long run, these requests for operational reporting on internal research can pave a path for companies to publish internal studies for peer review, which can, in turn, result in more rigorous and high-quality research about the impact of their interventions on threats like misinformation.

Category 5: External Requests for Intervention

Guiding Questions for External Requests for Interventions

Definition: External requests for interventions are requests to platforms from a third party to act on a user, an account, or a piece of content (e.g., user-flagged content, content removal requests, account suspension requests).

Official requests from government and law enforcement agencies (cases should be reported on individually)

1. What are the details of the specific internet referral unit or government department that has issued a request for intervention?
2. Are users able to request information about their government's applications for user data? And if yes, what is that process?
3. In which countries are platform-provided products and services subject to government-required monitoring, blocking, content filtering, or censoring?

Unofficial requests

1. Other than governments and law enforcement, who else issues requests for intervention?
    a. Individual users
        i. What terms of service violations are available for users to report (e.g., abuse, hateful conduct, self-harm)?
    b. Users en masse
        i. What class of users are being targeted?
        ii. What are users being reported on?
        iii. What is the volume of reports and how has that changed over time?

General information on both types of requests

1. What types of interventions have been requested by which type of requester?
2. Against what policy or legal basis was the request made?
3. Where is the requester located?
4. What language does the request pertain to?
5. What is the nature of the intervention (e.g., content removal)?
6. Why is the requestor requesting an intervention?
7. What percent of requests are complied with and what percent are rejected?
8. What percent of requests were appealed?
9. What percent of requests were reversed?

Information on request process

1. What teams are involved in making decisions about requests?
2. How is each type of requester informed about decisions about their requests?
3. What is the appeals process around requests?
4. What teams are involved in making decisions about appeals?

Operational reporting on external requests for intervention sheds light on which entities approach online services with requests about content, users, or capabilities; why they are doing so; and how companies respond to these requests. Reporting in this category includes qualitative information on the process of requesting interventions and quantitative reporting on trends in how these requests are handled.

While many online services already share information about their processes around external requests for interventions, this reporting tends to be limited to official requests from governments or law enforcement agencies. This reporting category can be expanded to include a separate category around unofficial intervention requests. This distinction matters as official requesters have more power and a stronger ability to abuse the system, with various implications for democracy and human rights. For example, a government requesting takedowns of activists and journalists can take platforms to court or threaten political ramifications if platforms do not cooperate. However, unofficial requests for intervention, such as mass individual user reporting as part of influence operations, or requests from powerful public figures outside official government channels, can be just as impactful and should be considered as part of a platform's overall response to external intervention. Reporting on official requests for intervention from government and law enforcement agencies should be conducted on a case-by-case basis, as these requests tend to come at a much lower volume and have a potentially higher impact than unofficial requests. This should include details on the specific government departments that have requested an intervention and the nature of the requests, such as content removal, monitoring, blocking, content filtering or censoring, and the duration of the request (e.g., one time, ongoing). These questions help shed light on the relationships between governments and law enforcement on the one hand, and online services on the other, helping researchers study human rights and democracy in the context of the information environment.

Online services may face certain obstacles in reporting on official requests. Gag orders may limit their ability to provide information on specific appeals. Similarly, when governments conduct appeals through legal means instead of through the company, there will be limits on what online services can report. Regulation governing operational reporting should be introduced to address these loopholes. Similarly, transparency reporting by democratic governments should also incorporate their interactions with online service providers.

Reporting on unofficial intervention requests should be aggregated and anonymized, and consist of two categories: requests from individual users and requests from users en masse. Aggregated reporting on intervention requests from individual users could include the frequency at which individual users report policy violations as part of their requests for intervention and the number of addressed requests. Such reporting would also cover the available options for users to report policy violations, which differ across online services. For example, the category that a user picks for a gender-based death threat may differ depending on the company's classifications. Operational reporting here examines whether existing categories for reporting are well defined, and whether there continue to be gaps or misclassifications that limit users' abilities to report their specific experiences.

Reporting from users en masse involves cases where a significant number of individual users make similar reports within a short time span. For example, the Syrian regime used Facebook reporting tools to have opposition activities blocked from the platform.[30] Reporting in this category would be conducted on a case-by-case basis as requests for intervention arise, covering the type of users targeted, volume of reporting, and decisions made about addressing the issue inside the online service.

Both official and unofficial requests will also address a series of more general questions, including, for example, the types of interventions requested, the policy or legal basis on which the request was made, the location of the requester, and language of the content. This reporting maps out trends in behavior across different countries and information ecosystems, and provides useful baselines for understanding how different users interact with intervention tools. In the long term, this can also inform how reporting tools could be studied to assess their impact on the information environment.

Reporting on intervention requests should also cover how platforms respond to external requests, such as which teams are involved in what steps along the process and how decisions are communicated to users.

Category 6: Data Access Requests and Tooling

Guiding Questions for Data Access Requests and Tooling

Definition: Data access requests and tooling are requests from third parties for access to the personal information of a user or group of users (e.g., law enforcement requests, court orders) and the tools created to facilitate access to data.

Official requests from governments and law enforcement agencies

1. Who requests data access?
2. Is law enforcement required to have a warrant before data is shared with them?
3. What team(s) respond to requests by governments and law enforcement agencies for data?

Unofficial requests

1. Who requests data access?
    a. Academics
    b. Civil society researchers
    c. Journalists
2. Are researchers prohibited from requesting certain data that could conflict with national security concerns?
3. How is data access provided on a regional level?
4. What team(s) respond to which types of requesters seeking data access?

General information on both types of requests

1. What levels of data access are available for the different types of requesters?
2. What criteria are used to decide who should have access to data?

Processing requests

1. Who reviews requests?
2. What is the capacity of the reviewers (e.g., language, cultural context)?
3. Is automation used in reviewing requests?
4. What percent of requests are complied with and what percent are rejected?

Information on specific requests

1. Where is the requesting entity located?
2. What types of data is the requesting entity requesting?
3. For what purposes and on what basis are the requests being made?

Reporting on data access requests and tooling to provide data would offer transparency on how online services engage the research community. This reporting category includes a mix of qualitative and quantitative information, looking into how external actors ask for data, their access level, and what companies' internal processes are for handling these requests.

Similar to requests for external interventions, reporting on data access requests and tooling should be conducted separately for official and unofficial requests. Data requests from law enforcement, for example, can raise various human rights and privacy concerns. Therefore, official requests require more scrutiny and targeted questions, such as whether tech companies require warrants before handing over data to law enforcement and which teams respond to requests by governments and law enforcement agencies for data. Such reporting seeks to increase public understanding and provide accountability on how tech companies work with governments and law enforcement to handle user data and to provide safeguards against possible data misuse and abuse.

For unofficial requests for data access, aggregated operational reporting can include information on the kinds of third parties that request data access, how data access is distributed on a regional level, and what teams respond to different types of requesters. These reports should also include the descriptions of the datasets made available to requesters, qualifications for requesters to receive data (such as institutional affiliation, previously published research, or other requirements), and any conditions placed on requesters to use the data in research, such as prepublication approval or a data deletion requirement. These questions provide transparency on the equality of data access among the research community and geographies.

Category 7: Terms of Service and Privacy Policies

Guiding Questions for Terms of Service and Privacy Policies

Definition: Terms of service refers to agreements between companies and users regulating the use of the service. A privacy policy is a statement that discloses how and when a company collects, uses, and shares user information. This information is typically shared with third-party services.

1. What are the terms of service for:
    a. Using the platform?
    b. Purchasing advertising?
    c. Collection and use of personal and nonpersonal data?

2. What is the privacy policy?
3. How have terms of service agreements changed over time?
    a. What triggers a change?
    b. How are these changes documented and communicated?
4. How has the privacy policy changed over time?
    a. What triggers a change?
    b. How are these changes documented and communicated?
    c. What tools are made available to users to protect their privacy?
        i. How are these tools communicated to users?
        ii. What is the percentage of uptake on these tools by users?
5. What rights do users have?
6. Are users given an opt-in to have their data studied?
7. How long does the average user spend reading the terms of service before agreeing?
8. What languages are the terms of service and privacy policy published in?
9. How do the terms of service and privacy policy vary across legal jurisdictions, if at all?

Operational reporting on platform terms of service and privacy policies aims to foster a better understanding about data collected on users and user awareness of that data collection. This category of reporting is primarily qualitative, outlining existing terms of service and privacy policies, how they have changed over time, and how that affects user rights, safety, and privacy across languages and jurisdictions.

Reporting on terms of service and privacy policies would help users and researchers understand and compare the content of these policies, how they change over time, and how these changes are communicated. Such documentation would promote a better understanding of how platforms collect user data and respond to legislative changes. For example, following the European Union's (EU's) introduction of the General Data Protection Regulation (GDPR), some companies introduced new data protections across their entire user base, while others only made changes for users in the EU. In the long term, operational reporting for this category could encourage platforms to explain how they perceive privacy risks and how they empower end users to protect themselves, possibly encouraging the creation of norms and standards that improve protections for user privacy. BOX

Category 8: Third-Party Relationships

Guiding Questions for Third-Party Relationships

Definition: Third-party relationships refer to arrangements between companies and third-party organizations regularly accessing user data.

1. Which third parties have ongoing or routine access to data, and for what purposes:
   a. Social media listening services?
   b. Ad-tech services?
   c. Online tracking services?
   d. API/other tooling?
   e. Fact-checkers?
   f. Researchers?
   g. Investigators? For example, DFRLab or GIFCT's Hash-Sharing Database.
2. What policies are in place that govern the collection, use, and disclosure of data by third parties?
3. What international fora/multistakeholder groups are the company part of?
4. Are third parties able to match data across platforms?
   a. What are the processes for cross-platform database sharing?

Operational reporting on third-party relationships provides information on how platforms share user data with third-party services. Reporting on this level is qualitative and covers which third parties have ongoing or routine access to user data (as opposed to singular requests for specific data), and for what purpose. It examines the policies in place governing the collection, use, and disclosure of this data and whether companies are involved in coordinated cross-platform efforts to increase third parties' access to data.

Third-party services are defined as external services that platforms work with to improve platform functionality, those to whom platforms sell data, and researchers who have ongoing data access. The first category includes external fact-checkers and investigators, while the second includes social media listening services, ad-tech services, and online tracking services. Researchers who have ongoing access may also receive discrete, time-bound datasets, as outlined in category 6 above. GDPR data processors would not fall into the category of third-party services.

Operational reporting on third-party relationships provides transparency on other actors beyond the online service that handles user data. In the status quo, there is a limited understanding of which third parties have access to user data and how that data is used. Greater scrutiny, and indeed outrage, about the lack of transparency around third-party data access followed the revelations of Cambridge Analytica's access to Facebook data during the 2016 U.S. presidential election.[31] Operational reporting on third-party access to data can address these concerns and help researchers understand what data nonresearcher third parties are able to access and could theoretically be made available for research purposes to make requests through the EU's Digital Services Act.

The third-party relationships category includes reporting on which third parties have ongoing or routine access to data, and the policies that govern the collection, use, and disclosure of this data. Beyond that, online services could also include information on the international fora and multistakeholder initiatives they are a part of and the processes through which they make data available in those relationships.

Next Steps

High-level harmonization of transparency reporting requirements across countries increases the likelihood that companies will adopt the practice, be it by law or voluntarily. A harmonized standard, adopted by enough countries, could benefit even smaller countries without their own law or regulation, so long as the standard is designed to facilitate this.

This paper is a beginning. It sets out eight categories for a broad reporting framework on the operations of online services, namely: (1) user level, (2) platform level, (3) policy development and enforcement, (4) internal research, (5) external requests for intervention, (6) data access request and tooling, (7) terms of service and privacy policies, and (8) third-party relationships. However, how that reporting would function in practice requires more details, which must still be worked out. Key questions include: Who should have access to what type of reporting, which categories of reporting should be prioritized, and how could each category of reporting be independently audited to assess the veracity of claims made?

Who should have access to what type of reporting?

Given the sensitivity around certain aspects of this reporting, more work must be done to determine what information should be made publicly available and what should only be provided to a limited audience. This will be integral in setting up a comprehensive and achievable set of reporting processes that can be used to improve our collective understanding of the information environment and the challenges within.

Different audiences should have different levels of access to operational reports. The general public will not require the same level of granularity as, say, regulators using operational reporting to assess harms to users. Reporting will also include sensitive information, which could be misused. For example, detailed information on internal company decisionmaking regarding adversarial actors could be leveraged to circumvent platform safeguards. Similarly, in authoritarian countries where there are few journalists and researchers reporting on the ruling regime, disclosing that data access is being granted to any researchers at all from that country could raise risks to their personal safety.

Therefore, more work must be done to determine what aspects of operational reporting should be made available to different audiences and under what circumstances. Such efforts might include conducting red team exercises playing through different scenarios around access to operational reporting to consider the consequences. Whatever approach is taken, considerations might include access benefits for different stakeholders, privacy, the potential for abuse, and more. A closer examination of possible safeguards against these risks, such as reporting on a regional level instead of a country-by-country level to prevent deanonymization, can also help mitigate specific concerns.

Which categories of reporting should be prioritized?

It's unlikely that any single online service will be able to introduce reporting across all these categories simultaneously. Thus, some prioritization of rollout will need to be made, ideally by a multistakeholder forum such as the Action Coalition on Meaningful Transparency, which aims to "bring together a wide range of academics, civil society organizations, companies, governments, and international organizations to work collaboratively on digital transparency."[32]

How can reporting be trusted?

Building onto this framework for operational reporting, an independent auditing process can also be established to ensure platform accountability and public trust. Audits can be used to assess the performance and accuracy of this reporting while also regularly reviewing and updating the processes involved. Further study will also be required to understand how these audits can be governed, structured, and reported on.

Annex 1: Eight Categories for Operational Reporting

Reporting Category	Reporting Details	Quantitative/ Qualitative
User level: aggregated information about different types of users.	1. Segment reports by user category: a. Individual accounts b. Collections of individual accounts c. Organizations/businesses	Quantitative
	2. Segment reports on user information by user category. Include information on: a. Demographics (e.g., age, gender, location) b. Psychographics (e.g., interests)	Quantitative
	3. Segment reports on user activity by user category. Include information on: a. Types of content of public posts, comments, and engagement b. Posting patterns, such as where users post and share c. Networks users form d. Users that purchase ads: i. The types of ads purchased by which types of users ii. Whom they target ii. In what languages and geographies, and when	Quantitative
	4. What services are available?	Qualitative
	5. What can users do with those services?	Qualitative
	6. What languages are those services offered in?	Qualitative
	7. As platforms grow their user base, what is their pathway of progression in terms of expanding available languages?	Qualitative

Reporting Category	Reporting Details	Quantitative/ Qualitative
Platform level: platform architecture and how platforms work.	1. What surfaces within the platform or online service recommend content to users?	Qualitative
	2. What is the organizing principle of platform algorithms? What does each algorithm in an interlocking set optimize for, and what is the system as a whole optimizing for (e.g., engagement, time spent on a surface, content quality, clicks, something else)?	Qualitative
	3. What internal research does the platform have about the impact and unintended consequences of the algorithm(s)? What trade-offs is the platform making in order to optimize for the goals as outlined in point 2? a. How is the platform mitigating the risks and externalities that were surfaced in the previously described internal research?	Qualitative
	4. What metrics are used to measure the accuracy and impact of user-facing algorithms? a. In general, how do changes to user-facing recommender algorithms impact the proliferation of disinformation and harmful content and/or incentivize bad behavior by users? b. What types of removed content has the algorithm optimized for?	Qualitative
	5. What is the workflow for a user-facing recommender algorithm's development, implementation, and validation? a. Who is involved in the decisionmaking process on a user-facing recommender algorithm's development, implementation, and validation? b. What guides decisions to develop user-facing algorithms? i. Which teams are responsible for these decisions? c. How are ads treated vis-à-vis other types of posts in terms of the parameters by which they are pushed to users? What mitigations are in place to address potential adverse effects of ads being displayed in such a way that violates laws protecting marginalized groups?	Qualitative

Reporting Category	Reporting Details	Quantitative/ Qualitative
Policy development and enforcement: internal policies that govern activity and use of the platform (e.g., content moderation policies).	1. What policies are in place that define the permitted activity and use of the service by a user?	Qualitative
	2. What is the decisionmaking process behind policy development? a. When and what changes are made to these policies? b. What triggers changes to policies? c. To whom are these policy changes communicated (e.g., public vs. user base vs. individual user)? d. Include an organizational map that outlines who makes policy decisions at what time during policy creation and implementation.	Qualitative
	3. What audits and risk assessments have been made on which policies?	Qualitative
	4. How frequently are audits/risk assessments carried out? a. Are they internal or undertaken by an external third party?	Qualitative
	5. What languages are policies and enforcement decisions posted in?	Qualitative
	6. How are policy development and enforcement mechanisms applicable to different types of users?	Qualitative
	7. How is freedom of expression safeguarded? a. How do company policies define, enumerate, and adjudicate freedom of expression?	Qualitative
	General policy enforcement	
	1. Which team(s) is responsible for enforcing policies?	Qualitative
	2. What types of expertise are applied in enforcement decisions? a. Are third parties involved in enforcement?	Qualitative
	3. What is the internal workflow for enforcement? a. How is automation involved in policy enforcement?	Qualitative
	4. What tools are used to enforce policies (e.g., takedowns, specific adjustments to a product)?	Qualitative
	5. On what basis is enforcement initiated (e.g., internal mechanism)?	Qualitative
	6. What are the rates of enforcement against all policies?	Quantitative
	7. What are the rates of false positives and negatives in enforcement decisions?	Quantitative
	8. What metrics and analyses are used to measure policy and enforcement impacts? a. What were the results of analyses of policy and enforcement impacts?	Qualitative
	9. Are there any complaint and appeal mechanisms associated with specific policies?	Qualitative

Reporting Category	Reporting Details	Quantitative/ Qualitative
Policy development and enforcement cont.	**Content moderation**	
	1. What is the process for content moderation? a. Is content moderation executed in house or contracted to external service providers? b. If external contractors moderate content, where are the contractors located? c. How much content moderation is human, how much is controlled by automation, and how much is hybrid?	Qualitative
	2. What languages are content moderators fluent in?	Qualitative
	3. How is data secured when shared with content moderators?	Qualitative
	4. What policies are in place to protect content moderators exposed to graphic or violent content?	Qualitative
	5. What metrics and analytics are used to monitor content moderation operations?	Qualitative
	6. What is the total amount spent on content governance?	Quantitative
Internal research: on the types and findings of research conducted inside the company to understand impacts on users and of interventions.	1. What internal studies have assessed the platform's impact, policies, or interventions?	Qualitative
	2. What teams conduct internal research? a. Who is leading the research? b. How is the internal research process organized within the company? c. How does the company efine research (e.g., does the company refer to A/B testing by a name other than "research")?	Qualitative
	3. What is the decisionmaking process behind undertaking and designing internal research? a. What ethical framework is applied in the design and execution of internal research?	Qualitative
	4. How are users informed about internal research?	Qualitative
	5. What are the findings of internal research? a. How are the findings of internal research used? b. How are the findings of internal research shared? i. Are there any studies that can be published for peer review? c. Exclude internal research used to increase market competitiveness.	Qualitative

Reporting Category	Reporting Details	Quantitative/ Qualitative
External requests for interventions: requests to platforms from a third party to act on a user, an account, or a piece of content (e.g., user-flagged content, content removal requests, account suspension requests).	**Official requests from government and law enforcement agencies (cases should be reported on individually)**	
	1. What are the details of the specific internet referral unit or government department that has issued a request for intervention?	Qualitative
	2. Are users able to request information about their government's applications for user data? And if yes, what is that process?	Qualitative
	3. In which countries are platform-provided products and services subject to government-required monitoring, blocking, content filtering, or censoring?	Qualitative
	Unofficial requests	
	1. Other than governments and law enforcement, who else issues requests for intervention? a. Individual users i. What terms of service violations are available for users to report (e.g., abuse, hateful conduct, self-harm)? b. Users en masse i. What class of users are being targeted? ii. What are users being reported on? iii. What is the volume of reports and how has that changed over time?	Qualitative
	General information on both types of requests	
	1. What types of interventions have been requested by which type of requester?	Quantitative
	2. Against what policy or legal basis was the request made?	Qualitative
	3. Where is the requester located?	Qualitative
	4. What language does the request pertain to?	Qualitative
	5. What is the nature of the intervention (e.g., content removal)?	Qualitative
	6. Why is the requestor requesting an intervention?	Qualitative
	7. What percent of requests are complied with and what percent are rejected?	Quantitative
	8. What percent of requests were appealed?	Quantitative
	9. What percent of requests were reversed?	Quantitative

Reporting Category	Reporting Details		Quantitative/ Qualitative
External requests for interventions cont.	**Information on request process**		
	1.	What teams are involved in making decisions about requests?	Qualitative
	2.	How is each type of requester informed about decisions about their requests?	Qualitative
	3.	What is the appeals process around requests?	Qualitative
	4.	What teams are involved in making decisions about appeals?	Qualitative
Data access requests and tooling: requests from third parties for access to the personal information of a user or group of users (e.g., law enforcement requests, court orders) and the tools created to facilitate access to data.	**Official requests from governments and law enforcement agencies**		
	1.	Who requests data access?	Qualitative
	2.	Is law enforcement required to have a warrant before data is shared with them?	Qualitative
	3.	What team(s) respond to requests by governments and law enforcement agencies for data?	Qualitative
	Unofficial requests		
	1.	Who requests data access? a. Academics b. Civil society researchers c. Journalists	Qualitative
	2.	Are researchers prohibited from requesting certain data that could conflict with national security concerns?	Qualitative
	3.	How is data access provided on a regional level?	Qualitative
	4.	What team(s) respond to which types of requesters seeking data access?	Qualitative
	General information on both types of requests		
	1.	What levels of data access are available for the different types of requesters?	Qualitative
	2.	What criteria are used to decide who should have access to data?	Qualitative

Reporting Category	Reporting Details	Quantitative/ Qualitative
Data access requests and tooling cont.	**Processing requests**	
	1. Who reviews requests?	Qualitative
	2. What is the capacity of the reviewers (e.g., language, cultural context)?	Qualitative
	3. Is automation used in reviewing requests?	Qualitative
	4. What percent of requests are complied with and what percent are rejected?	Quantitative
	Information on specific requests	
	1. Where is the requesting entity located?	Qualitative
	2. What types of data is the requesting entity requesting?	Qualitative
	3. For what purposes and on what basis are the requests being made?	Qualitative
Terms of service and privacy policies: Terms of service refers to agreements between companies and users regulating the use of the service. A privacy policy is a statement that discloses how and when a company collects, uses, and shares user information. This information is typically shared with third-party services.	1. What are the terms of service for: a. Using the platform? b. Purchasing advertising? c. Collection and use of personal and nonpersonal data?	Qualitative
	2. What is the privacy policy?	Qualitative
	3. How have terms of service agreements changed over time? a. What triggers a change? b. How are these changes documented and communicated?	Qualitative
	4. How has the privacy policy changed over time? a. What triggers a change? b. How are these changes documented and communicated? c. What tools are made available to users to protect their privacy? i. How are these tools communicated to users? ii. What is the percentage of uptake on these tools by users?	Qualitative
	5. What rights do users have?	Qualitative
	6. Are users given an opt-in to have their data studied?	Qualitative
	7. How long does the average user spend reading the terms of service before agreeing?	Quantitative
	8. What languages are the terms of service and privacy policy published in?	Qualitative
	9. How do the terms of service and privacy policy vary across legal jurisdictions, if at all?	Qualitative

Reporting Category	Reporting Details	Quantitative/ Qualitative
Third-party relationships: arrangements between companies and third-party organizations regularly accessing user data.	1. Which third parties have ongoing or routine access to data, and for what purposes: a. Social media listening services? b. Ad-tech services? c. Online tracking services? d. API/other tooling? e. Fact-checkers? f. Researchers? g. Investigators? For example, DFRLab or GIFCT's Hash-Sharing Database.	Qualitative
	2. What policies are in place that govern the collection, use, and disclosure of data by third parties?	Qualitative
	3. What international fora/multistakeholder groups are the company part of?	Qualitative
	4. Are third parties able to match data across platforms? a. What are the processes for cross-platform database sharing?	Qualitative
Auditing: an additional step whereby an internal or external body assesses information, policies, and practices for verification purposes. Depending on the focus, audits can include benchmarking against defined standards or requirements.	**Governance**	
	1. How are audits of transparency reporting or data-sharing initiatives governed?	Qualitative
	2. What is the board-level commitment to auditing?	Qualitative
	3. What is the executive responsibility for the implementation of audits? a. What are the structures in place for the oversight of auditing?	Qualitative
	4. What resources (financial and human) are dedicated to auditing, including size, roles, diversity, and team capacity?	Qualitative
	Performance	
	1. How is performance assessed during audits? a. In general, how does transparency reporting perform against internal and external objectives and metrics during auditing?	Qualitative
	2. How accurate and accessible is the information being reported or shared during audits?	Qualitative

About the Authors

Samantha Lai is a research analyst with the Partnership for Countering Influence Operations at the Carnegie Endowment for International Peace. Prior to joining Carnegie, Lai worked at the Brookings Institution's Center for Technology Innovation, the Foreign Policy Research Institute, and a boutique threat intelligence consulting firm. Her work has been featured by NPR, Lawfare, TechTank, and more. She holds a BA in political science from Wellesley College.

Naomi Shiffman is the head of data and implementation for the Meta Oversight Board, where she leads a team in assessing the Board's impact on Meta's content ecosystem. Naomi previously built the academic and research program at CrowdTangle, a Meta product. Before her work with CrowdTangle, Naomi was a policy researcher at Mozilla, focused on privacy policy, data protection, AI accountability, and misinformation. She is a nonresident fellow at the Atlantic Council's Digital Forensic Research Lab, a founding fellow at the Integrity Institute, and an adviser to Connect Humanity, a fund for digital equity. Naomi has a master's in public policy from UC Berkeley and a bachelor's of science from UC San Diego.

Alicia Wanless is the director of the Partnership for Countering Influence Operations at the Carnegie Endowment for International Peace, which aims to foster evidence-based policymaking for the governance of the information environment. In partnership with Princeton University, Alicia is developing the Institute for Research on the Information

Environment, a multinational, multistakeholder research facility. Alicia created a multistakeholder network in partnership with the G7 Rapid Response Network to support information integrity efforts in Ukraine. At King's College London in War Studies, she completed her PhD combining strategic theory and ecology in a new approach to understanding conflict within the information environment.

Acknowledgments

The authors would like to thank the contributions to this piece from Abrianna Rhodes, Adnan Raja, Benjamin Wittes, Collin Kurre, Daphne Keller, David Sullivan, Elonnai Hickok, Erin Saltman, Graham Brookie, Heidi Tworek, Jen Rosiere Reynolds, Jeremy West, Kamya Yadav, Laura Smillie, Lutz Guellner, Nicole Matejic, Rose Jackson, and members of the working group.

Notes

1 Mark MacCarthy, "Transparency is The Best First Step Towards Better Digital Governance," Centre for International Governance Innovation, May 5, 2022, https://www.cigionline.org/articles/transparency-is-the-best-first-step-to-better-digital-governance/; Aleksandra Urman and Mykola Makhortykh, "How Transparent Are Transparency Reports? Comparative Analysis of Transparency Reporting Across Online Platforms," *Telecommunications Policy* 47, no. 3 (April 2023): 102477, https://doi.org/10.1016/j.telpol.2022.102477; "Parents Want More Transparency From Tech Companies," ADL, May 31, 2022, https://www.adl.org/resources/report/parents-want-more-transparency-tech-companies; Andrew Puddephatt, "Letting the Sun Shine In: Transparency and Accountability in the Digital Age," UNESCO, 2021, https://unesdoc.unesco.org/ark:/48223/pf0000377231.

2 For example, in 2021, Tech Against Terrorism published "Guidelines on Transparency Reporting on Online Counterterrorism Efforts," https://transparency.techagainstterrorism.org. In 2022, the OECD published "Transparency Reporting on Terrorist and Violent Extremist Content Online 2022," OECD Digital Economy Papers, No. 334, https://doi.org/10.1787/a1621fc3-en. The Santa Clara Principles define best practices for transparency reporting of content moderation efforts. See "Santa Clara Principles on Transparency and Accountability in Content Moderation," https://santaclaraprinciples.org/. The EU Code of Practice on Disinformation commits signatories to public disclosure of political advertising. See: "2022 Strengthened Code of Practice on Disinformation," European Commission, June 16, 2022, https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation. The U.S. Social Media Disclosure and Transparency of Advertisements Act requires companies to establish ad libraries accessible to researchers. See: "Section-by-Section: Social Media DATA Act of 2021," Office of Congresswoman Lori Trahan, 2021, accessed April 27, 2023, https://trahan.house.gov/uploadedfiles/social_media_data_act_section_by_section.pdf. Also see: Social Media DATA Act, H.R.3451, 117th Cong. (2021), https://www.congress.gov/bill/117th-congress/house-bill/3451/text. In 2022, the European Parliament drafted a proposal laying down rules to prevent and combat child sexual abuse, which requires companies to proactively detect, report, and remove child sexual abuse content. See: "Proposal for a Regulation of the European Parliament and of the Council Laying Down Rules to Prevent and Combat Child Sexual Abuse," European Commission, COM/2022/209 final, May 11, 2022, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2022:209:FIN. In October 2022, the Tech Coalition published

"Trust: Voluntary Framework for Industry Transparency," which provides guidelines for tech companies to report on child sexual abuse material (CSAM) risks on their services. See: "TRUST: Transparency Reporting Implementation Guide," Technology Coalition, October 2022, https://paragonn-cdn.nyc3.cdn.digitaloceanspaces.com/technologycoalition.org/uploads/Tech_Coalition_Trust_Implementation_Guide_FINAL-1.pdf.

[3] For example, the German Network Enforcement Act requires social network providers that meet a certain threshold to publish biannual transparency reports. For more see: "Act to Improve Enforcement of the Law in Social Networks (Network Enforcement Act)," German Federal Ministry of Justice and Consumer Protection, July 7, 2017, https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf?__blob=publicationFile&v=2.

[4] "Transparency Reporting & Data Sharing," Partnership for Countering Influence Operations, accessed on April 27, 2023, https://ceip.knack.com/pcio-baseline-datasets#transparency--data-sharing/?view_69_page=1.

[5] Heidi Tworek and Alicia Wanless, "Time for Transparency From Digital Platforms, But What Does That Really Mean?," Lawfare, January 20, 2022, https://www.lawfareblog.com/time-transparency-digital-platforms-what-does-really-mean.

[6] Oversight Board, "2021 Annual Report," Meta, June 22, 2022, https://oversightboard.com/attachment/425761232707664/.

[7] Transparency Center, "Facebook Community Standards," Meta, accessed April 4, 2023, https://transparency.fb.com/policies/community-standards/; Transparency Center, "Hate Speech," Meta, accessed April 4, 2023, https://transparency.fb.com/policies/community-standards/hate-speech/.

[8] Alicia Wanless and Kamya Yadav, "What Do Transparency and Data Sharing Really Mean?," Lawfare, June 9, 2022, https://www.lawfareblog.com/what-do-transparency-and-data-sharing-really-mean.

[9] Sarah St. Vincent, "Human Rights and Surveillance: Governments Must Comply With Their Transparency Obligations," Center for Democracy and Technology, June 20, 2014, https://cdt.org/insights/human-rights-and-surveillance-governments-must-comply-with-their-transparency-obligations/; Daphne Keller, "Some Humility About Transparency," Center for Internet and Society at Stanford Law School, March 19, 2021, https://cyberlaw.stanford.edu/blog/2021/03/some-humility-about-transparency.

[10] Puddephatt, "Letting the Sun Shine In."

[11] "Information Requests," Twitter Transparency Center, accessed February 16, 2023, https://transparency.twitter.com/en/reports/information-requests.html; Transparency Center, "Government Requests for User Data," Meta, accessed February 16, 2023, https://transparency.fb.com/data/government-data-requests; "Law Enforcement Requests Report," Microsoft, Microsoft, accessed February 16, 2023, https://www.microsoft.com/en-us/corporate-responsibility/law-enforcement-requests-report; "Government Requests Report," LinkedIn, accessed February 16, 2023, https://about.linkedin.com/transparency/government-requests-report.

[12] See: "Removal Requests," Twitter Transparency Center, accessed February 16, 2023, https://transparency.twitter.com/en/reports/rules-enforcement.html#2:2021-jul-dec; Transparency Center, "Content Restrictions Based on Local Law," Meta, accessed February 16, 2023, https://transparency.fb.com/data/content-restrictions; "Reports Hub,", Microsoft, accessed February 16, 2023, https://www.microsoft.com/en-us/corporate-responsibility/reports-hub?activetab=pivot_1:primaryr3; "Transparency Report 2020," Reddit, January 2020, accessed February 16, 2023, https://www.redditinc.com/policies/transparency-report-2020-1.

[13] See: "Rules Enforcement," Twitter Transparency Center, accessed February 16, 2023, https://transparency.twitter.com/en/reports/rules-enforcement.html#2:2021-jul-dec; Transparency Center, "Transparency Reports," Meta, accessed February 16, 2023, https://transparency.fb.com/data/; "YouTube Community Guidelines Enforcement," Google Transparency Report, accessed February 16, 2023, https://transparencyreport.google.com/youtube-policy/removals?hl=en.

[14] See: "March 2021 Coordinated Inauthentic Behavior Report," Meta, April 6, 2021, https://about.fb.com/news/2021/04/march-2021-coordinated-inauthentic-behavior-report/; "Twitter Moderation Research Consortium," Twitter Transparency Center, accessed February 16, 2023, https://transparency.twitter.com/en/reports/moderation-research.html.

15 See: Transparency Center, "Internet Disruptions," Meta, accessed February 16, 2023, https://transparency.fb.com/data/internet-disruptions.

16 See: "Transparency Report, Government Requests," Apple, accessed February 16, 2023, https://www.apple.com/legal/transparency/.

17 See: Anna Stepanov, "Introducing the Widely Viewed Content Report," Facebook, August 18, 2021, https://about.fb.com/news/2021/08/widely-viewed-content-report/.

18 "TikTok Transparency Center," TikTok, accessed February 16, 2023, https://www.tiktok.com/transparency/en-us/; "Meta Transparency Center," Meta, accessed February 16, 2023, https://transparency.fb.com/; "Twitter Transparency Center," Twitter, accessed February 16, 2023, https://transparency.twitter.com/en.html; "Google Transparency Report," Google, accessed February 16, 2023, https://transparencyreport.google.com/?hl=en.

19 "2022 Strengthened Code of Practice on Disinformation," European Commission, June 16, 2022, https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation; "Voluntary Transparency Reporting Framework pilot for terrorist and violent extremist content," VTRF Pilot, OECD, accessed February 16, 2023, https://www.oecd-vtrf-pilot.org/.

20 "Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC," European Union, December 15, 2020, https://eur-lex.europa.eu/legal-content/en/ALL/?uri=COM:2020:825:FIN.

21 "Online Safety Bill," UK Parliament, last accessed February 16, 2023, https://bills.parliament.uk/publications/49376/documents/2822.

22 European Union, "Proposal for a Regulation of the European Parliament"; Congress.gov, PACT Act, S. 797, 117th Cong. (2021-2022), https://www.congress.gov/bill/117th-congress/senate-bill/797/text, Algorithmic Justice and Online Platform Transparency Act, H.R. 3611, 117th Cong. (2021-2022), https://www.congress.gov/bill/117th-congress/house-bill/3611/text; Congress.gov, Digital Services Oversight and Safety Act of 2022, H.R. 6796, 117th Cong. (2021-2022), https://www.congress.gov/bill/117th-congress/house-bill/6796/text#toc-H145117D5B12A4555B2BECEADDEC6BF78; Congress.gov, Platform Accountability and Transparency Act, S.5339, 117th Congress (2021-2022), https://www.congress.gov/bill/117th-congress/senate-bill/5339.

23 Megan A. Brown et al., "Echo Chambers, Rabbit Holes, and Ideological Bias: How YouTube Recommends Content to Real Users," Brookings Institution, October 13, 2022, https://www.brookings.edu/research/echo-chambers-rabbit-holes-and-ideological-bias-how-youtube-recommends-content-to-real-users/; Karen Kornbluh, "Disinformation, Radicalization, and Algorithmic Amplification: What Steps Can Congress Take?," Just Security, February 7, 2022, https://www.justsecurity.org/79995/disinformation-radicalization-and-algorithmic-amplification-what-steps-can-congress-take/; Brandy Zadrozny, "'Carol's Journey': What Facebook Knew About How It Radicalized Users," NBC News, October 22, 2021, https://www.nbcnews.com/tech/tech-news/facebook-knew-radicalized-users-rcna3581.

24 Anna-Katharina Meßmer and Martin Degeling, "Auditing Recommender Systems," Stiftung Neue Verantwortung, February 7, 2023, https://www.stiftung-nv.de/de/publication/auditing-recommender-systems.

25 See: Oversight Board Cross-Check Policy Advisory Opinion, pages 14 and 21: Oversight Board, "Policy Advisory Opinion on Meta's Cross-Check Program," Meta, December 6, 2022, 14, 21, https://oversightboard.com/attachment/512630074120983/.

26 See: Fair Housing Council v. Roommates.com, 521 F.3d 1157 (9th Cir. 2008); Henderson v. The Source for Public Data, L.P., 53 F.4th 110 (4th Cir. 2022); HUD v. Facebook, Inc., FHEO No. 01-18-0323-8 (HUD Mar. 28, 2019), https://www.hud.gov/sites/dfiles/Main/documents/HUD_v_Facebook.pdf.

27 "Ranking and Design Transparency: Data, Datasets, and Reports to Track Responsible Algorithmic and Platform Design," Integrity Institute, September 28, 2021, https://static1.squarespace.com/static/614cbb3258c5c87026497577/t/617834ea6ee73c074427e415/1635267819444/Ranking+and+Design+Transparency+%28EXTERNAL%29.pdf.

[28] Jeff Horowitz, "Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt.," *Wall Street Journal*, September 13, 2021, https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353.

[29] Georgia Wells, Jeff Horowitz, and Deepa Seetharaman, "Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show," *Wall Street Journal*, September 14, 2021, https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739?mod=hp_lead_pos7&mod=article_inline.

[30] Amar Toor, "Syrian Activists Are Being Muzzled on Facebook," Verge, February 5, 2014, https://www.theverge.com/2014/2/5/5381104/syrian-opposition-activists-muzzled-on-facebook.

[31] Carole Cadwalladr and Emma Graham-Harrison, "Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach," *Guardian*, March 17, 2018, https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election; Shiona McCallum, "Meta Settles Cambridge Analytica Scandal Case for $725m," BBC, December 23, 2022, https://www.bbc.com/news/technology-64075067.

[32] See: "The Action Coalition on Meaningful Transparency," accessed March 15, 2023, https://www.meaningfultransparency.tech/.

Carnegie Endowment for International Peace

The Carnegie Endowment for International Peace is a unique global network of policy research centers around the world. Our mission, dating back more than a century, is to advance peace through analysis and development of fresh policy ideas and direct engagement and collaboration with decisionmakers in government, business, and civil society. Working together, our centers bring the inestimable benefit of multiple national viewpoints to bilateral, regional, and global issues.

Partnership for Countering Influence Operations

Influence operations are a complex threat, and the community combating them—academics, social platforms, think tanks, governments—is broad. The goal of the Partnership for Countering Influence Operations (PCIO) is to foster evidence-based policymaking to counter threats in the information environment. Key roadblocks as found in our work include the lack of: transparency reporting to inform what data is available for research purposes; rules guiding how data can be shared with researchers and for what purposes; and an international mechanism for fostering research collaboration at-scale.

CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

CarnegieEndowment.org