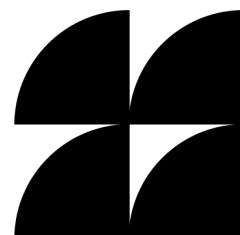


NOVEMBER 2025

Better Access: Data for the Common Good

*A Framework for Accessing
High-Influence Public Platform Data*

KGI EXPERT REPORT



KGI Expert Working Group on Public Platform Data

This framework is the product of the [KGI Expert Working Group on Public Platform Data](#). KGI convenes expert working groups (EWGs) that bring together relevant experts for a time-bound project to summarize knowledge and articulate policy options. KGI builds EWGs by inviting selected volunteer contributors from across academia, industry, civil society, journalism, and practitioner communities to engage in a focused collaboration on a specific policy topic over the course of 6-12 months or more. The goal of each EWG is to produce non-partisan resources of broad utility to policymakers and industry decision-makers. Learn more about [KGI Expert Working Groups](#).

The members of the KGI Expert Working Group on Public Platform Data who contributed to this report are:

Alex Abdo
Knight First Amendment Institute

Leticia Bode
Knight-Georgetown Institute

Angie Drobic Holan
International Fact-Checking Network

LK Seiling
DSA 40 Data Access Collaboratory

Brandon Silverman
*Knight Policy Fellow,
Formerly CrowdTangle*

Mark Scott
*Digital Forensic Research Lab,
Atlantic Council*

Cameron Hickey
National Conference on Citizenship

Nadine Farid Johnson
Knight First Amendment Institute

Daniel Arnaudo
University of Washington

Naomi Shiffman
*Meta Oversight Board,
Formerly CrowdTangle*

Jeremy Merrill
Washington Post

Nathaniel Persily
Stanford Law School

Josephine Lukito
University of Texas at Austin

Peter Chapman
Knight-Georgetown Institute

Justin Hendrix
Tech Policy Press

Rachelle Faust
Formerly National Democratic Institute

Kaitlyn Dowling
*Algorithmic Transparency Institute,
National Conference on Citizenship*

Laura Edelson
Northeastern University

Rebekah Tromble
Northeastern University

Tiago Ventura
Georgetown University

Acknowledgements

The framework was strengthened through feedback from a range of experts. The EWG would like to thank the reviewers listed below, as well as others who wished to remain unnamed:

- Anna Lenhart, The George Washington University
- Brandi Geurkink, Coalition for Independent Technology Research
- Claire Pershan, Mozilla Foundation
- Daphne Keller, Stanford Law School
- Duncan Watts, University of Pennsylvania
- Hilary Ross, Global Network Initiative
- Jerry Yu, Doublethink Lab
- Luka Bekavac, University of St. Gallen (HSG)
- Megan Iorio, Electronic Privacy Information Center
- Spence Purnell, R Street Institute

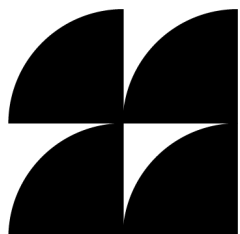
Review of the report does not constitute an endorsement of the report's conclusions or recommendations. Any errors or omissions are the responsibility of the authors.

The EWG would also like to recognize the contributions of the KGI team, including Alissa Cooper, Julie Anne Miranda-Brobeck, and Zander Arnao, as well as Patricia Ramírez, KGI's 2024-2025 graduate student research assistant.



About the Knight-Georgetown Institute

The Knight-Georgetown Institute (KGI) is dedicated to connecting independent research with technology policy and design. KGI serves as a central hub for the growing network of scholarship that seeks to shape how technology is used to produce, disseminate, and access information. KGI is designed to provide practical resources that policymakers, journalists, and private and public sector leaders can use to tackle information and technology issues in real time. Georgetown University and the Knight Foundation came together to launch the institute in 2024. Learn more about KGI at <https://kgi.georgetown.edu>.



Executive Summary

Online platforms and services shape what we know, how we connect, and who gets heard. From elections and public health to commerce and conflict, platforms are now indispensable infrastructure for civic life. Their influence is vast, and so is the need to understand them.

As critical conversations publicly unfold on digital platforms, the ability to study them at scale has steadily diminished. Platform interfaces and insight tools like CrowdTangle that once gave researchers, journalists, and civil society a window into public online discourse have disappeared.

Platforms restrict researcher access while public data is increasingly monetized for advertisers, data brokers, and training artificial intelligence (AI) systems. This imbalance – where companies profit while independent researchers are left in the dark – undermines transparency, limits free expression, and weakens oversight.

That is the reason for developing **Better Access**, a baseline framework for **independent access to public platform data**: the content, data, and information posted to platforms that anyone can access.

Published by the Knight-Georgetown Institute (KGI), this framework is the product of KGI's Expert Working Group (EWG) on Public Platform Data, a group of leaders from research, civil society, and journalism.

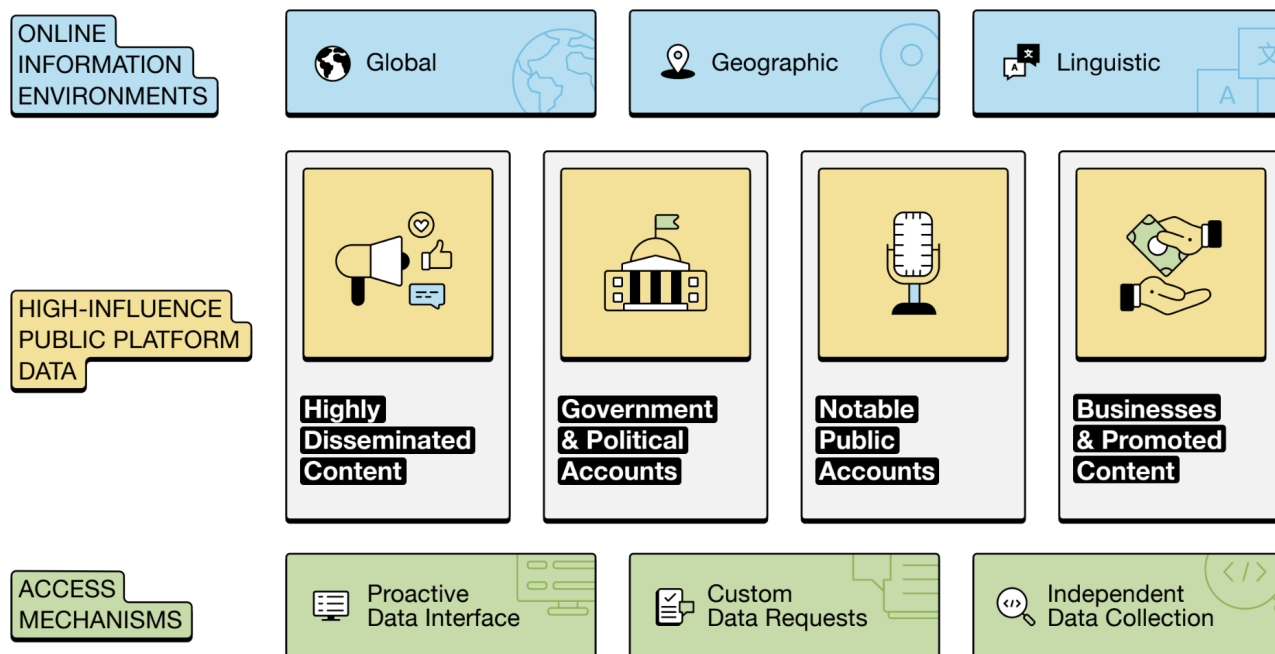
A Framework for High-Influence Public Platform Data

The framework offers a roadmap for expanding access to high-influence public platform data: the narrow slice of public platform data that has the greatest impact on civic life due to its reach, source, or role in shaping what people see online.

Public platform data does not all carry the same weight. Some accounts and content have far more influence than others. The *Better Access* framework focuses on the subset of public posts, accounts, and interactions that matter most for shaping public discourse.

Better Access defines the kinds of public platform data that researchers should be able to ethically use in their work. This definition sets a floor, not a ceiling – research with other public or non-public data is also important. The framework also sets out a model for data access mechanisms that platforms should implement in order to enable research, while acknowledging that permissionless, researcher-driven access to public platform data will remain necessary.

The Better Access Framework



Categories of High-Influence Public Platform Data

High-influence public platform data refers to the subset of public platform data that is most consequential, either because of its reach, its source, or its role in shaping what people see online. It is the data that tells us who holds influence, what content spreads, and how platforms themselves amplify certain voices.

Public posts vary in their significance. A family photo shared to a handful of followers is not the same, and should not be treated the same, as a head of state announcing a new policy or a viral video reaching millions. The *Better Access* framework focuses narrowly on the types of high-influence public platform data that shape civic discourse and access to information.

The framework identifies four categories of high-influence data:



Highly Disseminated Content: Posts or videos that achieve exceptional reach or engagement, shaping the public agenda.



Government and Political Accounts: Posts from accounts belonging to elected officials, candidates, political parties, and government institutions.



Notable Public Accounts: Content from accounts belonging to celebrities, journalists, civic leaders, or other public figures whose reach gives them outsized influence.



Business Accounts and Promoted Content: Advertising and commercial messaging, which can sway consumer behavior, public health, or public trust.

Defining Information Environments

What counts as influential varies dramatically across contexts. A viral post in the United States or India will tell us little about online discussion in Angola or Albania. For platform data access to be meaningful in different information environments, it must respond to local needs.

To account for these differences, the framework emphasizes three types of information environment:



Global: A platform's total user base across languages and geographies.



Geographic: Users of a platform in a particular geography.



Linguistic: Users of a platform engaging in a specific language or languages.

This approach ensures that vital research does not privilege only the largest countries or dominant languages.

How Access Should Work

To make public platform data meaningfully available, the framework defines three complementary data access mechanisms that platforms should support:



Proactive Data Interface: A structured, platform-supported access mechanism that provides ongoing, predictable access to high-influence public platform data. Examples include research Application Programming Interfaces (APIs), searchable archives, or downloadable datasets that are updated regularly without requiring researchers to make individual requests.



Custom Data Requests: Tailored datasets provided upon request to meet the specific local or thematic needs of researchers. These requests may be fulfilled through bespoke datasets, tailored APIs, archives, or other methods.



Independent Data Collection: Researcher-initiated access to high-influence public platform data, typically through automated collection such as scraping or crawling. This mode of access preserves researcher independence and enables validation of platform-provided data.

Together, these mechanisms can advance flexibility, relevance, and accountability across diverse research settings. Furthermore, this approach to high-influence public platform data can help to advance ethical research. The framework emphasizes access to public platform data categories where the public value of transparency is strongest. By focusing here, the framework sets clearer boundaries, lowers (but does not eliminate) privacy risks, and articulates a baseline expectation for consistent, responsible access.

The Way Forward

The stakes could not be clearer: when independent access to high-influence public platform data disappears, so does society's ability to understand and protect itself.

Regulators in Europe, the US, and beyond are already implementing or considering data access mandates. This framework offers a practical foundation: begin with the four categories of data that matter most for the public, ensure consistent and ethical access through multiple access mechanisms, and give researchers and the public the tools they need to understand the dynamics shaping public discourse in their own information environment.

High-influence public platform data is the minimum, most essential tier of data for accountability. This framework underscores that transparency is not a regulatory burden, but a democratic necessity. Adopting a uniform, cross-industry baseline would give the public the visibility it needs to see the digital public square clearly and to shape its future.

Table of Contents

I. Introduction.....	1
II. Terminology.....	2
III. The State of Platform Research.....	5
A. Overview of Platform Research.....	5
B. Barriers to Independent Research.....	11
C. An Incomplete Understanding of our Information Ecosystem.....	14
IV. Policy Landscape for Research.....	15
V. Core Concepts.....	16
A. Platforms.....	17
B. Information Environments.....	18
C. Access Mechanisms.....	19
D. Reasonable Understanding of Publicness.....	24
VI. Framework for High-Influence Public Platform Data.....	28
A. Highly Disseminated Content.....	29
B. Government and Political Accounts.....	34
C. Notable Public Accounts.....	42
D. Business Accounts and Promoted Content.....	45
VII. Baseline Data Access Fields.....	50
A. Overview of Data Types and Data Access Fields.....	50
B. Special Data Types.....	53
VIII. The Way Forward.....	56
Appendix A. The Policy Landscape for Public Platform Data.....	58
A. Voluntary Commitments.....	58
B. Regulatory Requirements.....	59
Appendix B. Research on Data Privacy.....	65
Appendix C. Proactive Data Interface Thresholds.....	67
Appendix D. Fields Included in Previous Platform Data Tools.....	70
Bibliography.....	72

I. Introduction

Online platforms and services shape what we know, how we connect, and who gets heard. From elections and public health to commerce and conflict, platforms are now indispensable infrastructure for civic life. Their influence is vast – and so is the need to understand them.

One critical way to understand platforms is through public platform data: the content, data, and information posted to platforms that anyone can access. From the statements a nation’s leaders share online to viral videos and images shared by influencers, data on public platforms shape our public discourse. These types of data offer a window into how information flows, how platforms amplify or suppress it, and what users see, engage with, and share.

Despite the importance of public platform data, there is no shared definition, framework, or approach to how research, journalism, and independent investigations (collectively referred to as “research” throughout the framework) should be conducted with public platform data, let alone consistent formats or predictable timelines for access. Meanwhile, platform companies increasingly leverage public platform data for a range of commercial purposes. They use public platform data to design services, target advertising, sell insights to third parties or data partners, and train large language models (LLMs).¹

While companies increasingly monetize users’ public platform data, researchers face barriers in accessing the same data. Platforms such as Meta, Reddit, and X (formerly Twitter) have eliminated data access tools that were once widely available for independent research. Researchers can face legal threats for accessing public platform data. Regulatory proposals in the United States (US) and frameworks like Article 40 of the European Union’s Digital Services Act (DSA) are beginning to require access to public platform data, but regulations remain limited and in the early stages of implementation.

Definitions of publicly available information differ across national legal frameworks and platform access mechanisms, making it difficult to find congruence across different contexts and frameworks. Indeed, researchers report that inconsistent and restricted access to platform data is one of the greatest obstacles to meaningful platform research.²

That is the reason for developing **Better Access**, a baseline framework for **independent access to public platform data**. Created by the Knight-Georgetown Institute (KGI) Expert Working Group on Public Platform Data, the framework specifically focuses on high-influence public platform data, a subset of public platform data that is most consequential, either because of its reach, its source, or its role in shaping what people see online. It is the data that tells us who holds influence, what content spreads, and how platforms themselves amplify certain voices.

¹ Paul and Tong, “Inside Big Tech’s Underground Race to Buy AI Training Data.”

² Coalition for Interdependent Technology Research, “Power in Numbers.”

The framework offers practical, cross-industry guidance for the minimum set of public platform data needed to begin to understand our digital environments. The framework is grounded in the principle that transparency is essential to accountability, and that enabling meaningful access to public platform data serves the common good.

At the same time, there are limits and risks of openness. Public platform data can raise privacy risks and ethical concerns, which is why this framework focuses on high-influence public platform data. Defining this narrow slice of public platform data as a baseline can help to advance more consistent and ethical research.

The framework is structured as follows:

- Section II defines key terms used in the framework.
- Section III provides an overview of platform research, including research collaborations, independent research, and ethical research frameworks.
- Section IV summarizes the policy landscape for platform research.
- Section V provides the core concepts of the framework.
- Section VI articulates the *Better Access* framework for high-influence public platform data.
- Section VII describes baseline fields for data access.
- Section VIII discusses the way forward.

II. Terminology

This section defines the key terms used within the *Better Access* framework.

Absolute threshold	An absolute value of total reach or engagement triggering inclusion in a data access mechanism.
Account	A distinct user identity, whether individual, organizational, or automated, through which content is created, shared, or engaged with on a platform. Platforms may require users to create an account. The framework distinguishes between an account and content shared or disseminated by the account.
Access mechanism	A mechanism by which researchers access high-influence public platform data. The access mechanisms discussed in the framework are: Proactive Data Interfaces, Custom Data Requests, and Independent Data Collection.
Automated collection	Techniques to extract data from websites or other web applications, including web scraping or web crawling.

Business Accounts	Accounts that are monetized and/or established primarily for a commercial purpose.
Custom Data Requests	Researcher-initiated requests for high-influence public platform data necessary for specific local or thematic needs of researchers. These requests may be fulfilled through bespoke datasets, tailored APIs, archives, or other methods.
Data	Platform content, data, and information.
Deleted content	Content removed from a platform by either the platform or the user who posted it.
Digital Services Act (DSA)	A European Union regulation adopted in 2022 that seeks to create a safer digital space where the fundamental rights of users are protected in the EU.
Engagement	User interaction with items on an online platform, including clicks, taps, comments, reshares, indications of approval or disapproval (such as likes, dislikes, upvotes, or downvotes), or other forms of interaction.
Ephemeral content	Content that is designed to be erased after being displayed for a limited period of time.
Framework	This document and the definitions and criteria it lays out.
General Data Protection Regulation (GDPR)	A European Union regulation adopted in 2016 that regulates how personal data of individuals within the EU is processed.
Government and Political Accounts	Accounts associated with major political parties, candidates, government officials, or government institutions acting in an official capacity.
Highly Disseminated Content (HDC)	Content that meets absolute or relative thresholds for reach or engagement as defined in this framework.
High-influence public platform data	The accounts, interactions, and content that are most likely to shape public life at scale, whether because of who is speaking, how widely content spreads, or how it is amplified.
Independent Data Collection	Researcher-initiated access to high-influence public platform data, typically through automated collection such as scraping or crawling.

	This mode of access preserves researcher independence and enables validation of platform-provided data.
Information environments	The global, geographic, and linguistic environments where information is produced, disseminated, and consumed.
Notable Public Account	Accounts belonging to notable public figures or entities, defined by reach or engagement, as well as public interest actors.
Online platforms and services (referred to as “platforms”)	Public-facing websites, online services, online applications, or mobile applications that predominantly store, aggregate, and disseminate user-generated information to the public, including text, videos, images, games, audio files, or other content.
Political party	A formal organization that organizes public opinion, communicates demands and priorities, and engages in political recruitment.
Proactive Data Interface	A structured, platform-maintained access mechanism that provides ongoing, predictable access to high-influence public platform data. Examples include research APIs, searchable archives, or downloadable datasets that are updated regularly without requiring researchers to make individual requests.
Promoted content	Content an account pays to make visible to a broad or specific audience.
Publicly available information	Legally defined in many jurisdictions around the world, it often encompasses a combination of government records, widely distributed media, and/or other legally mandated disclosures.
Reach	The number of unique platform users who have viewed, listened to, or downloaded a particular piece of content. Reach may also consist of other platform-specific metrics.
Reasonable understanding of publicness	The idea that a user interacting with a platform would reasonably understand that their data is or may be publicly accessible, taking into account platform features and policies.
Relative threshold	A relative value of total reach or engagement triggering inclusion in a data access mechanism.
Researcher	Individuals or organizations who facilitate the development of society’s collective knowledge of platforms. This includes a range of actors, including but not limited to platform users, journalists, civil society organizations, academics, and other types of researchers. Law

	enforcement, intelligence, and defense entities are not considered researchers. ³
Spotlighted content	Content that is posted by a user and then reposted by another user, changing its audience.

III. The State of Platform Research

Independent platform research is vital for understanding the digital systems that influence everything from public opinion to government policy. Independent research drives smarter platform design, better regulation, and more-informed societies.⁴ This research is undertaken by a wide range of actors, including groups of platform users, journalists, civil society organizations, academics, and more.⁵

This section maps the landscape of platform research: the types of research being conducted, the kinds of data it depends on, and the importance of public platform data within this research ecosystem. It outlines growing obstacles that researchers and the public face when studying online information systems, and examines the consequences of these constraints, including the rise of data vendors. The section then explores the ethical and privacy challenges of platform research, along with emerging frameworks aimed at addressing such risks.⁶

A. Overview of Platform Research

Platform research ranges from basic content analysis to sophisticated network experiments. While platforms historically have offered a variety of access mechanisms and tools to support research, these have become more restricted over time. This section summarizes common research tools, methods, and ethical frameworks, and describes both research collaborations and independent study.⁷

Research Collaborations

Large platforms including Facebook, Instagram, Twitter, and YouTube have developed various programs to support research, but access is increasingly limited, inconsistent, and controlled.

Programs include:

- **Ad Libraries:** Major platforms have offered ad libraries to researchers for years,⁸ including Meta,⁹ TikTok,¹⁰ and LinkedIn.¹¹ Under Article 39 of the DSA, very large platforms are required

³ Levinson-Waldman and Balkam, “The Government’s Growing Trove of Social Media Data”; Vogus, “Defending Data.”

⁴ Shiffman and Silverman, “The Case for Transparency.”

⁵ Coalition for Independent Technology Research, “Power in Numbers.”

⁶ Lenhart, “Ethical Use of Pervasive Data for Research.”

⁷ Bundtzen, “Data Access.” See also Initiative for Digital Public Infrastructure, “New Approaches to Platform Data Research.”

⁸ See European Commission, “Status Report”; Institute for Data, Democracy, & Politics, “Tracking Transparency to Enable Research.”

⁹ Meta, “Ad Library.”

¹⁰ TikTok, “Find Ads on TikTok.”

¹¹ LinkedIn, “Ad Library.”

to maintain ad libraries.¹² Some very large platforms provide libraries only in the EU to comply with the DSA (Pinterest,¹³ Snap¹⁴).

- **Application Programming Interfaces (APIs):** Most platforms have developed APIs that allow external access to platform data. In some cases, platforms direct researchers to APIs primarily intended for use by commercial developers, but at other times, platforms have built APIs specifically for research purposes.¹⁵ API access varies substantially from platform to platform, and within platforms over time. X, for example, eliminated free access to its API in 2023.¹⁶ Most platform APIs require registration and approval before use and researchers report barriers to access.¹⁷ Indeed, a 2025 survey found that platform API access faced “multiple barriers including complex API application processes, difficulties obtaining credentials, and limited API usability.”¹⁸
- **Data Archives:** Some platforms have offered archives of historical data. Twitter, starting in 2018, released a public archive of state-backed information operations.¹⁹ It also hosts a database of Community Notes from the platform.²⁰ The Social Media Archive (SOMAR) at the University of Michigan hosts data from X, Facebook, Instagram, and Reddit, among other platforms.²¹
- **Research Partnerships:** Platforms have occasionally launched partnerships with researchers – such as Twitter’s “Healthy Conversations”²² project or Meta’s 2020 election research partnership.²³ These collaborations are rare, often short-term, and have been criticized by participants.²⁴ Ad hoc researcher collaborations also occur, but formal collaboration typically requires agreement from platform leadership.²⁵

Partnerships come with strings attached. Companies are said to exert significant control over “research questions, methodology, workflow, and design choices” of research partnerships.²⁶ Communication of research may be controlled by the company, and there have been documented disagreements between researchers and companies over how to interpret research findings.²⁷ Furthermore, some have criticized partnerships for externalizing

¹² European Union, “Digital Services Act,” Article 39.

¹³ Pinterest, “Ads Repository.”

¹⁴ Snap, “Ads Gallery.”

¹⁵ Freelon, “Computational Research in the Post-API Age.”

¹⁶ Center for an Informed Public, “Twitter’s API Access Changes.”

¹⁷ Ayrshare, “Top 10 Social Media APIs for Developers.”

¹⁸ Mimizuka, “Post-Post-API Age.”

¹⁹ Gadde and Roth, “Enabling further research of information operations on Twitter”; Roth, “The Twitter Moderation Research Consortium is now open to researchers.”

²⁰ X, “Downloading data.”

²¹ SOMAR, “About.”

²² Gadde and Gasca, “Measuring healthy conversation.”

²³ Meta, “Research partnership to understand Facebook and Instagram’s role in the U.S. 2020 election.”

²⁴ Tromble, “Where Have All the Data Gone?”; Wagner, “Independence by Permission.”

²⁵ Scott, “Survey.”

²⁶ Iyer, “A Primer on the Meta 2020 U.S. Election Research Studies.”

²⁷ Hendrix and Barrett, “The Meta Studies.”

reputational and operational risks onto researchers. These dynamics have led some researchers to argue that collaborative research with platforms is fundamentally flawed, calling such research “independence by permission.”²⁸

Independent Study

Beyond formal collaboration with platforms, researchers also independently investigate platform dynamics using methods such as independent collection, user data donation, and whistleblower disclosures.

- **Independent collection:** Researchers have used automated collection – often web scraping or crawling – to independently gather public platform data. This work has been crucial to public understanding, shedding light on child safety issues, civic discourse, functioning of recommendation systems, and more.²⁹

Yet platforms try to restrict these efforts. Companies have suspended accounts, issued legal threats, and sued nonprofits for breaching terms of service.³⁰ While lawsuits targeting nonprofit researchers have failed on the merits as of 2025,³¹ the legal risks alone can chill research³² and strain institutional and financial support.³³ For many researchers, the threat of litigation raises the difficult question of whether research through automated collection is worth the risk.

- **User data donation:** Data donation involves users sharing their own data with third-party researchers for analysis.³⁴ Legal rights to data portability, including through Europe’s General Data Protection Regulation (GDPR) Article 20,³⁵ Digital Markets Act (DMA) Article 6(9),³⁶ and the California Consumer Privacy Act (CCPA),³⁷ have helped enable this research.

While data donation has strengths, the method also has limitations, including incomplete datasets and logistical and analytical challenges.³⁸ Data donation studies rely on voluntary participation and are sensitive to platform-side changes in data structure or format. Despite their utility, data donations offer only a partial view of the broader information ecosystem.

²⁸ Wagner, “Independence by Permission.”

²⁹ For many examples of social media research based on automated collection, see Brown et al., “Web Scraping for Research.”

³⁰ HateAid, “For Independent Research”; Klar, “Facebook Suspends Accounts of NYU Researchers Who’ve Criticized Platform”; United States District Court for the Northern District of California, “Demand for Jury Trial.”

³¹ Ingram, “Judge throws out Elon Musk’s X lawsuit against nonprofit.”

³² Abdo et al., “A Safe Harbor for Platform Research.”

³³ DiResta, *Invisible Rulers*.

³⁴ Lukito et al., “Enabling Independent Research without Unleashing Ethics Disasters”; Mozilla Foundation, “About YouTube Regrets”; Sanderson and Mohammed, “A Multi-Stakeholder Approach for Leveraging Data Portability to Support Research on the Digital Information Environment.”

³⁵ European Union, “General Data Protection Regulation,” Article 20.

³⁶ European Union, “Digital Markets Act,” Article 6(9).

³⁷ California, “California Consumer Privacy Act.”

³⁸ Tucker et al., “The Case for Open Data Access to Aid Tech Regulation.”

- **Whistleblower disclosures and leaked documents:** Whistleblowers and leaked documents have become a critical, if unpredictable, source of information. The FBarchive hosted by the Public Interest Tech Lab at Harvard,³⁹ for instance, hosts the trove of documents released by former Meta employee Frances Haugen.⁴⁰ These files, often redacted for privacy,⁴¹ have underpinned impactful research. Leaked documents have also revealed Meta’s internal rules for AI chatbots,⁴² as well as TikTok’s knowledge of how its platform could harm young users.⁴³

In response to staff leaks, companies have moved to further restrict data access. While access restrictions impact external researchers, they also reportedly make it harder for internal employees to assess platform decisions from inside the company.⁴⁴

Methodologies

Research collaborations and independent study utilize a range of methodologies. Most independent research of platforms is observational rather than experimental. It describes important patterns, but rarely establishes causality.⁴⁵ Experimental methods are important for understanding how platforms shape user behavior and societal outcomes.⁴⁶

While platforms routinely conduct their own experiments, they rarely share results externally.⁴⁷ Independent researchers face particular barriers in investigating causal relationships on platforms, relying primarily on leaked results from internal experiments. Researchers and legislators are increasingly requesting experimental data from platforms as a source of transparency.⁴⁸ The DSA requires qualifying researchers to be granted access to internal platform data, including, potentially, experiments.⁴⁹ Similar regulatory mechanisms have been proposed in the US.⁵⁰

Across methodologies, platform research is grappling with the rise of decentralized social media platforms and generative AI.⁵¹ However, at this stage, research access to decentralized platforms⁵² and LLMs⁵³ can best be described as mixed. Bluesky, for example, has enabled broad API access, including for researchers, journalists, and civil society organizations.⁵⁴ Independent access to data from new platforms and AI companies is an evolving challenge facing researchers.

³⁹ Public Interest Tech Lab, “About.”

⁴⁰ Public Interest Tech Lab, “Fbarchive.”

⁴¹ Public Interest Tech Lab, “Redaction Guidelines.”

⁴² Horwitz, “Meta’s AI rules have let bots hold ‘sensual’ chats with kids, offer false medical info.”

⁴³ Allyn et al., “States probed TikTok for years.”

⁴⁴ Horwitz, “The Facebook Files.”

⁴⁵ Mosleh et al., “Field Experiments on Social Media.”

⁴⁶ Thorburn et al., “Experiments Are the Best Kind of Transparency.”

⁴⁷ Quin et al., “A/B testing.”

⁴⁸ Grüning et al., “Independently testing prosocial interventions.”

⁴⁹ European Commission, “Delegated Regulation on data access provided for in the Digital Services Act”; European Union, “Digital Services Act,” Article 40.

⁵⁰ Thorburn et al., “Experiments Are the Best Kind of Transparency.”

⁵¹ Harrington and Vermeulen, “External Researcher Access to Closed Foundation Models.”

⁵² Guhl et al., “Researching the Evolving Online Ecosystem.”

⁵³ Anthropic, “Anthropic Economic Index.”

⁵⁴ Bluesky, “Bluesky Developer APIs.”

Ethical and Rights-Advancing Research Frameworks

Research using data from platforms can raise important ethical and personal privacy risks.⁵⁵ The scope and complexity of platform data has rapidly expanded. From social networking to livestreams, and from text to images and audio to video, researchers are adapting to new data formats and the related evolution of user behaviors and expectations.⁵⁶

The collection and curation of public and private data is foundational for the commercial model of many platforms,⁵⁷ and platforms often configure defaults to ensure that users engage in public ways.⁵⁸ With platforms collecting previously unthinkable quantities of user data, the research community has assessed and evolved ethical frameworks to keep pace.⁵⁹ Given that the dominant commercial models of platforms are rooted in data tracking and ad-driven personalization, research methods and ethical frameworks must inevitably address questions of consent, accountability, and risk.⁶⁰

There have been challenges in ensuring ethical research with public platform data.⁶¹ A widely cited 2010 study revealed how “de-identified” Facebook profile data could be re-identified, demonstrating the inadequacy of the safeguards used by researchers.⁶² The study signaled how existing ethical frameworks, including reliance on Institutional Review Boards (IRBs) for human subject research and user consent (already difficult to access in much of the world), would need to evolve alongside digital platforms.⁶³ In 2018, the company Cambridge Analytica drew global attention, revealing the scale of personal data collection by Facebook and how it can be misused by bad actors.⁶⁴ The resulting scandal corresponded with shifts in how platforms approached third-party access, including in relation to researchers.⁶⁵ Independent research that relies on the data generated through platforms must inevitably navigate a range of ethical and privacy risks.⁶⁶

In response, companies, regulators, and independent researchers have proposed new approaches. Platforms often cite privacy and ethical risks as reasons to restrict research access,⁶⁷ though these concerns can appear disingenuous given that the same companies monetize the same data with third-party data contracts.⁶⁸ Regulators are attempting to strike a new balance through frameworks like the DSA and proposals like the Platform Accountability and Transparency Act (PATA)⁶⁹ or the Digital

⁵⁵ Federal Trade Commission, “Complaint for Civil Penalties, Injunction, and Other Relief.”

⁵⁶ Brown et al., “Web Scraping for Research.”

⁵⁷ Zuboff, *The Age of Surveillance Capitalism*.

⁵⁸ Malik, “Are Social Media Settings Intentionally Confusing?”

⁵⁹ Hudson and Bruckman, “Go Away.”

⁶⁰ Shilton et al., “Excavating awareness and power in data science.”

⁶¹ Zimmer, ““But the data is already public””; Gilbert et al., “When research is the context.”

⁶² Zimmer, ““But the data is already public.””

⁶³ Metcalf and Crawford, “Where are the human subjects in Big Data research?”; Pater et al., “No Humans Here.”

⁶⁴ Confessore, “Cambridge Analytica and Facebook.”

⁶⁵ Freelon, “Computational Research in the Post-API Age.”

⁶⁶ National Telecommunications and Information Administration, “Ethical Research with Online/Pervasive Data”; PERVADE, “The Project.”

⁶⁷ European Commission, “Delegated Regulation on data access provided for in the Digital Services Act.”

⁶⁸ Hals, “Meta investors, Zuckerberg reach settlement to end \$8 billion trial over Facebook privacy violations.”

⁶⁹ United States, “Platform Accountability and Transparency Act.”

Services Oversight and Safety Act (DSOSA)⁷⁰ in the US, which are designed to enable researcher access while safeguarding privacy rights.

Another strategy is the use of Privacy-Enhancing Technologies (PETs), such as anonymization, differential privacy, and synthetic data generation. PETs seek to allow researchers to study platform data without exposing personal information. The promise of PETs, however, has not always matched their performance in practice.⁷¹ One notable case involved Facebook's release of a dataset to researchers that used differential privacy to manage personal data risks.⁷² Despite the project's ambitions and strong backing across industry and academia, it failed to deliver when, years into the project, an independent researcher found fundamental flaws in the underlying data provided by Facebook. This discovery invalidated years of research and undermined public trust in both Facebook and the research.⁷³ Without transparency or the ability for researchers to validate data, PETs alone may be insufficient to ensure rigorous and ethical research.

Meanwhile, civil society, academia, and journalism have developed frameworks to respond to new platforms and evolving ethical risks. The Coalition for Independent Technology Research (CITR), founded in 2022, brings together academics, journalists, and civil society actors to promote ethical standards and institutional support for independent research.⁷⁴ CITR emphasizes that maintaining public trust in research is central.⁷⁵ To maintain and deepen trust, CITR members promote ethical frameworks and standards to enable diverse forms of research while simultaneously ensuring users' rights are protected.

The Association of Internet Researchers' Ethical Guidelines, now in their third edition, articulate standards for ethical internet research.⁷⁶ The Pervasive Data Ethics for Computational Research (PERVADE) project at the University of Maryland has developed tools to advance ethical computational research.⁷⁷ A 2025 report from the Institute for Data Democracy and Politics (IDDP) at The George Washington University offers a comprehensive examination of the privacy and ethical challenges associated with research that relies on platform data and offers important recommendations to advance ethical research practices.⁷⁸ The report identifies how research and ethical review processes should evolve to keep pace with and enable ethical research with expanding platform data.

Journalists, too, have worked to develop responsible frameworks for the use and analysis of publicly available information. The International Fact-Checking Network (IFCN) at Poynter maintains a Code of

⁷⁰ United States, "Digital Services Oversight and Safety Act."

⁷¹ Shiffman, "Tools for Platform Research."

⁷² Alba, "Facebook sent flawed data to misinformation researchers"; Nayak, "New privacy-protected Facebook data for independent research on social media's impact on democracy."

⁷³ Timberg, "Facebook made big mistakes in data it provided to researchers, undermining academic work."

⁷⁴ Matias et al., "Manifesto."

⁷⁵ Shilton et al., "Excavating awareness and power in data science."

⁷⁶ franzke et al., "Internet Research."

⁷⁷ PERVADE, "The Project."

⁷⁸ Lenhart, "Ethical Use of Pervasive Data for Research."

Principles to promote accountable journalism.⁷⁹ News organizations also maintain their own ethical standards to guide analysis and use of digital platform data.⁸⁰

Ultimately, it is essential that research focused on public platform data is conducted in ethical and rights-advancing ways. Researchers, regulators, and platforms all should advance transparent, enforceable standards that evolve alongside digital systems.

B. Barriers to Independent Research

Independent research with public platform data faces significant barriers.⁸¹ Platforms have restricted research and many researchers are in turn relying on commercial data vendors for access to data. Research barriers are not new, but they are growing.⁸²

Restrictions on Data Access

Barriers to data access are substantial. A 2024 survey of over 100 researchers found that more than half had abandoned projects due to lack of data access.⁸³ CITR's 2025 State of Independent Tech Research report found that data access is a primary obstacle to independent platform research.⁸⁴

It might seem contradictory to suggest that public platform data is difficult to access. While data may be visible to thousands, millions, or even hundreds of millions of platform users, platforms do not always facilitate or enable collection of data at scale, which is required for real accessibility. Sophisticated technical skills should not be required for journalists, civil society groups, and academics to be able to analyze public platform data. Furthermore, legal risks, research restrictions, and the oversight of research collaborations by platform companies create access difficulties.⁸⁵

Many platforms initially tolerated independent platform research. In 2019, Facebook expanded access to CrowdTangle, a tool it had acquired in 2016, promising to help researchers bring “more transparency into the spread of public content across Facebook’s platforms.”⁸⁶ Twitter, too, actively promoted academic research, offering developer APIs⁸⁷ and dedicated tools for academics,⁸⁸ citing the platform’s belief in “the value of an open exchange of information.”⁸⁹

However, in 2023 and 2024, some of the world’s largest platforms significantly reduced researcher access. Meta shuttered CrowdTangle, the transparency tool frequently used by journalists and

⁷⁹ International Fact-Checking Network, “Commit to Transparency.”

⁸⁰ AFP, “AFP editorial standards and best practices”; New York Times, “Ethical Journalism.”

⁸¹ Freelon, “Computational Research in the Post-API Age.”

⁸² Bundtzen and Schwieter, “Access to Social Media Data for Public Interest Research.”

⁸³ Reynolds, “Findings from a Survey of Researcher Pain Points.”

⁸⁴ Coalition for Independent Technology Research, “Power in Numbers.”

⁸⁵ Agre, “Toward a Critical Technical Practice”; DeTar, “On Selling Out”; Matias, “Why We Need Industry-Independent Research on Tech & Society”; Vertesi and Matias, “Divesting from Big Tech”; Whittaker, “The Steep Cost of Capture.”

⁸⁶ Shiffman and Silverman, “CrowdTangle Opens Public Application for Academics.”

⁸⁷ Kim et al., “The Story of Goldilocks and Three Twitter’s APIs.”

⁸⁸ Roth and Gadde, “Expanding access beyond information operations”; Twitter, “Twitter data for academic research.”

⁸⁹ Twitter, “Twitter data for academic research.”

researchers all around the world.⁹⁰ X imposed steep fees on its API, effectively locking out most researchers.⁹¹ Reddit revised its Data API terms, introducing new rate limits that researchers claim undermine research.⁹² Many of these platforms have changed access rules repeatedly since 2006, disrupting research projects midstream.⁹³ Each of these moves inhibit research and require researchers to retool methodologies and reorient studies.⁹⁴

CrowdTangle was widely available for researchers that included journalists and civil society organizations. Its replacement, the Meta Content Library (MCL), is less accessible for some journalists and civil society organizations.⁹⁵ Platforms have also added new restrictions on how data may be used, refreshed, or shared.⁹⁶ In aggregate, these moves have placed some research access behind paywalls or within clean rooms controlled by platforms. Platforms have also introduced new requirements for approval of publications that rely on platform data, threatening the independence of research.

As independent researchers face growing barriers, many platform business models are built on the collection and monetization of extensive user data, including without informed consent of users.⁹⁷ This includes the provision of data to advertisers as well as for the training of LLMs.⁹⁸ Meta, for example, trains its AI systems with publicly available information, including public Facebook posts and comments.⁹⁹ Reporting suggests Meta is also seeking to access photos in users' camera libraries.¹⁰⁰

Even as regulation demands new forms of data access for researchers, platforms are complying in ways that seek to limit independent scrutiny of the platform.¹⁰¹ The European Commission, responsible for enforcement of the DSA data access requirements, opened proceedings against multiple platforms for alleged shortcomings in enabling researcher access to publicly available information. This includes investigations focused on X,¹⁰² TikTok,¹⁰³ AliExpress,¹⁰⁴ and Facebook/Instagram.¹⁰⁵ As of November 2025, the Commission has found both TikTok and Meta in preliminary breach of the DSA's obligations to provide researcher access to public data.¹⁰⁶

⁹⁰ Center for American Progress, "CrowdTangle Letter."

⁹¹ Coalition for Independent Technology Research, "Letter."

⁹² Gilbert et al., "Survey Report"; Reddit, "Creating a Healthy Ecosystem for Reddit Data and Reddit Data API Access."

⁹³ Observatory on Social Media, "OSoMe Awesome Speakers."

⁹⁴ Borchers et al., "Facebook Post Data."

⁹⁵ Gotfredsen and Dowling, "Meta Is Getting Rid of CrowdTangle—and Its Replacement Isn't as Transparent or Accessible."

⁹⁶ Ibid.; Counts, "TikTok's Rules Deter Researchers from Crunching Data on Users, Misinformation."

⁹⁷ Electronic Privacy Information Center, "Social Media Privacy."

⁹⁸ Chee, "X hit by complaints to EU over user data and targeted advertising."

⁹⁹ Associated Press, "Meta says it will resume AI training with public content from European users."

¹⁰⁰ Perez, "Facebook is asking to use Meta AI on photos in your camera roll you haven't yet shared."

¹⁰¹ Counts, "TikTok's Rules Deter Researchers from Crunching Data on Users, Misinformation."

¹⁰² European Commission, "Commission Sends Preliminary Findings to X for Breach of the Digital Services Act."

¹⁰³ European Commission, "Commission Opens Formal Proceedings against TikTok under the Digital Services Act."

¹⁰⁴ European Commission, "Commission Opens Formal Proceedings against AliExpress under the Digital Services Act."

¹⁰⁵ European Commission, "Commission Opens Formal Proceedings against Facebook and Instagram under the Digital Services Act."

¹⁰⁶ European Commission, "Commission preliminarily finds TikTok and Meta in breach of their transparency obligations under the Digital Services Act."

The Rise of Data Vendors

Given the lack of consistent, platform-provided data access, data vendors have emerged as key data intermediaries. Researchers increasingly turn to third-party commercial tools such as Meltwater, NewsWhip, Tubular Labs, and Bright Data,¹⁰⁷ or they build their own data pipelines.¹⁰⁸ This adds expense and complexity to research projects and results in a fragmented research ecosystem.

Platforms are racing to monetize their own private and public platform data¹⁰⁹ as well as additional data across the internet.¹¹⁰ Third-party data companies focused on automated collection and the sale of LLM training datasets are booming.¹¹¹ Actions taken by platforms to limit public data harvesting and funnel data collectors towards licensing agreements are creating barriers for independent researcher access as well.¹¹²

Data vendors have at times specifically sought to sell themselves as tools for researchers. For instance, NewsWhip, a commercial tool that was acquired by Sprout Social, actively marketed itself as a CrowdTangle replacement (“your only CrowdTangle alternative”)¹¹³ after Meta shut down the tool.¹¹⁴ CrowdTangle offered free access whereas NewsWhip charges for access.

Commercial vendors and civil society organizations provide access to a range of platform datasets, including public platform data and a broader array of publicly available information. Indeed, a 2025 report examining third-party social media research and monitoring tools found more than 250 tools that offer scraping, network analysis, and social media monitoring, many of which facilitate paid access to public platform data in one way or another.¹¹⁵

Commercial vendors may rely on commercial API access or independent data collection via scraping. Meltwater is a commercial tool that tracks information in near-real-time across platforms including X, Facebook, Instagram, YouTube, Reddit, Snap, Discord, and many other platforms.¹¹⁶ NewsWhip tracks information across platforms including Facebook, X, Instagram, YouTube, TikTok, web articles, and Reddit.¹¹⁷ Bright Data offers automated collection tools as well as precollected datasets.¹¹⁸

Civil society groups also offer data access tools. Junkipedia, built by the Algorithmic Transparency Institute, a project of the National Conference on Citizenship, monitors information across 15 different

¹⁰⁷ Meltwater, “The Complete Boolean Library”; Meltwater, “Social Listening & Analytics”; Miles, “From Dashboards to Data Acquisitions”; NewsWhip, “Stay ahead of emerging issues, identify critical moments, and track the spread of misinformation with NewsWhip.”

¹⁰⁸ Meyer et al., “Enhancing the ethics of user-sourced online data collection and sharing.”

¹⁰⁹ Reddit, “Addressing the community about changes to our API.”

¹¹⁰ Li et al., “LLM-PBE.”

¹¹¹ Nulty, “Top 5 AI Training Data Providers of 2025”; Perez, “Meta drops lawsuit against web-scraping firm Bright Data that sold millions of Instagram records”; Privacy International, “Large language models and data protection.”

¹¹² Reuters, “Reddit to update web standard to block automated website scraping.”

¹¹³ Quigley, “Six key features for newsrooms seeking a CrowdTangle replacement.”

¹¹⁴ NewsWhip, “NewsWhip, your only CrowdTangle alternative.”

¹¹⁵ See Miles, “From Dashboards to Data Acquisition.”

¹¹⁶ Meltwater, “The Complete Boolean Library”; Meltwater, “Social Listening & Analytics.”

¹¹⁷ Newswhip, “Stay ahead of emerging issues.”

¹¹⁸ Bright Data, “Limitless web data infrastructure for AI & BI.”

social media platforms, including Facebook, Instagram, YouTube, X, TikTok, and Telegram.¹¹⁹ Junkipedia explicitly focuses on supporting researchers, journalists, grassroots activists, and other institutions.¹²⁰ Other commercial platforms also provide researchers with free or discounted access. The Bright Initiative, from Bright Data, provides selected non-profit organizations, academic institutions, and public bodies with pro-bono data access.¹²¹

However, the legitimacy of data access by some of these third-party data providers continues to be contested by platforms and users, whose data has often been used without their knowledge or consent.¹²² Platforms have, at times, attempted to restrict data access for commercial vendors and introduce barriers to inhibit automated collection. In *LinkedIn v. HiQ*¹²³ and *Meta Platforms vs. Bright Data*, for example, platforms challenged commercial vendors for unauthorized collection of public platform data.¹²⁴ These tensions further underscore the need for standardized data categories and access mechanisms for independent researchers to access public platform data outside of commercial tools.

C. An Incomplete Understanding of our Information Ecosystem

Barriers to data access have created deep blind spots in understanding the digital information ecosystem. Systematic reviews of peer-reviewed publications further demonstrate that existing research on platforms is highly uneven, concentrated on a few platforms, in a few languages, and in a few countries.

A review of papers published between 2017 and 2021 in ten top academic journals found that nearly half of all studies relied exclusively on Twitter data, despite Twitter's smaller total user base than other major platforms.¹²⁵ This focus does not necessarily reflect Twitter's societal importance in particular information environments, but its historically generous data access policies.¹²⁶ Facebook has also received more attention in research largely due to CrowdTangle, before Meta shuttered it in 2024.¹²⁷ In contrast, commonly used platforms like YouTube, TikTok, Instagram, Pinterest (four of the five most used platforms in the US¹²⁸), and newer platforms have been far less studied.

Moves by X, Meta, and Reddit to restrict API access threaten the public's already-limited understanding. Without access, researchers cannot evaluate platform design, amplification, or risks. The voluntary nature of data access gives platforms a perverse incentive. If they restrict research, they

¹¹⁹ Junkipedia, "Untangle the online conversation across social media to decode what really matters."

¹²⁰ Ibid.

¹²¹ Bright Data, "The Bright Initiative."

¹²² Perez, "Meta drops lawsuit against web-scraping firm Bright Data that sold millions of Instagram records"; Privacy International, "Large language models and data protection."

¹²³ United States Court of Appeals for the Ninth Circuit, "Opinion."

¹²⁴ United States District Court for the Northern District of California, "Order Denying Meta's Motion for Partial Summary Judgment; and Granting Bright Data's Motion for Summary Judgment."

¹²⁵ Norton and Shapiro, "How to Better Study—and Then Improve—Today's Corrupted Information Environment."

¹²⁶ Tromble, "Where Have All the Data Gone?"

¹²⁷ Zhang et al., "Center-Left and Right-Wing News, YouTube, and Twitter as Key Connectors in the Social Media System."

¹²⁸ Gottfried, "Americans' Social Media Use."

may be able to reduce scrutiny.¹²⁹ For example, investigations into platform design decisions at Apple and Google suggest that platform design choices frustrate transparency and oversight.¹³⁰

Analysis of platforms also focuses primarily on Western countries and the English language.¹³¹ Two-thirds of papers from the 2017-2021 study focused only on Western countries and 60% of the studies were focused on English-language data.¹³² A study conducted by the International Panel on the Information Environment found significant gaps in the study of non-English information.¹³³ A further systematic review of worldwide causal and correlational evidence on digital media and democracy similarly found dozens of studies focused on the US, South Korea, and Europe, but just a handful exploring dynamics in Africa and Latin America.¹³⁴ Other commenters have argued that online information ecosystem research does not sufficiently capture Global Majority trends.¹³⁵

These imbalances leave many parts of the world and many digital platforms understudied. Knowledge gaps also weaken the evidence base for effective journalism and oversight.

IV. Policy Landscape for Research

Advocates have pushed for stronger transparency from digital platforms, leading to the development of both voluntary and regulatory transparency regimes. These efforts broadly aim to ensure platforms disclose information about their operations, policies, recommender systems, and moderation practices to users, researchers, and regulators. While some transparency practices are shaped by platforms' own policies, both voluntary commitments and regulation play a central role in defining how platform information is shared and accessed. Privacy laws and ethical research frameworks also influence both the scope and manner in which data can, and should, be made available to independent researchers. Appendix A describes these frameworks in more detail.

Digital platforms have experimented with voluntary transparency measures, though their consistency and depth vary significantly. Some companies publish transparency reports on content moderation or government requests, which are tracked by Access Now's Transparency Reporting Index.¹³⁶ Collaborative frameworks such as the Santa Clara Principles¹³⁷ the EU's Code of Conduct on

¹²⁹ Casas et al., "Commentary: How do we get platforms to share data with independent researchers?"

¹³⁰ Kollnig and Shadbolt, "How Decisions by Apple and Google obstruct App Privacy."

¹³¹ Tavishi, "Platform Transparency under the EU's Digital Services Act."

¹³² Norton and Shapiro, "How to Better Study—and Then Improve—Today's Corrupted Information Environment."

¹³³ A systematic review and meta-analysis of 4,798 peer-reviewed publications finds that "[o]ne of the critical next steps for both systematic reviews and meta-analysis is to incorporate research findings published in languages other than English, by researchers around the world, with data about users who consume content in other languages." See International Panel on the Information Environment, "Strategies for Improving the Global Information Environment."

¹³⁴ Budak et al., "Misunderstanding the Harms of Online Misinformation"; Lorenz-Spreen et al., "A Systematic Review of Worldwide Causal and Correlational Evidence on Digital Media and Democracy."

¹³⁵ Carnegie interviewed fifty-four experts on the information environment who have long-term, on-the-ground experience in twenty-nine countries across Asia, Latin America, Africa, and the Middle East. See Lai, "Understanding the Information Environment."

¹³⁶ Access Now, "Transparency Reporting Index."

¹³⁷ Access Now et al., "Santa Clara Principles on Transparency and Accountability in Content Moderation."

Disinformation,¹³⁸ and AI-focused efforts like the OECD AI Principles,¹³⁹ encourage platforms to disclose information about their systems, algorithms, and AI applications. However, because these initiatives are non-binding they may provide incomplete data,¹⁴⁰ and companies have retreated from commitments with little consequence.¹⁴¹

Given the limitations of voluntary action, governments have moved toward regulation to require independent access to platform data. The DSA requires large platforms and search engines to provide researchers access to both public and non-public platform data for the study of systemic risks.¹⁴² In the US, proposed bills such as PATA¹⁴³ or DSOSA¹⁴⁴ would create obligations for platforms to share certain kinds of public platform data. Other jurisdictions, including the UK, Brazil, and Singapore, have passed or proposed laws that emphasize transparency and safety but vary in whether they guarantee data access for researchers. Many of these laws include explicit carveouts for certain types of data that should never be treated as public. For example, as originally introduced, PATA excludes private messages, biometric information, and precise geospatial information.¹⁴⁵ Privacy laws in countries around the world further implicate access and use of platform data.

Taken together, voluntary initiatives and regulatory frameworks reflect a global effort to improve transparency in digital governance. Voluntary measures have fostered norms and encouraged experimentation but remain incomplete and inconsistent. By contrast, regulatory regimes – especially the DSA – are creating enforceable pathways for researcher and regulator access to platform data, though implementation is nascent and legal debates persist. As platforms continue to shape the flow of online information, striking a balance between transparency, privacy, and accountability will be critical to enabling independent oversight and debate.

V. Core Concepts

The ongoing lack of agreed-upon definitions and mechanisms undermines efforts to operationalize predictable access to public platform data. This framework describes how platforms, researchers, and regulators across jurisdictions can enable consistent access to a minimum set of public platform data.

The scope and definition of public platform data differ across national legal regimes and platform access mechanisms. Finding congruence across very different contexts and regimes is challenging. Within this broader and disparate landscape, this framework focuses on a specific subset of public platform data: **high-influence public platform data**.

¹³⁸ European Commission, “The Code of Conduct on Disinformation.”

¹³⁹ Organisation for Economic Co-operation and Development, “AI principles.”

¹⁴⁰ Albert, “Platforms’ Promises to Researchers”; Mündges and Park, “But Did They Really?”

¹⁴¹ Democracy Reporting International, “Big tech is backing out of commitments countering disinformation.”

¹⁴² European Union, “Digital Services Act”, Article 40.

¹⁴³ United States, “Platform Accountability and Transparency Act.”

¹⁴⁴ United States, “Digital Services Oversight and Safety Act.”

¹⁴⁵ United States, “Platform Accountability and Transparency Act.”

High-influence public platform data refers to the subset of public platform data that is most consequential, either because of its reach, its source, or its role in shaping what people see online. It is the data that tells us who holds influence, what content spreads, and how platforms themselves amplify certain voices. High-influence public platform data is a minimum subset of the universe of public platform data that researchers should be able to access. But this minimum or top tier of data has clear definitional boundaries – content from accounts that shape society, or that reaches thousands of people – and therefore offers a unifying starting point in the discussion of how to approach access to public platform data more broadly.

This section defines core concepts relevant to the framework:

- A. Platforms:** The platforms that should provide research access to high-influence public platform data.
- B. Information Environments:** Information environments are the global, geographic, and linguistic environments where information is produced, disseminated, and consumed.
- C. Access Mechanisms:** Access mechanisms are the mechanisms by which researchers access high-influence public platform data.
- D. Reasonable Understanding of Publicness:** How users understand that their data is or may be publicly accessible, taking into account platform features and policies.

A. Platforms

A range of platforms may host high-influence public platform data. This section defines platforms for the purpose of operationalizing researcher access to high-influence public platform data.

Definition

Online platforms and services are public-facing websites, online services, online applications, or mobile applications that predominantly store, aggregate, and disseminate user-generated information to the public, including text, videos, images, games, audio files, or other content. Such platforms and services include, but are not limited to, social media, social networks, and search engines. Other platforms, including video game sites, content-sharing sites, virtual reality environments, e-commerce sites, online booking sites, app stores, and payment platforms may also host high-influence public platform data.

The framework builds on existing definitions of platforms. Existing and proposed regulations across a range of countries establish definitions of platforms, including:

- The DSA in the EU,¹⁴⁶
- India’s Intermediary Guidelines and Digital Media Ethics Code,¹⁴⁷

¹⁴⁶ “Online platform” means a hosting service that, at the request of a recipient of the service, stores and disseminates information to the public, unless that activity is a minor and purely ancillary feature of another service or a minor functionality of the principal service and, for objective and technical reasons, cannot be used without that other service, and the integration of the feature or functionality into the other service is not a means to circumvent the applicability of this Regulation.” See European Union, “Digital Services Act,” Article 3(i).

¹⁴⁷ India, “Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules”; Sarkar et al., “On the Legality and Constitutionality of the Information Technology,” noting that “‘social media intermediary’ means an intermediary

- Nigeria’s Code of Practice for Interactive Computer Service Platforms/Internet Intermediaries,¹⁴⁸
- The Online Safety Act in Australia,¹⁴⁹
- The Online Safety Act in the UK,¹⁵⁰ and
- PATA as proposed in the US.¹⁵¹

Most of these regulations, proposals, and guidelines focus on social media.¹⁵² However, the framework adopts a definition that could include a range of platforms that host, aggregate, and disseminate high-influence public platform data.

Legislators and regulators around the world are also increasingly working to mandate transparency in AI systems alongside platforms. The EU’s AI Act,¹⁵³ as well as proposals in Canada,¹⁵⁴ Brazil,¹⁵⁵ and other countries would also require greater transparency related to the training and implementation of AI systems. These AI transparency efforts may also require access to public platform data.

B. Information Environments

What is considered high-influence public platform data can vary depending on the specifics of the information environment. Online information environments range from sprawling global communities that cut across languages, geographies, and cultures to specific countries, regions, or communities. Understanding how information is produced, disseminated, and consumed within different types of information environments is foundational for a wide range of research and policy goals.

which primarily or solely enables online interaction between two or more users and allows them to create, upload, share, disseminate, modify or access information using its services.”

¹⁴⁸ “Interactive Computer Service Platforms” otherwise referred to as a “Platform” in this Code, means any electronic medium or site where services are provided by means of a computer resource and on-demand and where Users create, upload, share, disseminate, modify, or access information, including websites that provide reviews, and gaming Platforms;” and “Large Service Platforms” (LSP) means an Interactive Computer Service Platform/Internet Intermediary whose registered Users in Nigeria are more than one million (1,000,000). Nigeria, “Code of Practice for Interactive Computer Service Platforms/Internet Intermediaries.”

¹⁴⁹ “For the purposes ... of the Act, this Part specifies the basic online safety expectations for the following: (a) a social media service; (b) a relevant electronic service of any kind; (c) a designated internet service of any kind.” Australia, “Online Safety (Basic Online Safety Expectations) Determination 2022, F2022L00062.”

¹⁵⁰ For the purposes of the Act “‘user-to-user services’, which is defined as internet services where content is user generated, uploaded or shared and may be encountered by another user of the services. This includes platforms like social media sites or apps, photo and video-sharing services, chat and instant messaging services, dating apps and an online or mobile gaming services, among others.” United Kingdom, “Online Safety Act 2023.”

¹⁵¹ “Platform” means any entity ... that — (A) operates a website, desktop application, augmented or virtual reality application, or mobile application that—(i) permits a person to become a registered user, establish an account, or create a profile for the purpose of allowing the user to create, share, and view user-generated content through such an account or profile; (ii) enables one or more users to generate content that can be viewed by other users of the platform; and (iii) primarily serves as a medium for users to interact with content generated by other users of the platform and for the platform to deliver ads to users; and (B) has at least 50,000,000 unique monthly users in the United States for a majority of the months in the most recent 12-month period.” United States, “Platform Accountability and Transparency Act”, § 2(5).

¹⁵² UNESCO, “Guidelines for the Governance of Digital Platforms”; United Nations, “Global Digital Compact.”

¹⁵³ European Union, “Artificial Intelligence Act.”

¹⁵⁴ Canada, “Artificial Intelligence and Data Act.”

¹⁵⁵ Brazil, “Projeto de Lei n° 2338, 2023.”

Definition

Information environments are defined as:	
	Global A platform’s total user base across languages and geographies.
	Geographic Users of a platform in a particular geography, including a region (e.g., Latin America), country (e.g., Brazil), or major subnational region within a country (e.g., a state, province, or large city).¹⁵⁶
	Linguistic Users of a platform engaging in a specific language or languages (e.g., Portuguese).

Research focused on a particular information environment will have different data needs. If high-influence public platform data is only defined at the global level, users, communities, and researchers in specific information environments would be unable to understand public conversations in their own community. Data with 10,000 total unique views in India or the US, for example, would be a far smaller percentage of total views than 10,000 total views in Solomon Islands or Malta, with much smaller total populations and user bases.¹⁵⁷ Likewise, languages with fewer numbers of speakers will require different thresholds than speakers of the world’s most common languages. As will be described below, the framework incorporates both absolute and relative thresholds for high-influence public platform data to enable more meaningful and consistent definition in each information environment. These three information environments are some of the most commonly considered in research, but are not meant to be exhaustive, and other information environments could be examined for specific research projects.¹⁵⁸

User privacy risks can arise in each type of information environment. Ethical research frameworks should be used to manage risks in each type of environment.

C. Access Mechanisms

The framework establishes three data access mechanisms to allow for researcher access to high-influence public platform data. Multiple access mechanisms are needed to enable ethical

¹⁵⁶ The working group recognizes the technical difficulties of delineating geographic boundaries for platform data. Data flows across borders and there are challenges associated with geolocation and virtual private networks. Cáceres and Grant, “Geolocation.”

¹⁵⁷ With 1.4 billion people in India, 345 million in the US, 820,000 in Solomon Islands, and 540,000 in Malta. See Worldometer, “Countries in the world by population (2025).”

¹⁵⁸ Wanless et al., “Assessing National Information Ecosystems.” Furthermore, certain research may also take place in particular thematic information environments (for example, global Real Madrid fans). Thematic research areas are expansive and the framework prioritizes the following environments as a starting point.

research in specific information environments, while also limiting infrastructure costs for platforms. While these mechanisms should allow for access to data beyond high-influence public platform data, this framework articulates a minimum set of expectations for each mechanism.

Definition

Access mechanisms are defined as:	
	Proactive Data Interface A structured, platform-supported access mechanism that provides ongoing, real time, predictable researcher access to high-influence public platform data. Examples include research APIs, searchable archives, or downloadable datasets that are updated regularly without requiring individual requests.
	Custom Data Requests Tailored datasets provided upon request to meet the specific local or thematic needs of researchers. These requests may be fulfilled through bespoke datasets, tailored APIs, archives, or other methods.
	Independent Data Collection Researcher-initiated access to high-influence public platform data, typically through automated collection such as scraping or crawling. This mode of access preserves researcher independence and enables validation of platform-provided data.

Note that Section VI uses the above icons to designate what types of high-influence public platform data should be enabled through which data access mechanism.

Some platform access mechanisms and regulations define the types of researchers eligible to access different types of platform data. Others introduce a purpose limitation on access, requiring specific agreement on how data will be used. By focusing on the top tier of public platform data, this framework seeks to enable a variety of researchers to access high-influence public platform data, without requiring expansive vetting of researchers.

Proactive Data Interface

Platforms should proactively establish efficient and effective data access mechanisms for researcher access to high-influence public platform data. They may develop these capacities internally, such as the MCL, or partner with a third party to develop them, such as CrowdTangle.

Proactive Data Interfaces should be provided through research APIs, libraries, archives, third-party mechanisms, or other mechanisms necessary to activate meaningful access. Proactive provision of

high-influence public platform data through these mechanisms is a vital form of access as it should require fewer technical skills or resources than other access mechanisms. This crucial point of access allows a broad range of researchers, including journalists and civil society organizations, to understand online information systems.

Platforms should provide access to high-influence public platform through a Proactive Data Interface, when platforms reach specific sizes in specific information environments:



Global

When a platform has 10 million monthly active global users, it should provide a Proactive Data Interface to provide researcher access to a global set of high-influence public platform data. Micro- or small-sized enterprises may be excluded from requirements to enable this form of data access.¹⁵⁹



Geographic

When a platform has monthly active users in a particular country that surpass 10% of the country's population, the platform should establish a Proactive Data Interface specific to that country. In this instance, the Proactive Data Interface should provide access to high-influence public platform data specific to that information environment across the four elements of the framework described below. For example, if a platform's user base includes more than 10% of the population of Albania, Algeria, or Argentina, it should create a Proactive Data Interface that allows for access to geographically relevant HDC, Government and Political Accounts, Notable Public Accounts, and Business Accounts and Promoted Content in each country.

The framework uses 10% of a country's population as a baseline in an attempt to strike a reasonable balance between emerging regulatory approaches and proposals mandating data access. Existing and proposed platform regulations establish differentiated data access expectations based on platform size. The DSA defines Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) as platforms which are used by at least or 45 million European citizens per month, corresponding to 10% of the EU population.¹⁶⁰ Given the outsized impact platforms of this size have on society, they are subject to heightened expectations.¹⁶¹ India's Intermediary Guidelines and Digital Media Ethics Code defines significant social media intermediaries as more than 5 million Indian users,

¹⁵⁹ The European Commission, for example, defines micro enterprise as fewer than 10 employees and an annual turnover or balance sheet below €2 million; small enterprise: fewer than 50 employees and an annual turnover or balance sheet below €10 million; and medium-sized enterprise: fewer than 250 employees and annual turnover below €50 million or balance sheet below €43 million. See European Commission, "Micro-, Small- and Medium-Sized Enterprises." See, also Keller, "Comment to Council of Europe Committee of Ministers on the Draft Recommendation on online safety and empowerment of content creators and users" suggesting data access requirements should be proportionate to the "size, role, and impact" of platforms.

¹⁶⁰ The specific terminology in the DSA is "recipients of a service" as opposed to users.

¹⁶¹ European Union, "Digital Services Act," Article 40.

or 0.03% of the Indian population.¹⁶² As drafted, PATA defines in-scope platforms as having a minimum of 50 million unique monthly users,¹⁶³ or roughly 15% of the population of the US. In Nigeria, Large Service Platforms are defined as more than 1 million users, or 0.4% of the population.¹⁶⁴



Linguistic

When a platform has monthly active users in a particular language that surpass 10 million monthly active users, a Proactive Data Interface should provide high-influence public platform data in that language. This should include data specific to that language across the four elements of the framework. For example, a platform that has 25 million monthly active users whose default language is French should provide a Proactive Data Interface that includes HDC, Government and Political Accounts (where relevant), Notable Public Accounts, and Business Accounts and Promoted Content in French. This mirrors the global threshold and seeks to account for researcher needs in particular linguistic environments.

Principles for the Proactive Data Interface

Proactive Data Interfaces should be offered consistent with the following principles:

1. Free to access
2. Searchable research interface
3. Real time data access
4. Downloadable data sets, as long as data protection, data security, and ethical safeguards are in place
5. Reasonable and transparent researcher rate limits
6. Non-proprietary
7. Shareable, as long as data protection, data security, and ethical safeguards are in place
8. Predictable, with any changes to the system communicated to users with sufficient notice
9. Auditable, to verify the accuracy of data provided through the mechanism (including through Independent Data Collection, discussed below)

Data accessed through Proactive Data Interfaces should be provided via a searchable interface, such that keywords or trending topics can be identified absent knowing which accounts to search for specifically, and in a way that allows for longitudinal analysis of data over time. Proactive Data Interfaces should allow for data to be downloadable to facilitate analysis, the format of the data should

¹⁶² India, “Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules”; Sarkar et al., “On the Legality and Constitutionality of the Information Technology,” noting that “‘significant social media intermediary’ means a social media intermediary having number of registered users in India above such threshold as notified by the Central Government [set at 5 million Indian users].”

¹⁶³ “Platform” means any entity ... that ... B) has at least 50,000,000 unique monthly users in the United States for a majority of the months in the most recent 12-month period.” United States, “Platform Accountability and Transparency Act,” § 2(5).

¹⁶⁴ “Large Service Platforms” (LSP) means an Interactive Computer Service Platform/Internet Intermediary whose registered Users in Nigeria are more than one million (1,000,000). Nigeria, “Code of Practice for Interactive Computer Service Platforms/Internet Intermediaries.”

not be proprietary, and researchers should be able to share data with one another as long as data protection and data security safeguards are met.¹⁶⁵

Given the importance of standardization across platforms, the largest platforms should use third-party verification audits to ensure the accuracy of data provision under the Proactive Data Interface.¹⁶⁶ The DSA, for example, requires VLOPs and VLOSEs to conduct annual audits of obligations, including in relation to data access under DSA Article 40. Audits of Proactive Data Interface accuracy could help ensure that all of the required fields of data are being provided, that the backend methodologies for gathering data are robust, and that access proceeds as expected. Without regular and robust audits, platform compliance could suffer.¹⁶⁷ Independent Data Collection should also serve as a check of accuracy.



Custom Data Requests

Specific research concerns will invariably include high-influence public platform data beyond what is offered through the Proactive Data Interface. For example, researchers working to understand child safety issues in a particular subnational region of a country might require high-influence public platform data beyond what is included in the Proactive Data Interface. For that reason, platforms should also respond to researcher requests for customized data sets across information environments, including cases where monthly active users fall below the established global, geographic, and linguistic mechanism thresholds for Proactive Data Interfaces.



Independent Data Collection

Platforms should enable Independent Data Collection of high-influence public platform data across all information environments. While Proactive Data Interfaces, including platform APIs, provide important access to public platform data, their accuracy is frequently questioned by researchers. This underscores the need for independent, automated collection as it provides “independence in how data is collected.”¹⁶⁸ By enabling multiple access mechanisms to high-influence public platform data, this framework ensures comprehensive and responsive data access. One form of data access cannot respond to all research questions.

Platforms and regulators should take steps to facilitate this access mechanism. First, legal protections should be in place to ensure researchers can safely engage in research with high-influence public platform data.¹⁶⁹ Second, platforms should not prohibit or restrict, either technically or via their terms

¹⁶⁵ Freelon, “Computational Research in the Post-API Age.”

¹⁶⁶ Pierce, “‘Views’ are lies.”

¹⁶⁷ Brown, “The Problem with TikTok’s New Researcher API is Not TikTok.”; Darius, “Researcher Data Access Under the DSA.”

¹⁶⁸ Scott, “Survey.”

¹⁶⁹ Abdo et al., “A Safe Harbor for Platform Research.”

of service,¹⁷⁰ independent access to high-influence public platform data.¹⁷¹ Lastly, as for all other access mechanisms, researchers using these tools should adhere to ethical and data protection standards.

Summary of Platform-enabled Access Mechanism Thresholds

	Access Mechanism		
	 Proactive Data Interface	 Custom Data Requests	 Independent Data Collection
 Global	>10 million monthly active global users	No minimum threshold, subject to a lower bound	
 Geographic	>10% of a country's population are monthly active users	No minimum threshold, subject to a lower bound	
 Linguistic	>10 million monthly active users in a particular language	No minimum threshold, subject to a lower bound	

D. Reasonable Understanding of Publicness

The *Better Access* framework is scoped to data that users would reasonably understand to be public. Because privacy is contextual – what a user considers to be private or public may depend on the circumstances – it is challenging to draw clear boundaries around this scope. This section explains the factors used to determine the absolute and relative thresholds for public platform data described in Section VI, and how those factors influence users' expectations about whether their information should be considered public and available for research. The purpose of defining what is meant by reasonable understanding of publicness is to align data access regimes with what an average user would reasonably expect, following existing privacy standards and policy best practices.

Definition

A **reasonable understanding of publicness** exists when a user interacting with a platform would reasonably understand that their data is or may be publicly accessible, taking into account platform features and policies.

¹⁷⁰ Facebook, “Terms of Service.”

¹⁷¹ Davidson et al., “Platform-controlled social media APIs threaten open science.”

Both privacy and risks are contextual, and preferences about both will vary across any given platform's user base. Considering multiple factors when determining reasonable understanding of publicness is an approach that seeks to advance user privacy, enable ethical research, and account for the ways in which information flows within our online information ecosystem. To read more about research on privacy and its contextuality, see Appendix B.

The following factors inform the absolute and relative thresholds for public platform data described in Section VI:

- **Platform Policies and User Settings:** Platforms may have terms of service and privacy policies that default towards broad public availability. User-facing policies are often written in technical language.¹⁷² Many users do not engage with these policies and the complexity of policies are a key reason users do not engage with or sufficiently understand policies and terms of service for platforms.¹⁷³ Platforms often default users to settings that enable the public visibility of data.¹⁷⁴ Such defaults are “sticky,”¹⁷⁵ given that platform design interfaces can make it difficult for users to modify their default settings.¹⁷⁶ Network effects and design patterns may further nudge users towards defaults for public availability.¹⁷⁷

That said, several platforms have worked to improve user-facing communication around platform policies. Meta, for example, has a central page to explain their privacy settings and tools.¹⁷⁸ Snap,¹⁷⁹ Spotify, and YouTube¹⁸⁰ all provide similar explainers. Platforms also maintain terms of service for developers and, occasionally, researchers who access platform-provided research tools. Nonetheless, even with improvements in how policies are communicated, users do not always understand when their data is private and when it is or may be publicly accessible.

Additionally, platforms' terms of service may also include provisions that establish expectations for permissible research. Users may be asked to agree to restrictions on data access that intersect with relevant national or regional regulation that requires data access (e.g., the DSA).

- **Platform User Experience:** Platform design offers users signals for whether data is or may be publicly available. Platforms communicate to brands about organic and inorganic ways to expand the reach of content, signalling that greater public visibility is expected and desirable.¹⁸¹ Similarly, platform marketing for businesses can be a signal of how content can be expected to be seen outside their immediate network, both through engagement and paid advertising.¹⁸²

¹⁷² Lomas and Dillet, “Terms and Conditions Are the Biggest Lie of Our Industry.”

¹⁷³ Benoliel and Becher, “The Duty to Read the Unreadable”; Obar and Oeldorf-Hirsch, “The Biggest Lie on the Internet.”

¹⁷⁴ Kelly, “Venmo privacy settings to change now.”

¹⁷⁵ Ben-Shahar and Pottow, “On the stickiness of default rules.”

¹⁷⁶ Beaumont, “GPEN Sweep 2024 on deceptive design patterns.”

¹⁷⁷ Marwick and boyd, “Networked Privacy”; Waldman, “Cognitive biases, dark patterns, and the ‘privacy paradox’.”

¹⁷⁸ Facebook, “Basic Privacy Settings & Tools.”

¹⁷⁹ Snap, “How Do I Change My Privacy Settings on Snapchat?”

¹⁸⁰ YouTube, “Privacy.”

¹⁸¹ LinkedIn, “Boost a post to reach a wider audience.”

¹⁸² TikTok, “Find and Reach Your Target Audience.”

- **Account Verification:** Platforms provide opportunities for certain categories of users to verify the authenticity of their accounts. For example, YouTube,¹⁸³ Meta,¹⁸⁴ Spotify,¹⁸⁵ TikTok,¹⁸⁶ and other platforms allow some individuals, public businesses, and content creators to verify their accounts, signaling intent for public reach. These platforms, alongside Bluesky,¹⁸⁷ may also verify government officials and journalists in some countries and some instances. Verification may also require accounts to have significant reach or engagement (YouTube¹⁸⁸), or to pay a fee (LinkedIn,¹⁸⁹ X¹⁹⁰). In both cases, these requirements can communicate a user's desire for, and understanding of, publicness.
- **Account Access Requirements:** Platforms may either be free for users, have a fee, or have both a free and a fee-based option. Both free and fee-based platforms may host public platform data. Similarly, some platforms require a login to the platform in order to view its content. These platforms also may host public platform data. A fee or requirement for a login does not mean information is necessarily private given some public platforms require accounts (e.g., X or Facebook).

Grounding definitions for public platform data in whether a typical user reasonably understands their information is public is an approach rooted in existing policy frameworks, proposals, and good practices.

Australia's Privacy Principles, for example, consider a user's reasonable expectations when assessing disclosure of personal information.¹⁹¹ The Brazilian General Data Protection Law speaks to the legitimacy of a data subject's expectation of privacy.¹⁹² In the US, a user's reasonable expectation of privacy is a key element in privacy law¹⁹³ that has been used in consumer protection¹⁹⁴ and legislative proposals around public access to platform data.¹⁹⁵ Europe's GDPR similarly references expectations

¹⁸³ Google, "YouTube Partner Program overview & eligibility."

¹⁸⁴ Meta, "Partner Monetization Policies."

¹⁸⁵ Spotify, "Spotify for Creators' Monetization Options."

¹⁸⁶ TikTok, "Verified Accounts on TikTok."

¹⁸⁷ Bluesky, "Bluesky for Journalists."

¹⁸⁸ YouTube, "Verification badges on channels."

¹⁸⁹ LinkedIn, "Verifications on your LinkedIn profile."

¹⁹⁰ X, "X Verification requirements."

¹⁹¹ Australia, "Australian Privacy Principles."

¹⁹² IAPP, "Brazilian General Data Protection Law."

¹⁹³ Busby, "expectation of privacy"; Supreme Court of the United States, "Katz v. United States."

¹⁹⁴ Federal Reserve, "Federal Trade Commission Act Section 5."

¹⁹⁵ The Platform Accountability and Transparency Act defines "reasonably public" as "information that the author made available in a manner and under such circumstances such that the author does not retain a reasonable expectation of privacy in the information. The fact that a user may need to register or create an account with a platform to view information does not preclude it [from] being deemed reasonably public." See United States, "Platform Accountability and Transparency Act," § 9(i)(6). The American Data Privacy and Protection Act defines "publicly available information" as "any information that a covered entity or service provider has a reasonable basis to believe has been lawfully made available to the general public from—...(iii) a website or online service made available to all members of the public, for free or for a fee, including where all members of the public, for free or for a fee, can log in to the website or online service." See United States, "American Data Privacy and Protection Act."

of data subjects, and sets standards for privacy safeguards in relation to disclosure.¹⁹⁶

These legal and legislative standards are mirrored by independent efforts to advance ethical and privacy respecting research on digital platforms. The European Digital Media Observatory (EDMO) Working Group on Platform-to-Researcher Data Access recommends consideration of whether a reasonable user would understand whether their data is publicly available when considering necessary ethical safeguards.¹⁹⁷ These recommendations include the consideration of “what a reasonable data subject would understand regarding the extent to which their data is publicly available, taking into account platform features and conventions.”¹⁹⁸ The Reasonable Expectations and Potential Impact Risk Framework included in the code suggest that the greater the expectation of publicness the lesser the privacy risk (while noting that some risk may nonetheless remain).¹⁹⁹

¹⁹⁶ European Union, “General Data Protection Regulation.”

¹⁹⁷ The Working Group on Platform-to-Researcher Data Access recommends considering: a) The likelihood that the data has reached or will reach a substantial public audience. (The posts of a notable public account, for example, might be considered more in the public domain than those of an ordinary user.) and b) What a reasonable data subject would understand regarding the extent to which their data is publicly available, taking into account platform features and conventions. See Working Group on Platform-to-Researcher Data Access, “Report of the European Digital Media Observatory’s Working Group on Platform-to-Researcher Data Access.”

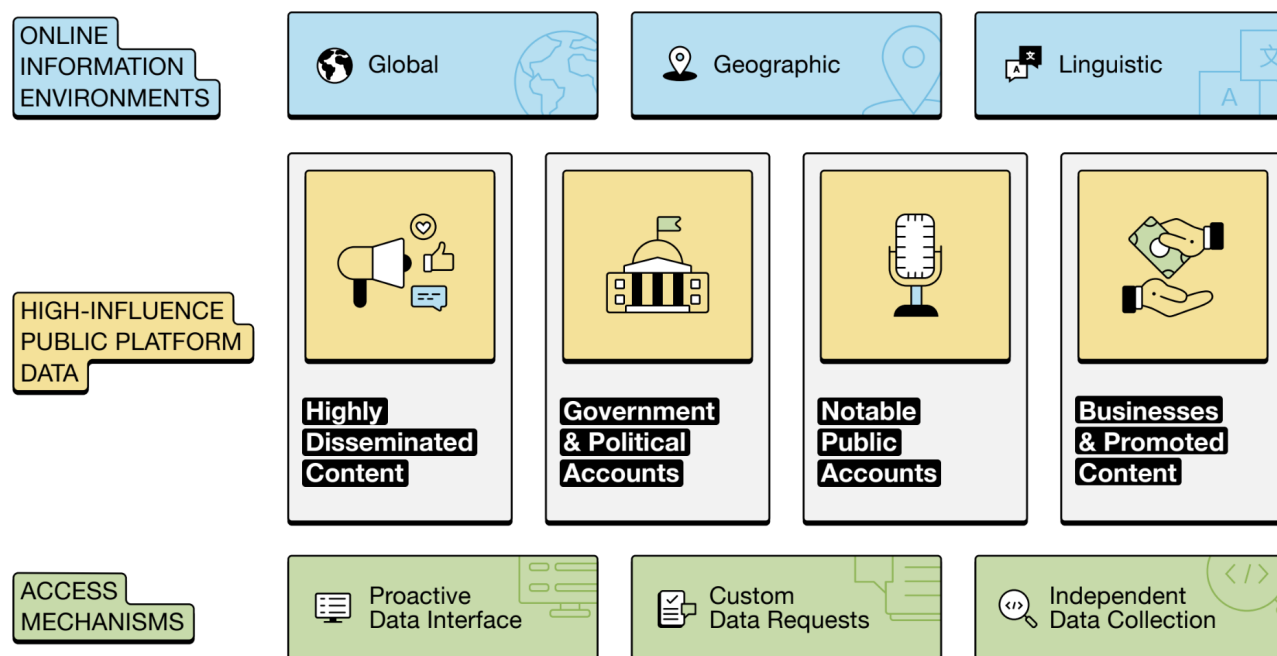
¹⁹⁸ Ibid.

¹⁹⁹ Ibid.

VI. Framework for High-Influence Public Platform Data

This framework establishes the minimum expectations for researcher access to high-influence public platform data. Platforms should be expected to enable researchers to access high-influence public platform data through each data access mechanism and online information environment.

The Better Access Framework



The framework identifies four categories of high-influence public platform data:



Highly Disseminated Content: Posts or videos that achieve exceptional reach or engagement, shaping the public agenda.



Government and Political Accounts: Posts from accounts belonging to elected officials, candidates, political parties, and government institutions.



Notable Public Accounts: Content from accounts belonging to celebrities, journalists, civic leaders, or other public figures whose reach gives them outsized influence.



Business Accounts and Promoted Content: Advertising and commercial messaging, which can sway consumer behavior, public health, or public trust.

For each category, this section spells out criteria for how relevant data is defined, as well as the ways it should be accessed. These categories are defined by clear numeric thresholds, while broadly taking account of a general user's reasonable understanding of publicness at scale.

The absolute and relative numeric thresholds specified for each category below are designed to provide clarity and predictability for researchers, policymakers, and platforms. They are informed by experiences conducting independent research as well as platform data access programs. They should be operationalized alongside implementation of strong researcher ethics and privacy safeguards. The definition of what is considered high-influence public platform data can be subject to a lower bound – specific information environments so small that their data should no longer be treated as public. This lower bound is necessarily contextual and should be informed by a user's reasonable understanding of publicness and ethical research frameworks.

For data to be made available via data access mechanisms, the accounts, groups, pages, or other platform-specific elements must be set to public. Data that a user restricts to a private audience is, by definition, not public platform data.

A. Highly Disseminated Content

HDC refers to posts or videos that achieve exceptional reach or engagement, shaping public discourse. Historically, platform data access mechanisms have enabled research with HDC. HDC provides a critical lens for understanding how information circulates online, including how platforms influence the visibility of information and how users encounter, interact with, and disseminate content. To support consistent access to HDC across different information environments, this framework defines HDC using clear absolute and relative thresholds for reach and engagement.

Definition

Content may be highly disseminated through either reach or engagement, defined as:	
Reach	Content that has either:  Proactive Data Interface Greater than 10,000 instances of total distribution, including unique views, listens, or downloads; or  Custom Data Requests  Independent Data Collection Top 2% of cumulative weekly platform reach, including unique views, listens, or downloads.
Engagement	Content that has either:  Proactive Data Interface Greater than 1,000 instances of aggregate platform engagement; or  Custom Data Requests  Independent Data Collection Top 2% of cumulative weekly platform engagement.

Research routinely finds that reach and engagement on many digital platforms follow power law or heavy-tailed distributions. A small minority of user accounts or content capture a disproportionately large share of total attention. Beginning in the early days of social media, empirical studies have noted power law distributions of account and content popularity.²⁰⁰ Recent studies have explored these dynamics on YouTube,²⁰¹ TikTok,²⁰² Facebook,²⁰³ among other platforms.

Rationale for Absolute Thresholds for HDC

Because of this power law distribution, for most platforms only a small percentage of content accounts for the majority of reach and engagement.²⁰⁴ Setting absolute thresholds at 10,000 or more instances of total distribution and 1,000 engagements captures this influential subset while excluding the vast majority of platform data. The absolute HDC threshold thus captures the small percentage of data that makes up the majority of reach and engagement for most platform users. This absolute HDC threshold enables clear, bright line boundaries for platforms to include in the Proactive Data Interface.

²⁰⁰ Johnson et al., “Emergence of Powers Laws in Online Communities.”

²⁰¹ McGrady et al., “Dialing for Videos”; Munger et al., “Pressing Play on Politics.”

²⁰² Steel et al., “Just Another Hour on TikTok.”

²⁰³ Edelson et al., “Measurement and Metrics for Content Moderation.”

²⁰⁴ Narayanan, “Understanding Social Media Recommendation Algorithms.”

Rationale for Relative Thresholds for HDC

Absolute thresholds, however, will not translate across information environments equally. 10,000 views in India or the US, for example, would be a far smaller percentage of total views than in Solomon Islands or Malta, with much smaller total population and user bases.²⁰⁵ As such, relying only on absolute HDC thresholds would prohibit effective research related to information environments – like small countries or languages with fewer speakers – where content would less frequently reach the threshold of 10,000 views.

For this reason, HDC is also defined through relative thresholds for reach and engagement, namely the top 2% of reach or engagement in a particular information environment. Relative HDC thresholds enable consistent definition of high-influence public platform data in each information environment.

When using a relative HDC threshold, however, there will be a point where the information environment is too small to be considered to contain HDC. Those accounts posting content in the top 2% of data by reach or engagement in a small information environment might be seen or engaged with by only a handful of people. Account holders likely would not expect such posts to be considered HDC, or their data to be high-influence in accordance with this definition. The framework does not recommend an absolute number for the minimum size of an information environment as this would be contextual on the facts of the research. Researchers are expected to carefully consider user privacy and to advance ethical practices when developing their methodologies.

Transparency and Verification of HDC Reach and Engagement Metrics

Platforms should publicly disclose their methodologies for determining reach and engagement. Some platforms describe how reach or engagement metrics are developed, but others are less transparent.²⁰⁶ Reach and engagement metrics can be inconsistent across platforms and have been contested.²⁰⁷ Advertisers, for example, sued Facebook in 2018 claiming that Facebook systematically inflated view counts to increase advertising revenue.²⁰⁸ Adversarial actors have taken note and studies have shown how bots can be used to increase reach²⁰⁹ as well as engagement metrics.²¹⁰ A 2025 study of TikTok, for example, found that 46% of *all* comments in a sample of 18.4 million comments consisted of only emojis, which the authors suggest “could imply large quantities of inauthentic activity, as emojis would be an easy way to boost a video’s engagement while requiring limited effort on behalf of the bot.”²¹¹ As such, disclosure of platform methodologies and regular accuracy audits, including through Independent Data Collection, will be critical for the effectiveness of the framework.

²⁰⁵ With 1.4 billion people in India, 345 million in the US, 820,000 in Solomon Islands, and 540,000 in Malta. Worldometer, “Countries in the world by population (2025).”

²⁰⁶ Apple, “Listener analytics”; X, “About view counts.”

²⁰⁷ Pierce, “‘Views’ are lies”; Silberling, “Spotify responds to backlash over public podcast play counts.”

²⁰⁸ Welch, “Facebook may have knowingly inflated its video metrics for over a year.”

²⁰⁹ Kuchhal and Li, “A View into YouTube View Fraud.”

²¹⁰ Aljabri et al., “Machine learning-based social media bot detection.”

²¹¹ Steel et al., “Just Another Hour on TikTok.”

Access Mechanisms for HDC

HDC should be accessible through multiple access mechanisms to enable research in different information environments. The following section describes the frameworks' HDC thresholds across two sub-elements: reach and engagement.



Proactive Data Interface

Reach – Greater than 10,000 instances of total distribution: A reach of greater than 10,000 strikes a balance between research goals, a user's reasonable expectations of privacy, and platform implementation considerations. The absolute threshold for the Proactive Data Interface allows for platforms to have clear boundaries for inclusion of data across information environments.

Research shows that content exceeding 10,000 views represents a small percentage of total content yet accounts for the majority of total public views. Studies focused on YouTube show a highly skewed distribution of views. For instance, a 2023 study analyzing a random sample of approximately 10,000 videos on YouTube found that only 3.67% of videos surpassed 10,000 views, but these accounted for 93.61% of all views.²¹² According to an analysis of data donated to the *Washington Post* of approximately 900 US users' TikTok watch histories, videos with a published view count greater than 10,000 (measured weeks-to-months after the fact) represent 91.4% of all videos served to users in the study.²¹³

Engagement – Greater than 1,000 engagements: Including content with greater than 1,000 distinct engagements in Proactive Data Interfaces balances research objectives and user expectations of privacy. The absolute engagement threshold also allows for clear boundaries for inclusion of data in the Proactive Data Interface.

The YouTube study referenced above found that just 1% of videos generated more than 100 comments, representing over half of all comments analyzed.²¹⁴ Reactions on YouTube (in this case, likes) are also not evenly distributed. The study found that the mean number of likes per video in the study was just over 16, with just eight videos (.08% reviewed) having nearly 55% of likes in the study.²¹⁵ A 2025 study of TikTok similarly identified an unequal distribution of engagement, with 51.9% of videos receiving no comments.²¹⁶

²¹² McGrady et al., "Dialing for Videos."

²¹³ The 95th percentile video by view count in the sample of videos shown to participants in that experiment has 4,400,000 views. Videos in the top 5% of videos by views represent 24.3% of all videos served to participants in the Post's TikTok data donation experiment. This is a suboptimal estimate of the distribution of views on TikTok videos because it is skewed by videos shown to participants in the study, where there are presumably many videos shown to effectively no one that should be used to calculate the distribution of view counts, but which are excluded from my sample.

²¹⁴ The mean number of comments per video in the sample is 5.32, with a median of 0, maximum of 856, and minimum of 0. Most videos (72.64%) had no comments. The 1.04% of videos with more than 100 comments (104) account for more than half (54.60%) of all comments in the sample. See McGrady et al., "Dialing for Videos."

²¹⁵ Ibid. The mean number of likes in the sample was 16.48, with a median of 0, maximum of 17,517, and minimum of 0. Even more videos have no likes than have no comments (8,884 or 88.71%). The log-transformed distribution is in Figure 8, and Q-Q plot in Figure 9. Eight videos (.08%) account for 54.91% of all likes in the sample.

²¹⁶ Steel et al., "Just Another Hour on TikTok."

The 1,000 engagements threshold allows for meaningful research on high-influence discourse while minimizing privacy risks by excluding the vast majority of user-generated content.



Custom Data Requests AND



Independent Data Collection

Reach – Content within the top 2% of weekly platform reach: A relative threshold of content in the top 2% of weekly platform reach will enable researchers to examine high-influence content representing a majority of total views in a given information environment. The YouTube study referenced above found that the top 3.67% of views accounted for 93.61% of the total views in the study.²¹⁷ A study of the distribution of views on Twitter similarly found a “top heavy” distribution of attention, where a small percentage of total content constituted the majority of views.²¹⁸

Engagement – Top 2% of cumulative weekly platform engagement: The distribution of engagement follows a broadly similar distribution to that of reach. The YouTube study confirms a skewed distribution of comments.²¹⁹ A 2024 study of YouTube found that the top 2% of commentators write 50% of all of the total comments on YouTube.²²⁰ The 2021 Twitter study found similar dynamics, with likes, retweets, replies, profile clicks and quote Tweets overwhelmingly clustered alongside the highest distribution of views.²²¹ A study of public content on Facebook published in 2025 also identifies a similar pattern across multiple language ecosystems, finding that the top 1% of content in U.S. English made up 58% of total dissemination (as well as 45% in Ukrainian and 57% in Russian).²²²

Guidance

Platforms have used absolute HDC thresholds for some time. This approach has enabled research and is operationally feasible for platforms. Appendix C summarizes thresholds for platforms’ Proactive Data Interfaces.

Policy proposals have incorporated both absolute and relative HDC thresholds. EDMO has recommended absolute thresholds to understand the platform data reach,²²³ but to estimate the prevalence of information types, EDMO recommends analyzing “a random sample of public content

²¹⁷ McGrady et al., “Dialing for Videos”; See also Munger et al., “Pressing Play on Politics.”

²¹⁸ Lazovich et al., “Measuring Disparate Outcomes of Content Recommendation Algorithms with Distributional Inequality Metrics.”

²¹⁹ The mean number of comments per video in the sample is 5.32, with a median of 0, maximum of 856, and minimum of 0. Most videos (72.64%) had no comments. The 1.04% of videos with more than 100 comments (104) account for more than half (54.60%) of all comments in the sample. See McGrady et al., “Dialing for Videos.”

²²⁰ Yang et al., “Coordinated link sharing on Facebook.”

²²¹ Lazovich et al., “Measuring Disparate Outcomes of Content Recommendation Algorithms with Distributional Inequality Metrics.”

²²² Edelson et al., “Measurement and Metrics for Content Moderation.”

²²³ Working Group on Platform-to-Researcher Data Access, “Report of the European Digital Media Observatory’s Working Group on Platform-to-Researcher Data Access.”

weighted by views (10,000 views – but to be adapted to population size of a [European] member state) in the monitored period, per member state and language.”²²⁴

Regulatory proposals also focus on enabling access to HDC. DSA Article 40 specifically requires platforms to provide access to “real-time” data that is “publicly accessible in their online interface,”²²⁵ including HDC (as well as other data). As originally drafted, PATA in the US includes an absolute threshold for HDC.²²⁶ The proposal for DSOSA would require platforms to provide certified researchers access to “high-reach” and “high-engagement” public content and tasks the FTC with defining these terms.²²⁷

Summary of Highly Disseminated Content Thresholds

	Access Mechanism		
	 Proactive Data Interface	 Custom Data Requests	 Independent Data Collection
Reach	>10k total views	Top 2% cumulative weekly views	
Engagement	>1k total engagements	Top 2% cumulative weekly engagement	

B. Government and Political Accounts

The important role government officials play in making decisions that impact people’s everyday lives makes it critical for the public to understand the types of information these officials are sharing in our online information ecosystems. For this reason, Government and Political Accounts are included in the framework, referring to content from accounts belonging to elected officials, candidates, political parties, and government institutions.

²²⁴ Nenadić et al., “Structural Indicators of the Code of Practice on Disinformation.”

²²⁵ European Union, “Digital Services Act”, Article 40.

²²⁶ United States, “Platform Accountability and Transparency Act,” § 9. Section 9 recommends that the Federal Trade Commission (FTC) define an appropriate definition of “highly disseminated,” that “may include engagement, views, reach, impressions, or other metrics, provided that a piece of content must have been viewed by at least 10,000 unique users to qualify.”

²²⁷ United States, “Digital Services Oversight and Safety Act.”

Definition

Data associated with an account belonging to a major political party, candidate, government official, or institution, defined as:		
 Proactive Data Interface	 Custom Data Requests	 Independent Data Collection
National level	National and Subnational level	
Persons elected, appointed, or assigned to a major government position		
Registered political candidates running for office		
Political parties and major party officials		
Government institutions and related officials		

Implementation Discussion

Government is defined to include all branches of government, and is inclusive of national, international, and regional governments. At the national level, government officials include:

- Heads of state: monarch, emir, president, etc.
- Executive leaders: president, prime minister, etc.
- Members of a national legislature: congress, parliament, national assembly, etc.
- Members of national court bodies: Supreme Court, High Court, Constitutional Court, etc.

Government officials increasingly use digital platforms²²⁸ to strategically communicate with citizens²²⁹ and signal political agendas.²³⁰ Furthermore, studying public communication from government officials is essential for transparency and public trust.

²²⁸ Connolly et al., "Explaining the varying levels of adoption of e-government services in American municipal government"; Norton and Shapiro, "How to Better Study—and Then Improve—Today’s Corrupted Information Environment."

²²⁹ Kreiss et al., "In Their Own Words."

²³⁰ Russell, *Tweeting Is Leading*.

A government official may use different types of accounts, including an official government account, a campaign account, and/or a personal account. Each of these accounts and related data can be considered high-influence public platform data, according to the definitions below.

It is also worth noting explicitly that it is generally easier to define these categories in democracies, in which specific political processes like elections, candidate registration, and reporting requirements exist, as compared to other forms of government where such processes are absent. The framework is equally applicable in non-democracies, but there may be increased difficulty in articulating who is a political entity in those contexts.

1. Persons elected or assigned through a political process to a major government position

Public office holders are included in Government and Political Accounts. It would be simplest to apply this directly: anyone elected or assigned to a government position should be included in the definition of high-influence public platform data, given their status and expectation that they speak publicly. However, such a comprehensive definition poses both implementation challenges and privacy and ethical trade-offs, given the varied nature of levels of government across countries.

Around the world, there are millions of international, national, and local government officials. In India, for example, there are over 3 million elected representatives at the local level,²³¹ and 800 members of the national parliament.²³² In Brazil, there are more than 5,000 local governments, each with mayors and municipal legislative chambers.²³³ In Myanmar, there are more than 600 members of the national legislature (Pyidaungsu Hluttaw), many of whom are unelected or are assigned the position through unelected means.²³⁴ This presents a logistical challenge for maintaining lists of current elected officials at every level of government on a global scale, and underscores that not all government officials would have the ability to speak in an official capacity nor reasonably believe they were doing so.

To operationalize a manageable definition of a major government position, this framework includes the following:



Proactive Data Interface

For the National Level: Platforms should identify and provide data related to major government officials who are elected or appointed to national office in the Proactive Data Interface. This includes any heads of state, executive leaders, any members of a national legislature, and any members of national court bodies.

²³¹ India Development Review, “Local Government in India.”

²³² Sabha, “Members of Parliament.”

²³³ Marengo, “Local Government, Brazil.”

²³⁴ Myanmar, “Pyidaungsu Hluttaw.”

**Custom Data Requests AND****Independent Data Collection**

For the National and Subnational Level: In many countries, there are subnational government designations (states, provinces, regions, cities, municipalities, districts, etc.). Provided they meet the requirements of publicness outlined above, the data produced by elected or appointed government leaders within these subnational designations are high-influence public platform data. Research projects may therefore necessitate the compilation of elected officials and government positions below the national delegation. In these instances, independent researchers should be able to identify relevant subnational accounts or content and pursue the associated data through Custom Data Requests and Independent Data Collection mechanisms.

As some individuals may use public personal or campaign accounts for official purposes,^{235,236} independent researchers should also be able to access such data if not provided by a platform.

2. **Registered political candidates running for office, beginning from when they register as a political candidate to when they are elected, or for up to 120 days after the election if candidate is not elected to position**

Data from campaign or public accounts of registered candidates who aspire to the governing positions outlined above including heads of state, executive leaders, members of a national legislature, and members of national court bodies should be considered high-influence public platform data.

**Proactive Data Interface**

For the National Level: Platforms should proactively identify registered candidates running for national office, and include data from those candidates in the Proactive Data Interface. This also includes any registered candidate running or those in a succession line for heads of state, executive leaders, or campaigning to be members of a national legislature or national court body.

**Custom Data Requests AND****Independent Data Collection**

For the National and Subnational Level: For research projects or information environments that require granular analysis of candidates beyond the national level, independent researchers should be able to identify relevant accounts and pursue associated data through Custom Data Requests and Independent Data Collection mechanisms. This category also includes other figurehead political

²³⁵ The United States Supreme Court has put forth a two-part test as to when a government official's personal account is used for official purposes, or "state action." See Supreme Court of the United States, "Lindke v. Freed." While *Lindke* addresses this use in the First Amendment context and does not extend to electioneering, the principle of access to information presented by public figures such as candidates for political office demonstrates a public interest in ensuring independent researcher access to campaign-related communications made via personal social media accounts.

²³⁶ Federal Election Commission, "Making Electioneering Communications."

figures, including religious figures in contexts in which those leaders have power over decision-making in a country.

Historically, communications from candidates in the US have been considered public in that they are disseminated to the public, and in that they are subject to public reporting requirements.²³⁷ Some platforms already publicly label candidate and campaign accounts and others track additional candidates and groups internally, but platforms have not used labels consistently across countries. Beginning in 2018, for example, Twitter introduced labels in the US for candidates running for governor, the US Senate, or US House of Representatives.²³⁸ YouTube includes candidate information panels for US federal candidates.²³⁹ Platforms should work with electoral management bodies to confirm candidate lists and validate information, particularly during election periods.²⁴⁰

Given the potential volatility of political transition, political candidates' accounts (individual or campaign) should remain accessible for up to 120 days following an election, if the candidate was not elected. The time period of 120 days is necessary to capture important post-election candidate conduct. Meta, for instance, considers candidates to be public officials for 30 days after an election result, even if they have lost.²⁴¹ Following that point, winning candidates should be incorporated into the framework as public officials (above), and losing candidates' data should be removed from being considered as public platform data by virtue of their political entity status (but could still qualify through other means, such as by being a Notable Public Account or by distributing HDC).

3. Political parties and party officials

Given the major role that political parties play in governments and political processes of many nations, their accounts should be included in high-influence public platform data. A political party is a formal organization that organizes public opinion, communicates demands and priorities, and engages in political recruitment.²⁴²



Proactive Data Interface

For Major Political Parties: To facilitate provision of data, the framework focuses on major political parties: any party that has won at least 5% of seats (or in non-proportional representation states, has

²³⁷ Ibid.

²³⁸ Coyne, "Introducing U.S. Election Labels for Midterm Candidates."

²³⁹ YouTube, "Election information panels."

²⁴⁰ The Carter Center, "Election Management."

²⁴¹ Facebook, "Policy Forum."

²⁴² This definition of political parties also requires a certain constancy: "(1) continuity in organization—that is, an organization whose expected life span is not dependent on the life span of current leaders; (2) manifest and presumably permanent organization at the local level, with regularized communications and other relationships between local and national units; (3) self-conscious determination of leaders at both national and local levels to capture and to hold decision-making power alone or in coalition with others, not simply to influence the exercise of power; and (4) a concern on the part of the organization for seeking followers at the polls or in some manner striving for popular support." See LaPalombara and Weiner, "The Origin and Development of Political Parties."

won at least 5% of the vote at a national level) in the national legislative body in the most recent election.²⁴³ This is a more conservative definition than some researchers have used.²⁴⁴

Platforms should proactively identify accounts associated with the primary leaders of major political parties at the national level and include data from those organizations in the Proactive Data Interface.



Custom Data Requests



AND Independent Data Collection

For Minor Political Parties: In principle, all political parties – even minor ones – are crucial to understanding an electoral or governing environment. Furthermore by definition, political parties and officials engage publicly and seek to engage constituencies and the wider public. For that reason, researchers should be able to identify relevant accounts of all political parties and their officials at both the national and subnational level, and they should be able to pursue the data produced by those accounts through Custom Data Requests and Independent Data Collection mechanisms.

Accounts associated with the regional or local leaders of major political parties, as well as accounts associated with leaders of minor political parties should also be considered public and should be accessible through Custom Data Requests and Independent Data Collection mechanisms.

4. Government institutions and related officials, such as agencies, ministries, committees, and departments from all branches of government

Within any national government exists a series of institutions – across all branches of government – that are critical for performing core government functions. The online data associated with these institutional and official accounts are also important for understanding our online environment. These include agencies, national ministries, courts, legislatures (including caucuses, commissions, and coalitions), departments, independent regulatory bodies, commissions, and state media, where applicable, as well as their senior officials, when operating in their capacity as government workers.



Proactive Data Interface

For the National Level: Platforms should proactively identify official accounts associated with national government institutions and make these data available through the Platform Data Interface. In addition, the senior officials (secretaries, ministers, or other leaders) of these institutions serve a crucial role in speaking for their respective institutions, so their accounts – both official and personal if those accounts are public – should also be included.

²⁴³ Cottrell and Ghai, “The Role of Constitution-Building Processes in Democratization.”

²⁴⁴ Comparative research on political parties defines a threshold for inclusion in a variety of ways, including parties that won “3% or more in the national election immediately prior to the survey or that elected at least one representative to the national or European parliament” (Jolly et al., “Chapel Hill expert survey trend file, 1999–2019.”) or all parties “represented in parliament with at least one seat (in established democracies) or with two seats (in young democracies or countries with highly fragmented party systems)” (Merz et al., “The Manifesto Corpus.”).



Custom Data Requests AND



Independent Data Collection

For the National and Subnational Level: Government institutions at the subnational level also conduct critical work and share public communications. Research needs within particular information environments may necessitate granular compilation of elected officials and government positions, or the collection of lower-level institutions (state, district, city, etc.) and their leadership. Independent researchers should be able to identify relevant accounts at the subnational level and pursue the data they share through Custom Data Requests and Independent Data Collection mechanisms.

Guidance

A 2024 judgment of the US Supreme Court considered under what circumstances public officials engage in state action online, requiring officials to have both the authority to speak on behalf of the state and to actually do so.²⁴⁵ In other words, simply because content is generated by a governmental official does not mean it is public content. However, this decision in the US suggests that both official public accounts (e.g., the US Secretary of State X account) as well as public personal accounts of public officials (e.g., current US Secretary of State Marco Rubio) could both be considered under the definition of government account. Researchers should be able to access data from personal accounts when they are used for campaign or official purposes.

Some digital platforms and social media companies have established processes to define government officials (as well as notable public figures, discussed in the following section). Meta provides specific tools to those in politics²⁴⁶ and the MCL includes Facebook and Instagram content from “widely-known individuals and organizations” such as government officials.²⁴⁷

X defines government officials through its public interest exception, where the platform may choose to not moderate posts from elected or government officials that would otherwise violate the platform’s policies. X lists categories for current or potential members of local, state, national, or supra-national governmental or legislative bodies,²⁴⁸ but the platform does not clarify what sources are used to determine status.

Google’s Political Content Policy²⁴⁹ includes election information panels on YouTube, which provide information about elections, candidates, voting, election integrity, election results, and other election topics from “non-partisan, third-party sources.”²⁵⁰ In March 2025, Google said it “is working to bring

²⁴⁵ Specifically the Court concluded a public official (1) possessed actual authority to speak on the State’s behalf on a particular matter, and (2) purported to exercise that authority when speaking in the relevant social-media posts. Supreme Court of the United States, “Lindke v. Freed.”

²⁴⁶ Meta, “Public Figures.”

²⁴⁷ Meta, “Data dictionary”; Meta, “Meta Content Library and API.”

²⁴⁸ The account represents a current or potential member of a local, state, national, or supra-national governmental or legislative body: 1. Current holders of an elected or appointed leadership position in a governmental or legislative body, OR 2. Candidates or nominees for political office, OR 3. Registered political parties. See X, “About public-interest exceptions on X.”

²⁴⁹ YouTube, “Political content.”

²⁵⁰ YouTube, “Election information panels.”

information panels to more countries/regions.”²⁵¹

TikTok allows government officials to register their account and maintains extensive definitions of its Government, Politician, and Political Party Accounts, including national or federal government-run entities, state, provincial and local governments, and national governmental officials, among others.²⁵²

In addition to platform approaches, there are a range of external resources which define and categorize public officials. These resources exist for government officials and candidates in a number of countries,²⁵³ including Argentina,²⁵⁴ Gabon,²⁵⁵ Germany,²⁵⁶ Indonesia,²⁵⁷ Italy,²⁵⁸ South Korea,²⁵⁹ the United Arab Emirates,²⁶⁰ and the US,²⁶¹ among others. Public Electoral Management Bodies also sometimes manage public databases, for example in Brazil.²⁶² These databases should be used for reference and confirmation of lists provided by other sources, or as primary sources themselves.

²⁵¹ YouTube, “Topical context in information panel.”

²⁵² TikTok includes: National/Federal Government-run entities such as agencies, ministries, or offices; State/Provincial and local government entities; Candidates and elected officials at the federal/national level; Government officials at the federal/national level, such as cabinet ministers and ambassadors; Official spokesperson or member of senior staff for a national/state level candidate or elected/appointed official (e.g. Chief of Staff, Campaign Director, or Digital Director); Official spokesperson, member of senior staff, or executive leaders for a political party (e.g. Party Chairman or Finance Director); Political parties; Royal family members with official government capacities; Political youth associations (for main political parties at the discretion of regional public policy); Former heads of state and/or heads of government; Political Action Committees (PACs) or any country-specific equivalents; Candidates and elected officials at the state/provincial and local levels; Government officials at the state/provincial and local levels; among others. See TikTok, “Government, Politician, and Political Party Accounts.”

²⁵³ Poynter, “Empowering fact-checkers worldwide.”

²⁵⁴ Argentina, “Listado Alfabético.”

²⁵⁵ Gabon, “Le Guide De L’Agent Public.”

²⁵⁶ Schmidt et al., “Die Datenbank Öffentlicher Sprecher.”

²⁵⁷ Firdaus, “Indonesia Presidential Candidate’s Dataset, 2024.”

²⁵⁸ Italy, “Archivio degli Spot politici.”

²⁵⁹ South Korea, “검색결과.”

²⁶⁰ United Arab Emirates, “حسابات التواصل الاجتماعي لحكومة الإمارات.”

²⁶¹ Ballotpedia, “Ballotpedia.” CandiData, an independent nonprofit, assembled a database of social media accounts for U.S. federal incumbents, federal candidates, and a selection of state candidates. See CandiData, “CandiData24.”

²⁶² Brazil, “Lista Candidatos Eleições 2024.”

C. Notable Public Accounts

Notable Public Accounts belong to celebrities, journalists, civic leaders, or other public figures whose reach gives them outsized influence. Content from these accounts provides a valuable window into the dynamics of the information ecosystem. Platform data access mechanisms have traditionally included some types of these accounts, recognizing the outsized role these individuals and institutions play in shaping public discourse. The framework defines Notable Public Accounts using both absolute and relative thresholds for reach and engagement, enabling consistent and meaningful definition across information environments and through each access mechanism.

Definition

Content shared by a public account, channel, group, or page representing a notable public figure or entity, defined by:				
Reach		Proactive Data Interface	25,000 followers or subscribers; or	
		Custom Data Requests		Top 2% of public accounts, channels, groups, and pages by weekly reach.
Engagement		Custom Data Requests		Top 2% of public accounts, channels, groups, and pages by weekly engagement.
Public Interest Actors		Proactive Data Interface		Custom Data Requests
				Independent Data Collection
	All public interest actors, including broadcast or news accounts, nonprofits, academic institutions, among others.			

Implementation Discussion

This section explains the rationale for absolute and relative thresholds for Notable Public Accounts. The use of absolute and relative thresholds enables policymakers and platforms to establish objective standards that can scale across platforms and different information environments. In calculating the absolute and relative thresholds for accounts, platforms and researchers should consider views, listens, streams, downloads, as well as any other platform-relevant metrics.

Content from Notable Public Accounts should be accessible through multiple access mechanisms to enable research in different information environments.



Proactive Data Interface

Reach – Public accounts, groups, and pages that have greater than 25,000 followers or subscribers: Public accounts, groups, and pages with more than 25,000 followers or subscribers constitute a significant proportion of total reach on a platform and should be accessible via a Proactive Data Interface. This threshold aligns broadly with several existing platform data access mechanism definitions (see Appendix C).

Defining accounts, groups, and pages with greater than 25,000 followers or subscribers as high-influence public platform data strikes an effective balance between public interest research and user expectations of privacy. Previous research focused on Notable Public Accounts has used thresholds as high as 100,000 followers,²⁶³ or as low as 1,000.²⁶⁴ Users operating accounts, channels, groups, or pages that are set to public and have more than 25,000 followers or subscribers would likely have reason to expect that their content is publicly available.

There are risks that follower or subscriber counts could be manipulated by bots or coordinated spamming. Platforms have existing tools to detect inauthentic accounts and should endeavor to maintain accurate follower or subscriber totals.



Custom Data Requests



AND Independent Data Collection

Reach – Top 2% of public accounts, channels, groups, and pages by weekly reach: Platforms should enable researchers to request or independently access user accounts, channels, groups, or pages that are in the top 2% of weekly platform reach in particular information environments through Custom Data Requests and Independent Data Collection.

The top 2% of accounts will likely represent a significant proportion of total platform distribution. In Meta's Q4 2024 Widely Viewed Content Report, for example, just the top 20 Facebook Pages collectively accounted for about 1% of *all* U.S. content views.²⁶⁵ While Facebook does not systematically publish the total number of pages on its platform, Facebook had 60 million business pages in 2016, suggesting the top 2% of reach represents only a tiny fraction of total pages.²⁶⁶ Consistent with this, a 2018 study of YouTube found that 85% of total views were of just 3% of all channels.²⁶⁷ These findings were broadly confirmed by a 2024 study of YouTube which found 3% of channels captured nearly 73% of total views.²⁶⁸

²⁶³ Starbird et al., "Influence and Improvisation."

²⁶⁴ Lutscher and Ketchley, "Online repression and tactical evasion"; Meta, "Data dictionary"; Meta, "Meta Content Library and API"; Tan et al., "The effect of wording on message propagation."

²⁶⁵ Meta, "Widely Viewed Content Report."

²⁶⁶ D'Onfro, "This group of 60 million Facebook users could be the secret to Facebook's next growth spurt."

²⁶⁷ Bärthel, "YouTube channels, uploads and views."

²⁶⁸ Munger et al., "Pressing Play on Politics."

Engagement – Top 2% of public accounts, channels, groups, and pages by weekly engagement:

Platforms should enable researchers to request or independently access user accounts, channels, groups, or pages that are in the top 2% of weekly platform engagement in particular information environments through Custom Data Requests and Independent Data Collection. Like reach, the distribution of engagement on platforms is highly concentrated, and the top 2% of engagement with public accounts, channels, groups, and pages will represent a narrow slice of overall accounts but a significant proportion of total engagement.

A 2025 study that relied on CrowdTangle and studied Facebook pages which referenced top U.S. politicians found that the top 6% of pages accounted for more than 91% of all engagement.²⁶⁹ Users and entities controlling accounts, channels, groups, and pages in the top 2% of aggregated weekly engagement would reasonably expect their data to be public.

**All Access Mechanisms**

Public Interest Actors (Including News Organizations and Journalists): User accounts, groups, or pages belonging to public broadcasters, news organizations, journalists, and other public interest actors should also be accessible via Proactive Data Interface, Custom Data Requests and Independent Data Collection mechanisms. This includes both registered or verified accounts as well as accounts from public interest actors who are not verified.

Platforms offer verification of accounts, including tools for proactive verification by media as well as platform-based identification of state-run media accounts. Facebook, for example, allows users to register as journalists in a range of countries, which makes accounts eligible for “stronger security features” and such accounts “may be eligible for other benefits, such as blue badge verification.”²⁷⁰ Bluesky also offers verification of journalists.²⁷¹

The rise of digital platforms has occurred alongside significant evolution of the media landscape. A range of individuals and institutions now play important roles in online information environments, and verification or employment by a major media company is by no means the only indicator of public interest actors.

In the Proactive Data Interface, platforms should incorporate accounts, groups, or pages from public interest actors when verified, as well as those significant accounts that are not. Researchers should also be able to use Custom Data Requests and Independent Data Collection to analyze data from identified public broadcasters, news organizations, journalists, and other public interest actors necessary for research in specific information environments.

²⁶⁹ Yang et al., “Coordinated link sharing on Facebook.”

²⁷⁰ Meta, “About journalist registration.”

²⁷¹ Bluesky, “Bluesky for Journalists.”

Guidance

Platforms have operationalized a range of programs relevant to the definition of Notable Public Accounts in public platform data access programs as summarized in Appendix C.

Policy proposals have also sought to establish defined Notable Public Accounts for inclusion in independent research. PATA would request the FTC to define “major public accounts” which “at a minimum ... are restricted to reasonably public accounts whose content is followed by at least 25,000 users or otherwise regularly reaches at least 25,000 users per month.”²⁷² The proposal for DSOSA similarly would require the FTC to clarify how platforms should provide certified researchers access to public content that includes “Public high-profile accounts involved in the spread of such pieces of content.”²⁷³ The DSA and delegated regulation does not define Notable Public Accounts.²⁷⁴

Summary of Notable Public Accounts Thresholds

	Access Mechanism		
	 Proactive Data Interface	 Custom Data Requests	 Independent Data Collection
Reach	>25,000 followers or subscribers	Top 2% by weekly reach	
Engagement		Top 2% by weekly engagement	
Public Interest Actors	All public interest actors included		

D. Business Accounts and Promoted Content

Platforms play a central role in advertising products and facilitating commercial transactions with the public. For this reason, the category of Business Accounts and Promoted Content – accounts that use advertising and commercial messaging, which can sway consumer behavior, public health, or public trust – is a key type of high-influence public platform data. Business Accounts are defined as accounts that are monetized and/or established for a primarily commercial purpose, and promoted content is defined as content that is paid to be made visible to a larger audience.²⁷⁵

²⁷² United States, “Platform Accountability and Transparency Act,” §(9)(i)(6). “REASONABLY PUBLIC.—The term “reasonably public” means information that the author made available in a manner and under such circumstances such that the author does not retain a reasonable expectation of privacy in the information. The fact that a user may need to register or create an account with a platform to view information does not preclude it [from] being deemed reasonably public.”

²⁷³ United States, “Digital Services Oversight and Safety Act.”

²⁷⁴ European Commission, “Delegated Regulation on data access provided for in the Digital Services Act.”

²⁷⁵ Business user is defined in the Digital Markets Act as “any natural or legal person acting in a commercial or professional capacity using core platform services for the purpose of or in the course of providing goods or services to end users.” See Waem, “Who’s who under the DMA, DSA, DGA and Data Act?”

While this definition incorporates a broader range of accounts than the previous three categories of data, these users have demonstrated a desire to reach a broader audience. Analysis of these sources of data, like all other categories, should be grounded in relevant ethical and privacy standards.

Definition

All data associated with an account belonging to a business enterprise, including:		
 Proactive Data Interface	 Custom Data Requests	 Independent Data Collection
Registered or verified business accounts		
Accounts that are part of a platform’s monetization program		
Accounts that pay to promote their account as a whole		
Any content that any account has paid to be promoted		
User-generated advertisements (e.g., externally sponsored content with #ad)		

Implementation Discussion

Business Accounts are an important part of the online information ecosystem, and fundamental to how the public makes sense of content dissemination.²⁷⁶ Indeed, many of the most widely used digital platforms are supported by targeted advertising and other types of commercial promotion.²⁷⁷ The framework’s focus on Business Accounts and Promoted Content is in complement to existing tools that enable access to advertising repositories, through both searchable interfaces as well as platform API access.²⁷⁸

Business Accounts are central for understanding the nature of what commercial activity transpires on such platforms. Because these accounts seek to engage public audiences through platforms, their data is reasonably understood to be publicly available, subject to certain limitations.

²⁷⁶ Silverman, “How to tell if a social media account is monetized.”
²⁷⁷ See Acemoglu, “Online Business Models, Digital Ads, and User Welfare.” Two cited examples include: digital ads made up 98% of Facebook’s revenue from 2017-2019. Trefis Team, “What is Facebook’s Revenue Breakdown?” And they were about 85% of YouTube’s revenue in 2020 despite its premium ad free subscription plan. Chapkanovska, “YouTube Revenue Statistics of 2024.”
²⁷⁸ See, e.g., European Commission, “Status Report.”

In practice, platforms also increasingly treat Business Account content as public by default. For example, Meta now indexes videos and photos shared by Instagram business users in external search engines by default.²⁷⁹ This shift reinforces the understanding that data from Business Accounts should be broadly discoverable, is meant for public consumption, and therefore aligns with the rationale for research accessibility.

Because Business Accounts are so clearly public, they should be accessible by all three access mechanisms. Because platforms internally categorize Business Accounts and Promoted Content, data should be provided through the Proactive Data Interface. In addition, specific data or use cases in different information environments may also require access through Custom Data Requests or Independent Data Collection.

The framework considers this concept at two levels – the content level and the account level, as detailed below.

1. Accounts

There are various ways that an account may clearly identify as being a Business Account, referred to here as a “business user.” A business user is defined in the Digital Markets Act as “any natural or legal person acting in a commercial or professional capacity using core platform services for the purpose of or in the course of providing goods or services to end users.”²⁸⁰ Three ways that business users are identified are outlined below.



All Access Mechanisms

Registered or verified Business Accounts: Many platforms offer opportunities for accounts to register or be verified as Business Accounts. Different platforms have different means by which they implement this verification. X offers users the opportunity to establish a professional account. This account option is expressly targeted at businesses, brands, creators, and publishers to have “a unique and clearly defined presence on the platform” and increase connections, promote posts, and monetize their account.²⁸¹ Business Accounts on Pinterest serve a similar function, allowing business users to analyze their account metrics and promote specific pins.²⁸² Spotify similarly allows users to establish an Artist profile in order to promote music and pitch to playlist editors.²⁸³ On some platforms, commercial nature is also self-identified at the time of account creation. Facebook, for example, allows users to register an online business profile which Facebook states will help the account “get discovered by fans and customers.”²⁸⁴

²⁷⁹ Meta, “Why search engines might index public Instagram photos and videos.”

²⁸⁰ Waem, “Who’s who under the DMA, DSA, DGA and Data Act?”

²⁸¹ X, “Professional Accounts.”

²⁸² Pinterest, “Get a business account.”

²⁸³ Spotify, “Spotify for Artists.”

²⁸⁴ Meta, “Facebook Page.”

Registering for these account statuses clearly indicates a desire to reach a public audience. When platforms formally register these profiles in their signup and operations systems, data should be included in a Proactive Data Interface. Where the registered or verified Business Accounts use the platform to sell goods or services, the platform should include relevant metadata such as data related to prices, sales, or rentals, among other relevant platform categories.

Registering through a platform monetization program: Use of platforms for brand awareness and monetization is an important signal of a user’s intent for public availability. A 2025 estimate suggests that platforms distribute more than \$20 billion USD annually through revenue distribution programs (distinct from branded content, tips, or subscriptions).²⁸⁵ Platforms offer a range of programs that typically include eligibility thresholds for reach and engagement, including the YouTube Partner Program,²⁸⁶ Meta’s Partner Monetization Policies,²⁸⁷ and Spotify for Creators,²⁸⁸ among others. Participating in such programs indicates a desire to reach more people – to engage publicly. When content creator accounts reach platform-established thresholds, the data from these accounts should be incorporated through a Proactive Data Interface.

Account holders that pay for promotion: Some platforms also have mechanisms by which account owners can pay to promote their entire account in some way, although this is less common than mechanisms to promote specific content. For example, Facebook offers the option to pay to promote an entire page,²⁸⁹ rather than a single piece of content. Paid promotion is an important indicator of account publicness – accounts are paying money to make the content *more* visible, or more public.

2. Content

While the previous categories focused on the account level, monetization or promotion can also happen at the content level. Even if an individual account might not fit the framework’s definition of high-influence public platform data, any content that is paid to be promoted should be considered high-influence public platform data.



All Access Mechanisms

Paying for promotion or increased visibility of a single piece of content: Many platforms have mechanisms by which an account owner can pay to promote a specific piece of content in some way. For instance, an account owner can pay to promote content on X,²⁹⁰ or to “boost” a post on a Facebook page.²⁹¹ TikTok also allows for promoting content, and makes it explicit that “You can only

²⁸⁵ WhatToFlx, “Social Media Monetization 2025.”

²⁸⁶ YouTube, “YouTube Partner Program overview & eligibility.”

²⁸⁷ Meta, “Partner Monetization Policies.”

²⁸⁸ Spotify, “Spotify for Creators’ monetization options.”

²⁸⁹ Meta, “About Page likes ads.”

²⁹⁰ X, “Promoted Tweets.”

²⁹¹ Meta, “Boost a post from your Facebook page.”

promote content that is public on TikTok,”²⁹² again underlying the explicit tie between advertising or promotion and content’s publicness.

Paid promotion is an important indicator that the content is commercial in nature, and indicates an expressed desire to expand the reach of that content with the public. Not all accounts that pay to promote individual pieces of content would assume that all of their data would be publicly available. Accordingly, the framework strikes a balance whereby *promoted content* is defined as high-influence public platform data whereas all of the data of the underlying account is not.

User-generated advertisements and externally boosted posts: Promotion also occurs through user-generated content that is externally sponsored, often marked with hashtags such as #ad, #sponsored, #paid, or #partner, or other disclosures.²⁹³ This includes posts by micro-influencers – typically accounts with fewer than 25,000 followers – who receive compensation from third parties to promote products or services.²⁹⁴ Although these posts are not promoted through the platform’s native paid promotion tools, they serve a clear commercial function and are directed at the public. From a technical perspective, platforms can identify these posts via standardized advertising disclosures, as required by national legislation (e.g., Germany’s influencer advertising laws²⁹⁵). The inclusion of such data aligns with the principle of treating publicly targeted, commercial messaging as data that should be accessible for accountability and research purposes.

Guidance

Commercial data is already included in public platform data access mechanisms offered by platforms. The MCL, for example, enables researchers to access:

- Public listings on Facebook Marketplace from Pages or profiles;
- Instagram posts from public business and creator accounts;
- Whether Facebook Groups are owned by registered professional accounts;
- Whether Facebook Events are hosted by registered professional accounts;
- Among other fields.

Existing laws and policy proposals have also sought to include data from Business Accounts and Promoted Content. As originally drafted, PATA included a section about advertising data requirements, acknowledging the importance of transparency in this area.²⁹⁶ The DSA includes requirements for transparency in paid advertising.²⁹⁷ The Brazilian version of the Digital Markets Act defines “professional users” as those who use the platforms to offer goods or services to final users, and can be considered differently.²⁹⁸

²⁹² TikTok, “Use Promote to grow your TikTok audience.”

²⁹³ Such disclosures are consistent with FTC guidelines (Federal Trade Commission, “Disclosures 101 for Social Media Influencers”) related to ad disclosures that are “clear and conspicuous” (Federal Trade Commission, “.com Disclosures.”).

²⁹⁴ Ershov, “Market incentives and advertising disclosure regulations.”

²⁹⁵ Media Authority of North Rhine-Westphalia, “Disclosing Sponsored Influencer Content”; Radtke, “Disclosure Requirements for Influencer Marketing in the U.S. and Germany.”

²⁹⁶ United States, “Platform Accountability and Transparency Act.”

²⁹⁷ European Union, “Digital Services Act”, Article 39.

²⁹⁸ Brazil, “Projeto de Lei N, DE 2022.”

VII. Baseline Data Access Fields

Establishing standardized data expectations across public platforms is critical for ensuring consistency in data access, enabling meaningful cross-platform analysis, and promoting a clear and more coherent understanding of the online information ecosystem.

This section establishes baseline expectations for the fields of high-influence public platform data that should be made available through platform data access mechanisms. Specifically, this section outlines data categories and data access fields that should be accessible via the appropriate access mechanism for each data category in the framework, and across different information environments. This includes a focus on historic data as well as secondary-level data, such as comments..

In addition to outlining key data access fields, this section addresses issues related to special data types, including ephemeral, deleted, and spotlighted content, which require additional care in implementation and research use.

A. Overview of Data Types and Data Access Fields

Advancing toward standardized data expectations across platforms is essential for enabling meaningful cross-platform comparison and ensuring research consistency. Platforms should make the same categories of high-influence public platform data available through the Proactive Data Interface. Researchers should be able to request these and other fields through the other access mechanisms.

At the same time, given the diversity of platform architecture and affordances, many fields will necessarily vary across platforms. Furthermore, the majority of research conducted on platforms has traditionally focused on social media platforms. As such, the below categories are most suitable for those platforms, but the framework expects these categories to be strengthened and applied across other types of platforms.

Data access mechanisms should include the following categories and data access fields for high-influence public platform data, where relevant. These fields were consolidated from existing and historical data access mechanisms, including those summarized in Appendix D. Other fields of relevant high-influence public platform data may be requested by researchers or accessed independently by researchers.

Type	Data Access Fields	
Account / Community	• Name and/or username • Account name change history • Profile picture	• List of followers, anonymized unless followers meet the framework’s criteria for public platform data • Account creation date

	• Poster bio, about, or description, including any links • Follower count	• Account location • Account default language
Content, including text, image, video, audio, etc.	• Content owner, including type • Content location surface • Language • URL of content • Type of content, including where relevant group, channel, or page location	• Content link, if applicable • Attachment link, if applicable • Date and time of creation • Date and time of any modification • Time series data, including for engagement
Sub-content / Affiliated Content*	• Anonymized commenter ID • Reshared or reposted content • Comment type • Language • Comment owner, if owner meets criteria for high-influence public platform data; if not owner anonymized • Content url	• Content link, if applicable • Comment content (text, image, etc) • Attachment link, if applicable • Date and time of creation • Date and time of any modification • Comment likes (number) • Comment shares (number) • Comment replies (number)
Metrics	• Number of views and view count date • Number of reactions (e.g., likes, favorites, upvotes, or other reactions as afforded by the platform)	• Number of downloads • Number of comments • Number of shares • Number of other reactions as afforded by the platform
Labels and Platform	• Verification status(es) • Profile categories (user self-identified) • Monetization, advertising, or sponsorship	• Public labels applied by the platform, users, or third parties e.g., community notes or AI-generated content
Removed Content	• User removed content, prior to removal • Platform removed content, prior to removal	• Content that is illegal in a particular jurisdiction should not be included in publicly available access mechanisms in that

	jurisdiction
Other	• Other fields as identified

Historic Data

Where an account crosses a definitional threshold to become high-influence, data associated with the account’s previous conduct is not included in the definition of high-influence public platform data. In other words, if an account or channel becomes a Notable Public Account by virtue of crossing into the top 2% of accounts, channels, groups, and pages, access mechanisms should only capture data from that moment forward.

Secondary-Level Data

Comments and other secondary-level interactions and data are also critical to include, given their centrality in online conversations.²⁹⁹ However, these secondary-level data may also raise heightened privacy concerns given user expectations of privacy.³⁰⁰

Empirical research shows that many users expect anonymity in research use of their data.³⁰¹ The poster of secondary-level data should not be identified in data access mechanisms and/or research outputs. Instead, an anonymized commenter ID specific to the poster of the comment should be used when possible. This allows comments from the same poster to be linked to one another, similar to the MCL’s approach in 2025.³⁰²

However, in some, if not many, cases, commenters will be identifiable simply through the platform interface or based on the style of their writing.³⁰³ This will particularly be true for the most prolific public commenters who write under their real names on multiple platforms. Nonetheless, anonymized comments on high-influence public platform data should be accessible. The importance of these data in the information ecosystem outweighs these potential privacy risks associated with access. Researchers should treat these data ethically and responsibly when considering potential publication.

Where a commenter or comment meets any of the criteria outlined in the four categories of the framework, their de-anonymized comment would be available under the framework’s access mechanisms. In other words, any comments that reach the thresholds defined by the framework for HDC, or that are posted by Government and Political Accounts, Notable Public Accounts, or Business Accounts, would be included as high-influence public platform data under those categories.

²⁹⁹ American Sunlight Project, “EcoBoost.”
³⁰⁰ Brown et al., “Web Scraping for Research.”
³⁰¹ Williams et al. “Users’ Views of Ethics in Social Media Research.”
³⁰² Meta, “Data Dictionary.”
³⁰³ Neumann et al., “An Examination of the Impact of Stylometry, Artificial Intelligence/Machine Learning (AI/ML) on Privacy in Social Media.”

B. Special Data Types

Several specific types of high-influence public platform data warrant additional discussion. This section briefly discusses guidelines around three such categories: ephemeral content, deleted content, and spotlighted bystander content. In addition to the recommendations for each type of content below, researchers should be particularly cautious in accessing, using, and publishing these types of high-influence public platform data given their inherent sensitivity.

Ephemeral Content

Ephemeral content is defined as “communication artifacts, including text, pictures, and videos, that [are designed to] be erased after being display[ed] for a limited period of time.”³⁰⁴ Ephemeral content is not inherently public or private. It may be public in some circumstances (e.g., when shared by a public official speaking in their official capacity) and private in others (e.g., a user without significant reach posting with the intent of reaching only their family for a limited period of time).

Despite its transitory nature, ephemeral content increasingly plays a significant role in shaping online discourse. Some 500 million users interact with the ephemeral Instagram Stories on a daily basis, contributing to the broader information ecosystem online, and the inherent risks therein.³⁰⁵ Social media platforms have intentionally expanded the use of ephemeral surfaces, such as Stories or time-limited posts, with leaked internal documents suggesting that these features are designed to encourage more frequent engagement by fostering a sense of urgency.³⁰⁶ And research suggests that users may actually devote more time and attention to ephemeral content, knowing that it will be unavailable in the future.³⁰⁷ As a result, ephemeral content has become a key component of platform design and user behavior. In determining whether ephemeral content can be considered high-influence public platform data, the central question to be addressed is not whether content is eventually deleted, but whether it is intended or understood to be *public* during the time it is visible.

There is ample evidence that ephemeral content is often treated as public by both its creators and consumers on some platforms (e.g., Instagram), but not always others (e.g., Snap). Ephemeral content across the four categories of high-influence public platform data is regularly used for research. Journalists regularly cite and incorporate ephemeral posts made by politicians and other public figures in their reporting, preserving screenshots as part of public discourse.³⁰⁸ Business Accounts frequently use ephemeral formats to share time-limited promotions, reflecting a clear intent to reach a broad audience.³⁰⁹ From a consumer protection standpoint, even the timing and existence of such promotional content is of public interest, because of laws concerning advertising discounts.³¹⁰

Highly disseminated ephemeral content – defined under this framework as content that has greater than 10,000 instances of total distribution or falls within the top 2% of platform distribution or

³⁰⁴ Chen and Cheung, “Unlocking the power of ephemeral content.” See also Bayer et al., “Sharing the small moments.”

³⁰⁵ Cozma, “91 Instagram Statistics to Track NOW.”

³⁰⁶ See, e.g., Kentucky, “Complaint”; Massachusetts, “Complaint and Jury Demand.”

³⁰⁷ Barnea et al., “The Effects of Content Ephemerality on Information Processing.”

³⁰⁸ Švelch, “Redefining screenshots.”

³⁰⁹ Barnea et al., “The Effects of Content Ephemerality on Information Processing.”

³¹⁰ United States, “16 CFR § 233.3.”

engagement – similarly appears to be considered public by internet users. Viral ephemeral content is regularly reported on,³¹¹ and while much of this content is entertainment-focused, ephemeral content is sometimes highly relevant to news events.³¹² That such content can achieve this level of reach, even within its limited lifespan, further underscores its public relevance and impact.

Ephemeral content can be considered high-influence public platform data when and if it meets the same criteria established for other content. The fleeting nature of this content does not appear to reduce its societal importance, and thus it should not be shielded from study on this basis.

Accordingly, content that is shared by Government and Political Accounts, Notable Public Accounts, or Business Accounts, or that qualifies as HDC or Promoted Content, should be treated as public *while it exists*, even if it is ephemeral. However, the framework does not expect that platforms should retain or archive ephemeral content beyond its intended lifespan, and encourages researchers to be judicious in what ephemeral content they republish.

Deleted Content

Similarly, deleted high-influence public platform data presents particular risks and challenges. Content is frequently deleted from platforms, with research suggesting a range of 10%³¹³ to 24%³¹⁴ of content on Twitter and perhaps as much as 50% on Instagram.³¹⁵

Data may be deleted from a platform in at least three ways. First, it may be removed by the platform due to moderation or policy enforcement, often after evaluation that the data is misaligned with community guidelines, terms of service, or legal requirements. Second, it may be deleted by users, who choose to delete either a specific post or an entire account on a platform. Finally, data may not be deleted, per se, but may be withdrawn from public view, by switching account settings to private, or archiving posts.

The first category of deleted content, when the deletion happens from the platform, generally occurs because the content or the account's behavior violates the terms of service of the platform, or its behavior guidelines. This can include deleting specific content, or deleting an account, sometimes referred to as deplatforming.³¹⁶ Platforms create their own rules (all have terms of service, YouTube³¹⁷ and TikTok³¹⁸ have community guidelines, Meta has community standards,³¹⁹ X has rules³²⁰), but they tend to have similar reasons for removing content – if it is illegal, if it is violent or threatening, or if it contains content that the platform considers harmful, then it is likely to violate platform rules. This type of deleted content, provided it fits the inclusion criteria already defined by the framework, is

³¹¹ Carroll, "Are they 'having an affair' or just shy? The couple caught on Coldplay's kiss cam."

³¹² Mathur, "Idaho shooter's chilling Instagram story goes viral; Here's what he posted hours before he ambushed firefighters."

³¹³ Potash et al., "Using Topic Modeling and Text Embeddings to Predict Deleted Tweets."

³¹⁴ Ferreyra et al., "Regret, Delete, (Do Not) Repeat."

³¹⁵ Tinati et al., "InstaCan," 267-271.

³¹⁶ Ali et al., "Understanding the Effect of Deplatforming on Social Networks."

³¹⁷ YouTube, "YouTube's Community Guidelines."

³¹⁸ TikTok, "Community Guidelines."

³¹⁹ Meta, "Community Standards."

³²⁰ X, "The X Rules."

particularly important for researchers to access, as it allows them greater information about what content is affected by platform rules.³²¹ As such, it should therefore generally be treated as public within the framework. There are, however, legal requirements which will govern the treatment of specific categories of data (e.g., Child Sexual Abuse Material) as well as user privacy. Researchers must comply with relevant laws.

The second category of deleted content occurs at the user level, when a user deletes either a post or their account on the platform. There is reason to believe that users do not want platforms to have access to data from deleted accounts.³²² There are also circumstances in which users may choose to delete data or their full account for their own protection,³²³ suggesting that real harm may occur if such data were to be accessed by malicious actors.³²⁴ Research suggests that on at least one platform (Twitter), the majority of deleted content may fall into this second category.³²⁵ Again, there are legal requirements – such as the GDPR³²⁶ – which will impact the feasibility of research with deleted data. Researchers should anticipate and align with these expectations.

Finally, users may choose to make their data less visible, which would have the effect of making it not public by the framework's definition of a reasonable understanding of publicness. This could happen if a user chooses to archive content, or to change their privacy settings to make it less public.³²⁷ For example, moving a post from publicly visible (20% of Facebook posts analyzed in one study) to visible only to friends (closer to 70% of Facebook posts in the same study) would be included in this third category.³²⁸ Such changes, while not technically deleted, would have the effect of making data not public by this framework's definition.

For all three categories of deleted content, if data had already been accessed (through any of the three data access mechanisms) prior to the point of deletion, the researcher would retain access to that data. However, data that was deleted or no longer visible on the platform would therefore no longer be public, and would then cease to be available through ongoing data provision from platforms. As in the case of ephemeral content, the framework does not require that platforms retain or archive deleted content. Furthermore, researchers should proceed with caution when analyzing and, especially, publishing any deleted content given ethical and legal requirements relevant to these types of data.

Spotlighting

Content flows across platforms in complicated ways. One form of this is so-called spotlighting: when content that is posted by a user (that is not subject to any of the criteria of the framework) is then reposted by another user who does meet one or more of the criteria articulated in the framework. A

³²¹ Chancellor et al., “This Post Will Just Get Taken Down’.”

³²² Schaffner, “Understanding Account Deletion and Relevant Dark Patterns.”

³²³ Petrovic, “I Wish I Didn't Say That!”

³²⁴ Wang et al., “I regretted the minute I pressed share’.”

³²⁵ This study of Twitter suggested 85% of deleted tweets were deleted by the user (as opposed to a deleted or protected account). Petrovic, “I Wish I Didn't Say That!”

³²⁶ Working Group on Platform-to-Researcher Data Access, “Report of the European Digital Media Observatory’s Working Group on Platform-to-Researcher Data Access.”

³²⁷ Mondal et al., “Moving Beyond Set-It-And-Forget-It Privacy Settings on Social Media.”

³²⁸ Ibid.

classic example of this is when prominent political commentary accounts repost content from non-notable cross-partisans (including from minority groups) to portray them as unreasonable (sometimes maliciously), completely changing the context and audience for the content.

In most cases, the originally posted content would have to be publicly shared in order to be identified by the spotlihter. So the original poster could, in some instances, have a reasonable expectation that their content is public as opposed to private.

While the expectations of that user with regard to the audience for their content may be violated, if the spotlighted content fits within the framework's criteria – that is, if it is shared by Government and Political Accounts, Notable Public Accounts, or Business Accounts, or that qualifies as HDC or Promoted Content – then researchers may treat the content as high-influence public platform data. However, in this instance, underlying account information should not be treated as high-influence public platform data unless the account also meets the framework's criteria for public platform data. As in the case of ephemeral content and deleted content, researchers should proceed with caution when analyzing and, especially, publishing content that includes spotlighting.

Spotlighted users also have important options to change the visibility of their content, which could remove it from being included as public platform data once that change is made. Some research suggests, however, that spotlighted users tend to post *more* after being spotlighted,³²⁹ but that tendency is almost certainly dependent on the context of the spotlighting. Researchers would not be obligated to delete data that had already been accessed prior to its visibility being changed.

VIII. The Way Forward

The stakes could not be clearer: when independent access to high-influence public platform data disappears, so does society's ability to understand and protect itself.

Independent research with public platform data is increasingly constrained. Once supportive of independent research, major platforms have reversed course: Meta closed CrowdTangle, X priced most researchers out of its API, and Reddit imposed new limits that undermine research. At the same time, AI firms and data brokers capture vast quantities of public platform data at scale, deepening inequities between commercial and research use cases.

Emerging regulatory proposals in the US and frameworks in the EU and elsewhere aim to improve access to public platform data. But operationalization and enforcement remains uneven. As a result, researchers must rely on costly third-party tools or navigate legal and technical uncertainties associated with automated collection. These trends combine to create a fragmented and inequitable ecosystem of study of online platforms.

³²⁹ Schafer and Starbird, "Post-Spotlight Posts."

The *Better Access* framework offers a practical foundation: begin with the four categories of data that matter most for the public, ensure consistent and ethical access through multiple access mechanisms, and give researchers and the public the tools they need to understand the dynamics shaping public discourse in their own information environment.

High-influence public platform data is the minimum, most essential tier of data for accountability. This framework underscores that transparency is not a regulatory burden, but a democratic necessity. Adopting a uniform, cross-industry baseline would give the public the visibility it needs to see the digital public square clearly and to shape its future.

Appendix A. The Policy Landscape for Public Platform Data

Thanks to transparency advocates across diverse fields, various digital platform transparency regimes have emerged, including both voluntary and regulatory. These regimes expect or require platforms and online service providers to share information about their activities, algorithms, and processes with the public as well as vetted entities like researchers, regulators, and, sometimes, business competitors. Terms of service for platforms often govern whether and how independent research can be carried out.³³⁰ However, voluntary initiatives and regulation shape these requirements.

This section describes two interrelated strands of policy that impact public platform data: voluntary commitments and regulatory requirements. Privacy laws, including data protection regulation, and ethical research frameworks are also central factors in determining access to public platform data.

A. Voluntary Commitments

Platforms have a long, if uneven, track record of engaging with data access. Many large digital platforms have conducted transparency reporting on content moderation and, in some cases, widely distributed content.³³¹ However, the scope, frequency, and effectiveness of transparency reporting varies widely. Access Now's Transparency Reporting Index, launched in 2014, tracks how platforms report on how they respond to government requests and enforce their own policies.³³² The 2024 Index, which documents transparency reporting from 84 platforms, including social media companies, telecommunications companies, and internet infrastructure companies, notes that some companies have paused reporting and taken down previously available data.³³³

Third-party and collaborative initiatives, including the Santa Clara Principles³³⁴ and the EU's Code of Conduct on Disinformation,³³⁵ have also played roles in shaping norms around transparency and data access. These efforts have focused on providing transparency on content moderation decisions, the use of algorithms in recommender systems, and proactive detection of policy violating content, among others. Given the non-binding nature of these regimes, researchers contend that companies often provide incomplete data,³³⁶ and companies may and do retreat from commitments with little consequence.³³⁷

³³⁰ Clark, "Research Cannot Be the Justification for Compromising People's Privacy."

³³¹ Trust & Safety Professional Association, "History of Transparency Reports."

³³² Access Now, "Transparency Reporting Index."

³³³ Ibid.

³³⁴ Access Now et al., "Santa Clara Principles on Transparency and Accountability in Content Moderation."

³³⁵ European Commission, "The Code of Conduct on Disinformation."

³³⁶ Albert, "Platforms' Promises to Researchers"; Mündges and Park, "But Did They Really?"

³³⁷ Democracy Reporting International, "Big tech is backing out of commitments countering disinformation."

Transparency has also featured prominently in emerging voluntary frameworks around AI. The OECD AI Principles emphasize transparency and responsible disclosure, enabling users to recognize when they are interacting with AI and to question or challenge its outcomes.³³⁸ The 2023 US Executive Order on Safe, Secure, and Trustworthy AI directed federal agencies to adopt transparency measures in their AI use, including publishing reports and conducting structured risk evaluations.³³⁹ The 2025 America's AI Action Plan from the White House similarly includes references to transparency.³⁴⁰

Despite gains, these transparency reporting and voluntary initiatives have been criticized by researchers and regulators as providing incomplete information and an inconsistent picture.³⁴¹ As a result, increasing jurisdictions are taking steps to mandate data access through regulation.

B. Regulatory Requirements

Regulators have begun to establish requirements for platform transparency, including access to non-public data and public platform data. Broadly, regulations and proposed regulations focus on enabling two types of data access: privileged access to platform data and access to public platform data.³⁴² Regimes like the DSA that seek to enable analysis of private platform data require researchers to comply with a number of provisions, including affiliation to a research organization and the publication of the findings. Regulations requiring access to public platform data generally have fewer requirements for researchers, but also may mandate safeguards such as independence from commercial interests and proportionate data protection and confidentiality. Regulation may also establish requirements for the research purpose, namely that research is in the public interest or related to specific areas of inquiry (e.g. protection of monies). This section summarizes regulatory requirements and proposals, highlighting connections to high-influence public platform data.

Regulation and Regulatory Proposals in the US

A wide range of legal frameworks in the US impact access to public platform data, including the Computer Fraud and Abuse Act (CFAA),³⁴³ Children's Online Privacy Protection Rule (COPPA),³⁴⁴ among others.

There have been multiple efforts to mandate platform transparency at both the federal and state levels. At the federal level, PATA was first introduced in 2022 and has been reintroduced in subsequent congressional sessions.³⁴⁵ Subject to ongoing debate,³⁴⁶ PATA would require platforms with at least 50 million unique monthly users to make publicly accessible data available through a dedicated

³³⁸ Organisation for Economic Co-operation and Development, "AI principles."

³³⁹ United States Executive Office of the President, "Executive Order 14110."

³⁴⁰ The White House, "America's AI Action Plan."

³⁴¹ 55% of required qualitative responses were incomplete or missing, and 64% of quantitative data points were either absent or not robust. See Mozilla, "Full Disclosure: Stress testing tech platforms' ad repositories"; Mündges and Park, "But Did They Really?"

³⁴² Husovec, *Principles of the Digital Services Act*.

³⁴³ United States, "Computer Fraud and Abuse Act," §1030.

³⁴⁴ Federal Trade Commission, "Children's Online Privacy Protection Rule."

³⁴⁵ United States, "Platform Accountability and Transparency Act."

³⁴⁶ Eastwood, "Imperfect Legislative Solutions for Researcher Access to Social Media Data."

repository.³⁴⁷ PATA introduces the concept of “reasonably public” content, emphasizing both absolute thresholds and that the public nature of data depends on the circumstances of its creation and whether the data generator has a reasonable expectation of privacy. PATA defines reasonably public content as content that has been highly disseminated or originated from major public accounts.³⁴⁸ PATA excludes private messages, biometric, and precise geospatial information from accessible data by definition.³⁴⁹

Expanding access to public platform data also was included in DSOSA, introduced in 2022.³⁵⁰ DSOSA would create a Bureau of Digital Services Oversight and Safety at the Federal Trade Commission (FTC) and references “public content,” defined as “information on a covered platform that is available to a potentially unlimited number of third parties.”³⁵¹ DSOSA would require platforms to provide certified researchers access to a high-reach public content stream and enable the FTC to require platforms to provide a publicly-available public content stream.³⁵²

State level laws and proposals have also focused on platform transparency, research ethics, and privacy. In California, for example, Assembly Bill No. 587 would have required platforms to communicate their terms of service as well as content moderation actions.³⁵³ This bill was successfully enjoined by X in 2024, which argued that the obligation to publish detailed statistics on content moderation would violate the First Amendment.³⁵⁴ Indeed, a review of transparency regulation in the US finds that some laws have been successfully challenged by industry on First Amendment grounds.³⁵⁵ However, scholars have pointed out that better-designed transparency laws in the US should be able to survive constitutional scrutiny.³⁵⁶

Alongside platform transparency laws, privacy laws in several US states define publicly available information. These laws typically include reference to information that consumers make available to the general public through widely distributed media as well as whether the data controller has a “reasonable basis” to believe that the user has made the information available to the general public.³⁵⁷

Regulation in Europe

The DSA is the world’s most comprehensive platform transparency regime. It entered into force in early 2024 and requires VLOPs and VLOSEs³⁵⁸ to provide researchers with access to data to study systemic risks and measures taken to mitigate such risks. Systemic risks include the dissemination of illegal

³⁴⁷ This repository must be maintained on an ongoing basis and be accessible through a searchable tool that allows multi-criteria queries as well as through application programming interfaces (APIs). United States, “Platform Accountability and Transparency Act.”

³⁴⁸ Ibid.

³⁴⁹ United States, “Platform Accountability and Transparency Act.”

³⁵⁰ United States, “Digital Services Oversight and Safety Act.”

³⁵¹ Ibid.

³⁵² Ibid.

³⁵³ California, “AB-587.”

³⁵⁴ United States Court of Appeals for the Ninth Circuit, “Combined Opinion.”

³⁵⁵ Keller, “Platform Transparency and the First Amendment.”

³⁵⁶ Ibid.

³⁵⁷ See, e.g., Colorado, “Colorado Privacy Act”; Delaware, “Delaware Personal Data Privacy Act.”

³⁵⁸ Those having an average of more than 45 million monthly active users in the EU.

content, threats to fundamental rights, negative effects on civic discourse or electoral processes, and negative consequences for public health and physical and mental well-being.³⁵⁹ The DSA data access regime includes a focus on paid advertising (Article 39), privileged data access (40.4), access to publicly accessible data (40.12), among others.³⁶⁰

Platforms have built tools to further transparency around paid advertising. These efforts have been expanded by Article 39 of the DSA.³⁶¹ Due to this requirement, most major platforms (e.g., Meta,³⁶² TikTok,³⁶³ LinkedIn³⁶⁴) maintain some sort of access to ad archives, though sometimes that is limited to ads targeted in the E.U. (Pinterest,³⁶⁵ Snap³⁶⁶).

Data access requirements in the DSA make explicit reference to obligations in the GDPR, which would apply to personal data and transfers of data to third countries.³⁶⁷ The GDPR, which came into effect in May 2018, governs the collection, processing, storage, and transfer of personal data of individuals located in the EU.³⁶⁸ When platforms grant researcher access to data under Article 40 of the DSA, any personal data disclosed must comply with GDPR standards, such as implementing data minimization, purpose limitation, and appropriate measures for secure international data transfers.

Research based on privileged data access under the DSA must be focused on understanding systemic risks in the EU or the adequacy of the platforms' mitigation measures required under the Act. Researchers are vetted by the relevant supervisory authority and must demonstrate their affiliation to a research organization, demonstrate their independence from commercial interests, disclose their research funding, possess the capability to fulfill data security and confidentiality requirements, justify that their requested data access and timeframes are necessary and proportionate for their research, and commit to making the results of research publically accessible, free of charge, and within a reasonable period after the completion of the research.³⁶⁹

For access to public platform data under the DSA, researchers are expected to meet fewer requirements.³⁷⁰ Article 40 also specifies that research with public platform data should be related to "the detection, identification and understanding of systemic risks" as laid out in Article 34(1) of the DSA. This also means that a broader range of researchers may analyze publicly available information under 40.12, so long as they are independent from commercial interests, disclose funding, can maintain data integrity and protect personal data, and justify the necessity and proportionality of their request.

³⁵⁹ European Union, "Digital Services Act," Article 34.

³⁶⁰ Husovec, *Principles of the Digital Services Act*.

³⁶¹ European Union, "Digital Services Act," Article 39.

³⁶² Meta, "Ad Library."

³⁶³ TikTok, "Find Ads on TikTok."

³⁶⁴ LinkedIn, "Ad Library."

³⁶⁵ Pinterest, "Pinterest Ads Repository."

³⁶⁶ Snap, "Ads Gallery."

³⁶⁷ European Centre for Algorithmic Transparency, "FAQs."

³⁶⁸ European Union, "General Data Protection Regulation."

³⁶⁹ European Union, "Digital Services Act," Article 40.

³⁷⁰ Ibid.

Beyond the text of the DSA, the European Commission and relevant Digital Service Coordinators (DSCs) have provided limited public guidance to platforms in how to interpret and operationalize these requirements. Platforms have wide discretion in interpreting DSA requirements, including the degree to which research proposals contribute to “detection, identification and understanding” of systemic risks. The DSA 40 Data Access Collaboratory, an initiative of the European New School and the Weizenbaum Institute, has found that insufficient connection to systemic risks defined in Article 34 is a reason for rejection of data access requests.³⁷¹

While operationalizing data access under the DSA is still in its infancy, an independent and systematic review of public platform data access has found that mechanisms provided by platforms are uneven and often insufficient to meet the needs of researchers.³⁷² Questions remain around the ongoing costs of operationalizing data access programs under the DSA and other regulations.³⁷³ The European Commission has initiated formal proceedings against several major platforms due to alleged shortcomings in facilitating researcher access to platform data.³⁷⁴ As of November 2025, the Commission has found both TikTok and Meta in preliminary breach of the DSA’s obligations to provide researcher access to public data.³⁷⁵ This highlights the ongoing challenges in ensuring that platforms genuinely comply with the spirit of Article 40.12 and provide meaningful access to public platform data for research purposes.

Nonetheless, platforms have taken affirmative steps to enable access under the DSA. The MCL provides information from Threads, Facebook, and Instagram.³⁷⁶ YouTube provides an API that enables researchers to analyze non-private pages and data, though the API has specific rate limits on what can be accessed.³⁷⁷ TikTok created a new researcher API which differentiates between academic and non-academic researchers.³⁷⁸ TikTok’s and Meta’s research offerings have been criticized by researchers.³⁷⁹ Other platforms are building out regimes that enable researchers to apply to access public platform data in systematic ways.³⁸⁰ In June 2025, the European Commission announced that

³⁷¹ DSA40 Data Access Collaboratory, “Tracker Insights.”

³⁷² Hickey et al., “Public Data Access Programs.”

³⁷³ Edelson et al., “Access to Data and Algorithms.”

³⁷⁴ This includes investigations focused on AliExpress, Facebook, Instagram, TikTok, and X: European Commission, “Commission Opens Formal Proceedings against AliExpress under the Digital Services Act”; European Commission, “Commission Opens Formal Proceedings against Facebook and Instagram under the Digital Services Act”; European Commission, “Commission Opens Formal Proceedings against TikTok under the Digital Services Act”; European Commission, “Commission Sends Preliminary Findings to X for Breach of the Digital Services Act.”

³⁷⁵ European Commission, “Commission preliminarily finds TikTok and Meta in breach of their transparency obligations under the Digital Services Act.”

³⁷⁶ Meta, “Data Dictionary”; Meta, “Meta Content Library and API.”

³⁷⁷ Google, “Implementation Guide.”

³⁷⁸ TikTok, “Research Tools,” stating that access is limited to those affiliated with “academic institutions in the US, EEA, UK or Switzerland; or Not-for-profit and/or independent research institution, organization, association, or body in the EU...”

³⁷⁹ Alvarado et al., “Unpacking TikTok’s Data Access Illusion”; Gottfredsen and Dowling, “Meta Is Getting Rid of CrowdTangle—and Its Replacement Isn’t as Transparent or Accessible.”

³⁸⁰ For example:

- Booking.com: Researchers who are eligible under the Digital Services Act can apply for access to publicly available information, but the platform does not define the thresholds for data inclusion. See Booking.com, “DSA Data Researchers Request Form.”

AliExpress had made binding commitments to provide multiple forms of independent data access, including through automated collection (referencing ‘data scraping’), a dedicated API, and the provision of customized data sets upon request.³⁸¹ Other platforms, including Booking, Google, Wikipedia, and Amazon appear to authorize scraping for non-commercial purposes.³⁸²

The EU AI Act imposes transparency obligations on high-risk AI systems, requiring documentation, traceability, and clearly defined human oversight mechanisms.³⁸³ These systems must also be designed to inform users when they are interacting with AI.³⁸⁴ Additionally, providers of general-purpose AI models must disclose documentation of their training data and system capabilities.³⁸⁵ These provisions align with the goals of the DSA and reflect the EU’s broader approach to promoting transparency, accountability, and public interest oversight in digital governance.³⁸⁶

Other Regulations

While other jurisdictions’ social media and digital platform regulations include transparency elements, they do not mandate access to public platform data.

The UK’s Online Safety Act of 2023 introduces transparency, risk assessment, and content-related requirements. However, it does not explicitly mandate public access to platform data. The British government is contemplating new legislation, known as the Data Use and Access Bill, to provide outside researchers with mandatory data access to supervised companies.³⁸⁷ Ofcom, the UK’s regulator for communication services, conducted a call for evidence in early 2025 focused on how it could enable access to data for researchers.³⁸⁸ The UKRI, a public-funded research group, is similarly supporting a UK taskforce to provide guidance to the British government on how its potential data access provisions should be developed.

Singapore’s 2022 Online Safety (Miscellaneous Amendments) Act (OSMAA) and corresponding Code of Practice include requirements for inscope companies to publish annual online safety reports.³⁸⁹ For years, Brazil has debated a Law on Freedom, Responsibility and Transparency on the Internet (Bill

-
- Snapchat: Snap provides a contact for global researchers to apply for access to publicly available information, but does not detail thresholds for inclusion. See Snap, “Researcher Data Access Instructions.”
 - LinkedIn: LinkedIn offers a beta Researcher Access Program that offers access to data related to systemic risks in the European Union. LinkedIn does not detail thresholds for inclusion. See LinkedIn, “Data Access for Researchers.”
 - AliExpress: AliExpress offers an Open Research and Transparency portal to provide approved researchers, under Article 40(12) of the Digital Services Act, with access to publicly available information. See AliExpress, “Open Research.”
 - Bing: Bing’s Qualified Researcher Program makes the Bing Search API, Bing Webmaster Tools, and datasets available to researchers “free from commercial interests” globally. See Microsoft, “Bing Qualified Researcher Program.”

³⁸¹ European Commission, “Commission makes AliExpress’ commitments under the Digital Services Act binding.”

³⁸² Booking.com, “Customer terms of service.”

³⁸³ European Union, “Artificial Intelligence Act,” Articles 11, 12, and 14.

³⁸⁴ European Union, “Artificial Intelligence Act,” Article 50(1).

³⁸⁵ European Union, “Artificial Intelligence Act,” Article 53.

³⁸⁶ European Union, “Digital Services Act,” Article 40.

³⁸⁷ United Kingdom, “Data (Use and Access) Act 2025.”

³⁸⁸ Ofcom, “Call for Evidence.”

³⁸⁹ Singapore, “Broadcasting Act 1994.”

2630/20), colloquially called the “Fake News Bill.”³⁹⁰ While these regulations do not specifically seek to enable access to public platform data, they are part of a growing global movement to increase transparency of platforms through regulation.

³⁹⁰ Dias Silveira et al., “Internet Impact Brief.”

Appendix B. Research on Data Privacy

This appendix briefly describes research that explores how users understand the publicness of digital platforms and the data they share on those platforms.

Research with public platform data necessarily implicates diverse types of user data and may raise ethical and privacy risks, including specific risks associated with the potential for bad actors to exploit data access programs.³⁹¹

Prior studies have investigated how users understand the publicness of different platforms. A 2018 study of young adults highlighted that “young people are socialized to anticipate digital surveillance from various social institutions.”³⁹² Other studies have explored whether a user’s “imagined audience” – that is, a user’s “mental conceptualization of the people with whom he or she is communicating” – actually aligns with the actual audience for their digital platform data.³⁹³ This work underscores the need to proactively consider whether users reasonably understand their data to be public.

Understanding that one’s data is public is distinct from understanding that one’s data may be used for research. Some studies suggest that users are unaware of and somewhat disapproving of researcher use of public platform data.³⁹⁴ Other studies find that users are “generally okay” with research use cases.³⁹⁵ Users generally seem to be more comfortable with researcher use of data in larger information environments, consistent with the *Better Access* approach.³⁹⁶ The public interest value of researchers studying the online information environment must be balanced with users’ hesitation regarding such use.

A user’s reasonable understanding of whether their data is public and visible is a foundational consideration for determining whether that data should be made available for research. Contextual integrity³⁹⁷ and related privacy scholarship, outline a number of different scenario-specific parameters upon which privacy should be determined.³⁹⁸ This collective work encourages the understanding of privacy as inherently context-dependent, and empirical work shows that what users consider “private” tends to depend on the circumstances.³⁹⁹

³⁹¹ Working Group on Platform-to-Researcher Data Access, “Report of the European Digital Media Observatory’s Working Group on Platform-to-Researcher Data Access.”

³⁹² Duffy and Chan, “‘You never really know who’s looking’.”

³⁹³ Litt, “Knock, Knock. Who’s There? The Imagined Audience.” See also Mark and boyd, “I tweet honestly, I tweet passionately.”

³⁹⁴ Fiesler and Proferes, “‘Participant’ Perceptions of Twitter Research Ethics”; Gilbert et al., “Measuring Americans’ Comfort With Research Uses of Their Social Media Data.”

³⁹⁵ Hemphill et al., “Comparative sensitivity of social media data and their acceptable use in research.”

³⁹⁶ Hudson and Bruckman, “Go Away.”

³⁹⁷ Barth et al., “Privacy and Contextual Integrity.”

³⁹⁸ Nissenbaum, “Contextual integrity up and down the data food chain.” Specifically, “privacy norms prescribe information flows with reference to five parameters — sender, recipient, subject, information type, and transmission principle.”

³⁹⁹ Gerdon, “Attitudes on data use for public benefit”; Martin and Nissenbaum, “Measuring Privacy”; Vitak et al., “When do data collection and use become a matter of concern?”

While frameworks exist for operationalizing contextual integrity,⁴⁰⁰ applying this concept (which is largely decided on a case-by-case basis) at the scale of platforms is a significant challenge. Platforms have billions of users across diverse cultural, linguistic, educational, and geographic contexts. Assessing each user's contextually dependent privacy preferences to determine access to data, as opposed to what data to use and how to communicate, is impractical. However, contextual assessments should be done by researchers as they are analyzing data in accordance with relevant ethical and privacy standards.

⁴⁰⁰ Barth et al., "Privacy and Contextual Integrity."

Appendix C. Proactive Data Interface Thresholds

Platform Thresholds for Highly Disseminated Content in Proactive Data Interfaces

Platform	Reach Threshold	Engagement Threshold
CrowdTangle		Included posts to public Facebook pages with more than 25,000 Page likes or followers ⁴⁰¹
Facebook		MCL includes posts to public Pages with 15,000 or more likes or followers. ⁴⁰²
YouTube	YouTube Data API enables analysis of all non-private pages and data, irrespective of reach or engagement. ⁴⁰³	
TikTok	TikTok's research API includes all public TikTok videos posted by a public creator (who is aged 18 and over). ⁴⁰⁴	
Reddit	Reddit's API enables research on public (or non-private) data without specific reach or engagement thresholds. ⁴⁰⁵	

⁴⁰¹ CrowdTangle, "What data is CrowdTangle tracking?"

⁴⁰² Meta, "Data Dictionary"; Meta, "Meta Content Library and API."

⁴⁰³ Google, "Implementation Guide."

⁴⁰⁴ TikTok, "Codebook."

⁴⁰⁵ Reddit, "Addressing the community about changes to our API."

Platform Thresholds for Notable Public Accounts in Existing Proactive Data Interfaces

Platform	Reach	Engagement	Status
CrowdTangle	Facebook Pages with more than 25,000 page likes or followers, all global public Facebook groups with more than 95,000 members, US-based public groups with more than 2,000 members, and all verified profiles. ⁴⁰⁶ Instagram all verified public accounts or accounts with more than 50,000 followers. ⁴⁰⁷		
Facebook ⁴⁰⁸	MCL enables researchers to: - View public profiles with a verified badge or more than 1,000 followers; - Download posts from public profiles with a verified badge or more than 25,000 followers or posts from public Pages with more than 15,000 followers.	MCL enables researchers to download posts from public Pages with more than 15,000 likes.	
Instagram ⁴⁰⁹	MCL enables researchers to: - View posts from public personal users accounts with more than 1,000 followers; - Download posts from personal accounts set to public with a verified badge or more than 25,000 followers.		MCL enables researchers to: - View Instagram posts from public business and creator accounts; - Download Instagram posts from public business and creator accounts with a verified badge or more than 25,000 followers.
Threads	MCL enables researchers to access Threads shared by public profiles with more than 1,000 followers		

⁴⁰⁶ CrowdTangle, “What data is CrowdTangle tracking?”

⁴⁰⁷ CrowdTangle, “What data is CrowdTangle tracking?”

⁴⁰⁸ Meta, “Public figures safety.”

⁴⁰⁹ Ibid.

X ⁴¹⁰	X's Public Interest policy applies to accounts having more than 100,000 followers.		
YouTube ⁴¹¹	YouTube Data API (v3) enables analysis of all non private pages and data, irrespective of status.		
TikTok ⁴¹²	TikTok's research API includes all public TikTok videos posted by a public creator (who is aged 18 and over), who wants to expose their videos to all users of TikTok ⁴¹³	TikTok includes media verification, requiring users to "provide credible written media coverage, such as news articles." ⁴¹⁴	
Reddit ⁴¹⁵	Reddit's API enables research on public (or non-private) data without specific reach or engagement thresholds.		

⁴¹⁰ X, "Defining public interest on Twitter."

⁴¹¹ YouTube, "Implementation Guide."

⁴¹² TikTok, "Verified accounts on TikTok."

⁴¹³ TikTok, "Codebook."

⁴¹⁴ TikTok, "Codebook."

⁴¹⁵ Reddit, "Addressing the community about changes to our API."

Appendix D. Fields Included in Previous Platform Data Tools

CrowdTangle

The CrowdTangle tool formerly acquired and hosted by Meta included multiple fields of data, including:⁴¹⁶

- Date-time of creation
- Type of post
- # Likes
- # Comments
- # Shares
- # Love
- # Wow
- # Haha
- # Sad
- # Angry
- # Thankful
- Video share status
- Post views
- Total views
- Post URL
- Message (Post text)
- Link
- Link text
- Score (level at which a post overperformed)
- For an individual post, time-series engagement data

Importantly, CrowdTangle also included benchmarks of typical metrics for fields. Benchmarks were calculated based on the same post type from the same page, they were not broadly applicable across different pages or groups. However, time-series data (amount of engagement recorded every 15 min for the first 8 hours and then in lengthening intervals subsequently), which was used to calculate the benchmarks, was one of the most-used features by researchers. The MCL includes similar fields,⁴¹⁷ but with greatly restricted access to researchers and journalists.⁴¹⁸

Twitter/X

Twitter's Firehose API initially provided users with tremendously broad data access. It provided researchers with *all* tweets on the platform in real-time.⁴¹⁹ The Twitter/X API now returns the following fields of data:⁴²⁰

- Tweet ID
- Tweet text
- Tweet edit history
- Attachments
- Author ID
- Context annotations
- Conversation ID
- Data of creation
- Entities mentioned
- Reply to user
- Organic metrics
- Promoted metrics
- Possibly sensitive indicator
- Public metrics (including retweets, replies, likes, and quote tweets)
- Referenced tweets
- Reply settings

⁴¹⁶ Garmur et al., "CrowdTangle Codebook."

⁴¹⁷ Meta, "Data Dictionary."

⁴¹⁸ Gotfredsen and Dowling, "Meta Is Getting Rid of CrowdTangle—and Its Replacement Isn't as Transparent or Accessible."

⁴¹⁹ Hofer-Shall, "Working Directly With the Twitter Data Ecosystem."

⁴²⁰ X, "X API v2 data dictionary."

TikTok

In 2023,⁴²¹ TikTok provided researcher access to its API.⁴²² The API includes public TikTok video posted by a public creator (who is aged 18 and over), who wants to expose their videos to all users of TikTok and returns the following fields of data, among others:⁴²³

- ID
- Create Time
- Username
- Region Code
- Video Description
- Music ID
- Like Count
- Comment Count
- View Count
- Video ID/URL link
- Hashtags, ID, and description
- Video_mention_list
- Video_label
- Playlist_ID
- Voice_to_text
- Video_duration
- Favorites_count
- Comment create time
- Comment ID
- Comment like count
- Parent comment ID and parent video ID
- Comment reply count
- Comment text
- User following, like, and video count
- User bio description and URL
- User display name
- User follower count
- User Avatar URL
- User verification
- Liked videos from user, including array of parent content statistics
- Reposted videos from user, including array of parent content statistics
- Pinned videos from a user, including array of parent content statistics
- Followers list
- Following list
- Playlist info

⁴²¹ Lurie and Schroeder, “TikTok just announced the data it’s willing to share.”

⁴²² Edelson, “Getting to know the TikTok Research API.”

⁴²³ TikTok, “Codebook.”

Bibliography

- Abdo, Alex, Ramya Krishnan, Stephanie Krent, Evan Welber Falcón, and Andrew Keane Woods. “A Safe Harbor for Platform Research.” Knight First Amendment Institute at Columbia University, January 19, 2022. <https://knightcolumbia.org/content/a-safe-harbor-for-platform-research>.
Access Now. “Transparency Reporting Index.” Accessed May 21, 2025.
<https://www.accessnow.org/campaign/transparency-reporting-index/>.
- Access Now, ACLU Foundation of North California, ACLU Foundation of North Carolina, et al. “Santa Clara Principles on Transparency and Accountability in Content Moderation.” 2018.
<https://santaclaraprinciples.org/>.
- Acemoglu, Daron, Daniel Huttenlocher, Asuman Ozdaglar, and James Siderius. “Online Business Models, Digital Ads, and User Welfare.” National Bureau of Economic Research, October 2024.
<https://www.nber.org/papers/w33017>.
- AFP. “AFP editorial standards and best practices.” December 3, 2024.
https://www.afp.com/sites/default/files/afp_ethic_march_2024.pdf.
- Agre, Philip E. “Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI.” In *Social Science, Technical Systems, and Cooperative Work*, edited by Geoffrey C. Bowker, Les Gasser, Leigh Star, and William Turner. Psychology Press, 1998.
- Alba, Davey. “Facebook sent flawed data to misinformation researchers.” *New York Times*, September 10, 2021.
<https://www.nytimes.com/live/2020/2020-election-misinformation-distortions#facebook-sent-flawed-data-to-misinformation-researchers>.
- Albert, John. “Platforms’ Promises to Researchers: First Reports Missing the Baseline.” *AlgorithmWatch*, February 2023.
<https://algorithmwatch.org/en/platforms-promises-to-researchers/>.
- Ali, Shiza, Mohammad Hammas Saeed, Esraa Aldreabi, et al. “Understanding the Effect of Deplatforming on Social Networks.” *Proceedings of the 13th ACM Web Science Conference* (2021). <https://seclab.bu.edu/people/gianluca/papers/deplatforming-websci2021.pdf>.
- AliExpress. “Open Research & Transparency: Guidelines for Researchers to Apply Public Data Access.” Accessed May 21, 2025.
<https://www.aliexpress.com/p/transparencycenter/openResearch.html>.
- Aljabri, Malak, Rachid Zagrouba, Afrah Shaahid, Fatima Alnasser, Asalah Saleh, and Dorieh M. Alomari. “Machine learning-based social media bot detection: a comprehensive literature review.” *Social Network Analysis and Mining* 13, no. 20 (2023).
<https://link.springer.com/article/10.1007/s13278-022-01020-5>.
- Allyn, Bobby, Silvia Goodman, and Dara Kerr. “States probed TikTok for years. Here are the documents the app tried to keep secret.” *OPB*, October 11, 2024.
<https://www.opb.org/article/2024/10/11/tiktok-knows-its-app-is-harming-kids-new-internal-documents-show/>.

- American Sunlight Project. “EcoBoost: A malign influence network targeting the environmental and tech sectors in Western democracies.” March 4, 2025.
<https://static1.squarespace.com/static/6612cbdfd9a9ce56ef931004/t/67c72bac9804be60f4251e3a/1741106103480/ASP+EcoBoost.pdf>.
- Anthropic. “Anthropic Economic Index: Insights from Claude 3.7 Sonnet.” March 2025.
<https://www.anthropic.com/news/anthropic-economic-index-insights-from-claude-sonnet-3-7>.
- Apple. “Listener analytics.” Accessed May 21, 2025.
<https://podcasters.apple.com/support/5392-listener-analytics>.
- Argentina. “Listado Alfabético.” Accessed May 21, 2025.
<https://www.senado.gob.ar/senadores/listados/listaSenadoRes>.
- Associated Press. “Meta says it will resume AI training with public content from European users.” *Associated Press*, April 14, 2025.
<https://apnews.com/article/facebook-instagram-meta-ai-europe-c785dc3591ae3c49543c435fc15379fb>.
- Australia. “Australian Privacy Principles.” Office of Information Commissioner. Accessed May 21, 2025.
<https://www.oaic.gov.au/privacy/australian-privacy-principles>.
- — —. “Online Safety (Basic Online Safety Expectations) Determination 2022, F2022L00062.” Federal Register of Legislation, January 2022. <https://www.legislation.gov.au/F2022L00062/latest/text>.
- Ayrshare. “Top 10 Social Media APIs for Developers.” February 2025.
<https://www.ayrshare.com/top-10-social-media-apis-for-developers/>.
- Ballotpedia. “Ballotpedia.” Accessed May 21, 2025. www.ballotpedia.org.
- Barth, Adam, Anupam Datta, John C. Mitchell, and Helen Nissenbaum. “Privacy and Contextual Integrity: Framework and Applications.” *IEEE Symposium on Security and Privacy* (2006).
<https://ieeexplore.ieee.org/document/1624011>.
- Bärtl, Mathias. “YouTube channels, uploads and views: A statistical analysis of the past 10- years.” *Convergence* 24, no. 1 (2018): 16-32.
https://www.researchgate.net/publication/322431948_YouTube_channels_uploads_and_views_A_statistical_analysis_of_the_past_10_years.
- Barnea, Uri, Robert J. Meyer, and Gideon Nave. “The Effects of Content Ephemerality on Information Processing.” *Journal of Marketing Research* 60, no. 4 (2023): 750-766.
<https://journals.sagepub.com/doi/10.1177/00222437221131047>.
- Bayer, Joseph B., Nicole B. Ellison, Sarita Y. Schoenebeck, and Emily B. Falk. “Sharing the small moments: ephemeral social interaction on Snapchat.” *Information, Communication & Society* 19, no. 7 (2016): 956-977. <https://www.tandfonline.com/doi/abs/10.1080/1369118X.2015.1084349>.
- Beaumont, Julie. “GPEN Sweep 2024 on deceptive design patterns.” Global Privacy Enforcement Network, July 9, 2024.
<https://www.privacyenforcement.net/content/2024-gpen-sweep-deceptive-design-patterns>.
- Ben-Shahar, Omri, and John AE Pottow. “On the stickiness of default rules.” *Florida State University Law Review* 33 (2006). <https://ir.law.fsu.edu/lr/vol33/iss3/5/>.
- Benoliel, Uri, and Shmuel I. Becher. “The Duty to Read the Unreadable.” *Boston College Law Review* 60, no. 8. (2019): 2255. <https://bclawreview.bc.edu/articles/320/files/63a92425205eb.pdf>.
- Bluesky. “Bluesky Developer APIs.” Accessed October 21, 2025. <https://docs.bsky.app/>.
- — —. “Bluesky for Journalists.” October 2023. <https://bsky.social/about/blog/press-faq>.

Booking.com. “Customer terms of service.” November 2023.

<https://www.booking.com/content/terms.html>.

— — —. “DSA Data Researchers Request Form.” Accessed May 21, 2025.

<https://dsarequests.moderation.booking.com/hc/en-gb/articles/23752208094996-DSA-Data-Researchers-Request-form>.

Borchers, Conrad, Joshua M. Rosenberg, and Rita M. Swartzentruber. “Facebook Post Data: A Primer for Educational Research.” *Educational Technology Research and Development* 71, no. 6 (2023): 2345–64. <https://doi.org/10.1007/s11423-023-10269-2>.

Brazil. “Lista Candidatos Eleições 2024.” Accessed May 21, 2025.

<https://listadoscandidatos.com.br/candidatos>.

— — —. “Projeto de Lei N, de 2022.” Accessed May 21, 2025.

https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=2214237&filename=PL%202768%2F2022.

— — —. “Projeto de Lei n° 2338, 2023.” 2023.

<https://www25.senado.leg.br/web/atividade/materias/-/materia/157233>.

Bright Data. “The Bright Initiative.” Accessed September 18, 2025. <https://brightinitiative.com/>.

— — —. “Limitless web data infrastructure for AI & BI.” Accessed September 18, 2025.

<https://brightdata.com/>.

Brown, Megan A. “The Problem with TikTok’s New Researcher API is Not TikTok.” *Tech Policy Press*, March 2023.

<https://www.techpolicy.press/the-problem-with-tiktoks-new-researcher-api-is-not-tiktok/>.

Brown, Megan A., Andrew Gruen, Gabe Maldoff, Solomon Messing, Zeve Sanderson, and Michael Zimmer. “Web Scraping for Research: Legal, Ethical, Institutional, and Scientific Considerations.” October 2024. <https://www.meganbrown.org/publications/b-web-scraping/>.

Budak, Ceren, Brendan Nyhan, David M. Rothschild, Emily Thorson, and Duncan J. Watts.

“Misunderstanding the Harms of Online Misinformation.” *Nature* 630, no. 8015 (2024): 45–53.

<https://doi.org/10.1038/s41586-024-07417-w>.

Bundtzen, Sara. “Data Access.” Institute for Strategic Dialogue, accessed May 21, 2025.

<https://www.isdglobal.org/explainers/data-access/>.

Bundtzen, Sara, and Christian Schwieter. “Access to Social Media Data for Public Interest Research: Lessons Learnt and Recommendations for Strengthening Initiatives in the EU and Beyond.” Institute for Strategic Dialogue, May 2023.

<https://www.isdglobal.org/isd-publications/researcher-access-to-social-media-data-lessons-learn-t-recommendations-for-strengthening-initiatives-in-the-eu-beyond/>.

Busby, John C. “expectation of privacy.” Legal Information Institute, December 2022.

https://www.law.cornell.edu/wex/expectation_of_privacy.

Cáceres, Marcos, and Reilly Grant. “Geolocation.” World Wide Web Consortium, July 11, 2025.

<https://www.w3.org/TR/2025/REC-geolocation-20250711/>.

California. “AB-587 Social Media Companies: terms of service.” 2022.

https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=202120220AB587.

— — —. “California Consumer Privacy Act.” Office of the Attorney General, March 2024.

<https://oag.ca.gov/privacy/ccpa>.

Canada. “Artificial Intelligence and Data Act.” Last modified September 27, 2023.

<https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act>.

CandiData. “CandiData24.” Accessed May 21, 2025. <https://www.candidata24.org/>.

Carroll, Rory. “Are they ‘having an affair’ or just shy? The couple caught on Coldplay’s kiss cam.” *The Guardian*, July 18, 2025.

<https://www.theguardian.com/music/2025/jul/18/couple-caught-coldplay-kiss-cam-affair-very-shy>.

Casas, Andreu, Georgia Dagher, and Ben O’Loughlin. “Commentary: How do we get platforms to share data with independent researchers? Regulation alone will not cut it: a commentary on Livingston et al. (2023), Bourgaize et al. (2025).” *Child Adolescent Mental Health* 30, no. 3 (2025).

<https://pubmed.ncbi.nlm.nih.gov/40685332/>.

The Carter Center. “Election Management.” Accessed August 15, 2025.

<https://eos.cartercenter.org/parts/8>.

Center for American Progress. “CrowdTangle Letter.” May 2024.

<https://www.americanprogress.org/wp-content/uploads/sites/2/2024/05/CAP-ISD-letter2.pdf>.

Center for an Informed Public. “Twitter’s API Access Changes Could Mark ‘End of an Era’ in Academic Research on the Platform.” February 2023.

<https://www.cip.uw.edu/2023/02/02/twitters-api-access-changes-academic-research/>.

Chancellor, Stevie, Zhiyuan Lin, and Munmun De Choudhury. “‘This Post Will Just Get Taken Down’: Characterizing Removed Pro-Eating Disorder Social Media Content.” *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016).

<https://dl.acm.org/doi/10.1145/2858036.2858248>.

Chapkanovska, Evangelina. “YouTube Revenue Statistics of 2024: Play Video!” SpendMeNot, July 2025. <https://spendmenot.com/blog/youtube-revenue-statistics/>.

Chen, Kuan-Ju, and Hoi Ling Cheung. “Unlocking the power of ephemeral content: The roles of motivations, gratification, need for closure, and engagement.” *Computers in Human Behavior* 97 (2019): 67–74. <https://www.sciencedirect.com/science/article/abs/pii/S0747563219300998>.

Chee, Foo Yun. “X hit by complaints to EU over user data and targeted advertising.” *Reuters*, July 15, 2025.

<https://www.reuters.com/sustainability/boards-policy-regulation/x-hit-by-complaints-eu-over-user-data-targeted-advertising-2025-07-15/>.

Clark, Mike. “Research Cannot Be the Justification for Compromising People’s Privacy.” Meta, August 3, 2021.

<https://about.fb.com/news/2021/08/research-cannot-be-the-justification-for-compromising-peoples-privacy/>.

Coalition for Independent Technology Research. “Letter: Imposing Fees to Access the Twitter API Threatens Public-Interest Research.” February, 2023.

<https://independenttechresearch.org/letter-twitter-api-access-threatens-public-interest-research/>.

— — —. “The State of Independent Technology Research 2025: Power in Numbers.” August, 2025.

<https://independenttechresearch.org/wp-content/uploads/2025/08/The-State-of-Independent-Technology-Research-Power-in-Numbers.pdf>.

Colorado. “Colorado Privacy Act.” July 8, 2021.

https://coag.gov/app/uploads/2022/01/SB-21-190-CPA_Final.pdf.

- Confessore, Nicholas. "Cambridge Analytica and Facebook: The Scandal and the Fallout So Far." *New York Times*, April 2018.
<https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html>.
- Connolly, Jennifer M., Leticia Bode, and Ben Epstein. "Explaining the varying levels of adoption of e-government services in American municipal government." *State and Local Government Review* 50, no. 3 (2018): 150-164.
<https://journals-sagepub-com.proxy.library.georgetown.edu/doi/abs/10.1177/0160323X18808561>.
- Cottrell, Jill, and Yash Ghai. "The Role of Constitution-Building Processes in Democratization." International Institute for Democracy and Electoral Assistance, 2004.
<https://constitutionnet.org/sites/default/files/CBP-Fiji.pdf>.
- Counts, Aicha. "TikTok's Rules Deter Researchers from Crunching Data on Users, Misinformation." *Bloomberg*, September 2023.
<https://www.bloomberg.com/news/articles/2023-09-21/tiktok-terms-of-service-strict-for-researchers>.
- Coyne, Bridget. "Introducing U.S. Election Labels for Midterm Candidates." X, May 2018.
https://blog.x.com/en_us/topics/company/2018/introducing-us-election-labels-for-midterm-candidates.
- Cozma, Ioana. "91 Instagram Statistics to Track NOW: Strategic Implications and Our Agency's Perspective." InBeat, January 4, 2025. <https://inbeat.agency/blog/instagram-statistics>.
- CrowdTangle. "What data is CrowdTangle tracking?" Accessed May 21, 2025.
<https://web.archive.org/web/20231001153923/https://help.crowdtangle.com/en/articles/1140930-what-data-is-crowdtangle-tracking>.
- D'Onfro, Jillian. "This group of 60 million Facebook users could be the secret to Facebook's next growth spurt." *Business Insider*, July 2016.
<https://www.businessinsider.com/facebook-targets-small-businesses-2016-7>.
- Darius, Philipp. "Researcher Data Access Under the DSA: Lessons from TikTok's API Issues During the 2024 European Elections." *Tech Policy Press*, September 24, 2024.
<https://www.techpolicy.press/-researcher-data-access-under-the-dsa-lessons-from-tiktoks-api-issues-during-the-2024-european-elections/>.
- Davidson, Brittany I., Darja Wischerath, Daniel Racek, et al. "Platform-controlled social media APIs threaten open science." *Nature Human Behaviour* 7, no. 12 (2023): 2054-2057.
<https://pubmed.ncbi.nlm.nih.gov/37919445/>.
- DeTar, Charlie. "On Selling Out: How the Conditions of Technological Design Perpetuate the Status Quo." 2016. <https://tirl.org/writing/cfd-on-selling-out.pdf>.
- Delaware. "Delaware Personal Data Privacy Act." September 11, 2023.
<https://legis.delaware.gov/json/BillDetail/GenerateHtmlDocument?legislationId=140388&legislationTypeId=1&docTypeId=2&legislationName=HB154>.
- Democracy Reporting International. "Big tech is backing out of commitments countering disinformation—What's Next for the EU's Code of Practice?" Democracy Reporting International, July 2024.
<https://democracy-reporting.org/en/office/EU/publications/big-tech-is-backing-out-of-commitments-countering-disinformation-whats-next-for-the-eus-code-of-practice>.

- Dias Silveira, Ana Carolina Rodrigues, Anala Paula Camelo, et al. “Internet Impact Brief: Proposals to Regulate Digital Platforms in Brazil, Potential Impacts in the Internet.” Internet Society, February 19, 2025.
<https://www.internetsociety.org/resources/doc/2025/proposals-to-regulate-digital-platforms-in-brasil-potential-impacts-in-the-internet/>.
- DiResta, Renée. *Invisible Rulers: The People Who Turn Lies into Reality*. PublicAffairs, June 11, 2024.
- DSA40 Data Access Collaboratory. “Tracker Insights.” Last modified May 26, 2025.
<https://dsa40collaboratory.eu/tracker-insights/>.
- Duffy, Brooke Erin, and Ngai Keung Chan. “You never really know who’s looking’: Imagined surveillance across social media platforms.” *New Media & Society* 21, no. 1 (2018).
<https://journals.sagepub.com/doi/10.1177/1461444818791318>.
- Eastwood, Brent M. “Imperfect Legislative Solutions for Researcher Access to Social Media Data.” R Street Institute, December 12, 2023.
<https://www.rstreet.org/commentary/imperfect-legislative-solutions-for-researcher-access-to-social-media-data/>.
- Edelson, Laura. “Getting to know the TikTok Research API.” Cybersecurity for Democracy, May 2024.
<https://cybersecurityfordemocracy.org/getting-to-know-the-tiktok-research-api>.
- Edelson, Laura, Borys Kovba, Hanna Yershova, Austin Botelho, Damon McCoy, and Tobias Lauinger. “Measurement and Metrics for Content Moderation: The Multi-Dimensional Dynamics of Engagement and Content Removal on Facebook.” *Journal of Online Trust and Safety* 2, no. 5 (April 2025). <https://tsjournal.org/index.php/jots/article/view/220/87>.
- Edelson, Laura, Inge Graef, and Filippo Lancieri. “Access to Data and Algorithms: For an Effective DMA and DSA Implementation.” Centre on Regulation in Europe, March 2023.
<https://cerre.eu/wp-content/uploads/2023/03/CERRE-Access-to-Data-Algorithms.pdf>.
- Electronic Privacy Information Center (EPIC). “Social Media Privacy.” Accessed May 21, 2025.
<https://epic.org/issues/consumer-privacy/social-media-privacy/>.
- European Centre for Algorithmic Transparency. “FAQs: DSA data access for researchers.” July 3, 2025.
https://algorithmic-transparency.ec.europa.eu/news/faqs-dsa-data-access-researchers-2025-07-03_en.
- European Commission. “The Code of Conduct on Disinformation.” February 2025.
<https://digital-strategy.ec.europa.eu/en/library/code-conduct-disinformation>.
- — —. “Commission makes AliExpress’ commitments under the Digital Services Act binding.” June 18, 2025.
<https://digital-strategy.ec.europa.eu/en/news/commission-makes-aliexpress-commitments-under-digital-services-act-binding>.
- — —. “Commission Opens Formal Proceedings against AliExpress under the Digital Services Act.” March 2024. https://ec.europa.eu/commission/presscorner/detail/en/ip_24_1485.
- — —. “Commission Opens Formal Proceedings against Facebook and Instagram under the Digital Services Act,” April 2024. https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2373.
- — —. “Commission Opens Formal Proceedings against TikTok under the Digital Services Act.” February 2024. https://ec.europa.eu/commission/presscorner/detail/en/ip_24_926.

- — —. “Commission preliminarily finds TikTok and Meta in breach of their transparency obligations under the Digital Services Act.” October 24, 2025.
https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2503.
- — —. “Commission Sends Preliminary Findings to X for Breach of the Digital Services Act.” July 2024. https://ec.europa.eu/commission/presscorner/detail/en/ip_24_3761.
- — —. “Delegated Regulation on data access provided for in the Digital Services Act.” Accessed May 21, 2025.
https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13817-Delegated-Regulation-on-data-access-provided-for-in-the-Digital-Services-Act_en.
- — —. “Micro-, Small- and Medium-Sized Enterprises: Definition and Scope.” November 2016.
<https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=legisum%3An26026>.
- — —. “Status Report: Mechanisms for Researcher Access to Online Platform Data.” April 5, 2024.
<https://digital-strategy.ec.europa.eu/en/library/status-report-mechanisms-researcher-access-online-platform-data>.
- Ershov, Daniel. “Market incentives and advertising disclosure regulations.” In *The Hashtag Hustle: Law and Policy Perspectives on Working in the Influencer Economy*, edited by Taylor Annabell, Christian Fieseler, Catalina Goanta, and Isabelle Wildhaber. Edward Elgar Publishing, January 21, 2025.
- European Union. “Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation).” 2016. <https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng>.
- — —. “Regulation (EU) 2022/1925 of European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act).” 2022. <https://eur-lex.europa.eu/eli/reg/2022/1925/oj/eng>.
- — —. “Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act).” 2022. <https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng>.
- — —. “Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).” 2024.
<https://artificialintelligenceact.eu/ai-act-explorer/>.
- Facebook. “Basic Privacy Settings & Tools.” Accessed May 21, 2025.
<https://www.facebook.com/help/325807937506242/>.
- — —. “Policy Forum.” October 6, 2020.
https://about.fb.com/wp-content/uploads/2021/10/Facebook_PolicyForum_Definition-of-Public-Figures.pdf.
- — —. “Recommendation: Definition of Public Figures.” October 6, 2020.
https://about.fb.com/wp-content/uploads/2021/10/Facebook_PolicyForum_Definition-of-Public-Figures.pdf.
- — —. “Terms of Service.” Accessed January 24, 2022. <https://perma.cc/LWX7-JH6Z>.
- Federal Election Commission. “Making Electioneering Communications.” Accessed May 20, 2025.
<https://www.fec.gov/help-candidates-and-committees/other-filers/making-electioneering-communications/>.

- Federal Reserve. “Federal Trade Commission Act Section 5: Unfair or Deceptive Acts or Practices Background.” Accessed May 21, 2025.
<https://www.federalreserve.gov/boarddocs/supmanual/cch/200806/ftca.pdf>.
- Federal Trade Commission. “.com Disclosures: How to Make Effective Disclosures in Digital Advertising.” March 2013.
<https://www.ftc.gov/sites/default/files/attachments/press-releases/ftc-staff-revises-online-advertising-disclosure-guidelines/130312dotcomdisclosures.pdf>.
- — —. “Children’s Online Privacy Protection Rule.” January 17, 2023.
<https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa>.
- — —. “Complaint for Civil Penalties, Injunction, and Other Relief.” July 24, 2019.
https://www.ftc.gov/system/files/documents/cases/182_3109_facebook_complaint_filed_7-24-19.pdf.
- — —. “Disclosures 101 for Social Media Influencers.” November 2019.
<https://www.ftc.gov/business-guidance/resources/disclosures-101-social-media-influencers>.
- Ferreira, Díaz, Nicolás Emilio, Gautam Kishore Shahi, Catherine Tony, Stefan Stieglitz, and Riccardo Scandariato. “Regret, Delete, (Do Not) Repeat: An Analysis of Self-Cleaning Practices on Twitter After the Outbreak of the COVID-19 Pandemic.” *CHI Conference on Human Factors in Computing Systems* (2023). <https://dl.acm.org/doi/abs/10.1145/3544549.3585583>.
- Fiesler, Casey, and Nicholas Proferes. “‘Participant’ Perceptions of Twitter Research Ethics.” *Social Media + Society* 4, no. 1 (2018): 1-14.
<https://journals.sagepub.com/doi/10.1177/2056305118763366>.
- Firdaus, Asno. “Indonesia Presidential Candidate’s Dataset, 2024.” December 12, 2023.
<https://data.mendeley.com/datasets/7w5zvr8jgp/5>.
- franzke, aline, Anja Bechmann, Michael Zimmer, et al. “Internet Research: Ethical Guidelines 3.0.” Association of Internet Researchers, October 6, 2019. <https://aoir.org/reports/ethics3.pdf>.
- Freelon, Deen. “Computational Research in the Post-API Age.” *Political Communication* 35, no. 4 (2018): 665–68. <https://doi.org/10.1080/10584609.2018.1477506>.
- Gabon. “Le Guide De L’Agent Public.” Accessed June 23, 2025.
<https://fonction-publique.gouv.ga/object.getObject.do?id=1869>.
- Gadde, Vijaya, and David Gasca. “Measuring healthy conversation.” X, July 30, 2018.
https://blog.x.com/en_us/topics/company/2018/measuring_healthy_conversation.
- Gadde, Vijaya, and Yoel Roth. “Enabling further research of information operations on Twitter.” Twitter, October 17, 2018.
https://blog.x.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.
- Garmur, Matthew, Gary King, Zagreb Mukerjee, Nathaniel Persily, and Brandon Silverman. “CrowdTangle Codebook.” *Social Science One*, October 15, 2019.
<https://socialscience.one/crowdtangle-codebook>.
- Gordon, Frederic. “Attitudes on data use for public benefit: Investigating context-specific differences across Germany, Spain, and the United Kingdom with a longitudinal survey experiment.” *Social Media + Society* 10, no. 4 (2024).
<https://journals.sagepub.com/doi/10.1177/20563051241301202>.

- Gilbert, Sarah, J. Nathan Mathias, Ethan Zuckerman, Yellowmix, and James Mickens. “Survey Report: Reddit’s Actions Continue to Undermine Moderation & Research.” Coalition for Independent Technology Research, July 12, 2023. <https://independenttechresearch.org/reddit-survey-results/>.
- Gilbert, Sarah, Katie Shilton, and Jessica Vitak. “Measuring Americans’ Comfort With Research Uses of Their Social Media Data.” *Social Media + Society* 7, no. 3 (2021): 1-13. <https://journals.sagepub.com/doi/10.1177/20563051211033824>.
- — —. “When research is the context: Cross-platform user expectations for social media data reuse.” *Big Data & Society* 10, no. 1 (2023): 1-14. <https://journals.sagepub.com/doi/10.1177/20539517231164108>.
- Gotfredsen, Sarah Grevy, and Kaitlyn Dowling. “Meta Is Getting Rid of CrowdTangle—and Its Replacement Isn’t as Transparent or Accessible.” *Columbia Journalism Review*, July 9, 2024. https://www.cjr.org/tow_center/meta-is-getting-rid-of-crowdtangle.php.
- Gottfried, Jeffrey. “Americans’ Social Media Use.” Pew Research Center, January 31, 2024. <https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/>.
- Grüning, David J., Julia Kamin, Emily Saltz, et al. “Independently testing prosocial interventions: Methods and recommendations from 31 researchers.” *Annals of the New York Academy of Sciences* (2025): 1-14. <https://nyaspubs.onlinelibrary.wiley.com/doi/pdf/10.1111/nyas.15393>.
- Guhl, Jakob, Oliver Marsh, and Henry Tuck. “Researching the Evolving Online Ecosystem: Barriers, Methods and Future Challenges.” Institute for Strategic Dialogue, July 21, 2022. <https://www.isdglobal.org/isd-publications/researching-the-evolving-online-ecosystem-barriers-methods-and-future-challenges/>.
- Harrington, Esme, and Mathias Vermeulen. “External Researcher Access to Closed Foundation Models: State of the Field and Options for Improvement.” Mozilla Foundation, August 21, 2024. <https://blog.mozilla.org/wp-content/blogs.dir/278/files/2024/10/External-researcher-access-to-closed-foundation-models.pdf>.
- HateAid. “For Independent Research: Landmark Case against X.” Accessed May 21, 2025. <https://hateaid.org/en/for-independent-research-landmark-case-against-x/#case-travis-brown>.
- Hemphill, Libby, Angela Schöpke-Gonzalez, and Anmol Panda. “Comparative sensitivity of social media data and their acceptable use in research.” *Scientific Data* 9, no. 1 (2022). <https://www.nature.com/articles/s41597-022-01773-w>.
- Hendrix, Justin, and Paul M. Barrett. “The Meta Studies: Nuanced Findings, Corporate Spin, and Media Oversimplification.” *Tech Policy Press*, August 2, 2023. <https://www.techpolicy.press/the-meta-studies-nuanced-findings-corporate-spin-and-media-oversimplification/>.
- Hickey, Cameron, Kaitlyn Dowling, Isabella Navia, and Claire Pershan. “Public Data Access Programs: A First Look.” Mozilla Foundation, 2024. https://assets.mozoprod.net/network/documents/Public_Data_Access_Programs_A_First_Look_Final_cExWfcH.pdf.
- Hofer-Shall, Zach. “Working Directly With the Twitter Data Ecosystem.” Gnip Company Blog, April 10, 2015. <https://web.archive.org/web/20150701021530/https://blog.gnip.com/twitter-data-ecosystem/>.
- Hals, Tom. “Meta investors, Zuckerberg reach settlement to end \$8 billion trial over Facebook privacy violations.” *Reuters*, July 17, 2025.

- <https://www.reuters.com/sustainability/boards-policy-regulation/meta-investors-zuckerberg-reach-settlement-end-8-billion-trial-over-facebook-2025-07-17/>.
- Horwitz, Jeff. “The Facebook Files.” *Wall Street Journal*, September 15, 2021.
<https://www.wsj.com/articles/the-facebook-files-11631713039>.
- — —. “Meta’s AI rules have let bots hold ‘sensual’ chats with kids, offer false medical info.” *Reuters*, August 14, 2025.
<https://www.reuters.com/investigates/special-report/meta-ai-chatbot-guidelines/>.
- Hudson, James M., and Amy Bruckman. “Go Away”: Participant Objections to Being Studied and the Ethics of Chatroom Research.” *The Information Society* 20, no. 2 (2004): 127-139.
<https://www.tandfonline.com/doi/abs/10.1080/01972240490423030>.
- Husovec, Martin. *Principles of the Digital Services Act*. Oxford University Press, 2024.
- IAPP. “Brazilian General Data Protection Law (LGPD, English Translation).” October 2020.
<https://iapp.org/resources/article/brazilian-data-protection-law-lgpd-english-translation/>.
- India. “Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules.” Ministry of Electronics and Information Technology, 2023.
<https://www.meity.gov.in/static/uploads/2024/02/Information-Technology-Intermediary-Guidelines-and-Digital-Media-Ethics-Code-Rules-2021-updated-06.04.2023-.pdf>.
- India Development Review. “IDR Explains | Local Government in India.” January 28, 2020.
<https://idronline.org/idr-explains-local-government-in-india/>.
- Ingram, David. “Judge throws out Elon Musk’s X lawsuit against nonprofit.” *NBC News*, March 25, 2024. <https://www.nbcnews.com/tech/tech-news/elon-musk-x-twitter-ccd-h-lawsuit-rcna144960>.
- Initiative for Digital Public Infrastructure. “New Approaches to Platform Data Research.” University of Massachusetts Amherst, February 9, 2021.
<https://publicinfrastructure.org/2021/02/09/new-approaches-to-platform-data-research/>.
- Institute for Data, Democracy, & Politics. “Tracking Transparency to Enable Research.” The George Washington University. Accessed August 6, 2025.
<https://iddp.gwu.edu/platform-transparency-tools-brussels-effect>.
- International Fact-Checking Network. “Commit to Transparency.” Poynter. Accessed August 6, 2025.
<https://ifcncodeofprinciples.poynter.org/>.
- International Panel on the Information Environment. “Strategies for Improving the Global Information Environment: Results from a Systematic Review and Meta-Analysis.” July 2023.
<https://www.ipie.info/research/sfp2023-1>.
- Italy. “Archivio degli Spot politici.” Accessed May 21, 2025. <https://www.archivispolitici.it/>.
- Iyer, Prithvi. “A Primer on the Meta 2020 U.S. Election Research Studies.” *Tech Policy Press*, May 6, 2025. <https://www.techpolicy.press/a-primer-on-the-meta-2020-us-election-research-studies/>.
- Johnson, Steven L., Samer Faraj, and Srinivas Kudaravalli. “Emergence of Power Laws in Online Communities: The Role of Social Mechanisms and Preferential Attachment.” *MIS Quarterly* 38, no. 3 (September 2014): 795-808. <https://www.jstor.org/stable/26634997>.
- Jolly, Seth, Ryan Bakker, Liesbet Hooghe, et al. “Chapel Hill expert survey trend file, 1999–2019.” *Electoral Studies* 75 (2022).
<https://www.sciencedirect.com/science/article/abs/pii/S0261379421001323>.
- Junkipedia. “Untangle the online conversation across social media to decode what really matters.” Accessed May 22, 2025. <https://www.junkipedia.org/>.

- Keller, Daphne. “Comment to Council of Europe Committee of Ministers on the Draft Recommendation on online safety and empowerment of content creators and users.” August 12, 2025. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5384268.
- — —. “Platform Transparency and the First Amendment.” *Journal of Free Speech Law* (2023). <https://www.journaloffreespeechlaw.org/keller2.pdf>.
- Kelly, Heather. “Venmo privacy settings to change now.” *The Washington Post*, July 28, 2024. <https://www.washingtonpost.com/technology/2024/venmo-privacy-settings-oversharing/>.
- Kentucky. “Complaint.” Scott County Circuit Court, October 8, 2024. <https://files.lbr.cloud/public/2024-10/Kentucky%20Complaint.pdf?VersionId=5QnqTcUw5gMrx1pQbvoEg3Y8nLtKKi8>.
- Kim, Yoonsang, Rachel Nordgren, and Sherry Emery. “The Story of Goldilocks and Three Twitter’s APIs: A Pilot Study on Twitter Data Sources and Disclosure.” *International Journal of Environmental Research and Public Health* 17, no. 3 (2020). <https://pmc.ncbi.nlm.nih.gov/articles/PMC7037495/>.
- Klar, Rebecca. “Facebook Suspends Accounts of NYU Researchers Who’ve Criticized Platform.” *The Hill*, August 4, 2021. <https://thehill.com/policy/technology/566332-facebook-suspends-accounts-of-nyu-researchers-who-criticized-platform/>.
- Kollnig, Konrad, and Nigel Shadbolt. “How Decisions by Apple and Google obstruct App Privacy.” *Technology & Regulation* (2023). <https://techreg.org/article/view/13254>.
- Kreiss, Daniel, Regina G. Lawrence, and Shannon C. McGregor. “In Their Own Words: Political Practitioner Accounts of Candidates, Audiences, Affordances, Genres, and Timing in Strategic Social Media Use.” In *Studying Politics Across Media*, edited by Raymond J. Pingree and Jennifer VanHeerde-Hudson. Routledge, 2019.
- Kuchhal, Dhruv, and Frank Li. “A View into YouTube View Fraud.” *Proceedings of the ACM Web Conference* (2022): 555-563. <https://dl.acm.org/doi/abs/10.1145/3485447.3512216>.
- LaPalombara, Joseph, and Myron Weiner. “The Origin and Development of Political Parties.” In *The Origin and Development of Political Parties*, edited by Joseph LaPalombara and Myron Weiner. Princeton University Press, 1966.
- Lai, Samantha. “Understanding the Information Environment: Insights from the Majority World.” Carnegie Endowment for International Peace, December 12, 2024. <https://carnegieendowment.org/research/2024/12/global-north-majority-world-misinformation?lang=en>.
- Lazovich, Tomo, Luca Belli, Aaron Gonzalez, et al. “Measuring Disparate Outcomes of Content Recommendation Algorithms with Distributional Inequality Metrics.” *Patterns* 3, no. 8 (2022). <https://www.sciencedirect.com/science/article/pii/S2666389922001799>.
- Li, Qinbin, Junyuan Hong, Chulin Xi, et al. “LLM-PBE: Assessing Data Privacy in Large Language Models.” September 6, 2024. <https://arxiv.org/abs/2408.12787>.
- Litt, Eden. “Knock, Knock. Who’s There? The Imagined Audience.” *Journal of Broadcasting & Electronic Media* 56, no. 3 (2021). <https://www.tandfonline.com/doi/full/10.1080/08838151.2012.705195#d1e173>.
- LinkedIn. “Ad Library.” Accessed May 22, 2025. <https://www.linkedin.com/ad-library/home>.

- — —. “Boost a post to reach a wider audience.” Accessed May 22, 2025.
<https://business.linkedin.com/marketing-solutions/cx/22/06/boosting-ff-d>.
- — —. “Data Access for Researchers.” Accessed May 22, 2025.
<https://www.linkedin.com/help/linkedin/answer/a1645616>.
- — —. “Verifications on your LinkedIn profile.” Accessed November 1, 2025.
<https://www.linkedin.com/help/linkedin/answer/a1359065>.
- Lenhart, Anna. “Ethical Use of Pervasive Data for Research: Actions for Academia and Civil Society.” The George Washington University Institute for Data, Democracy, & Politics, July 22, 2025.
<https://iddp.gwu.edu/ethical-use-pervasive-data-research>.
- Levinson-Waldman, Rachel, and Naz Balkam. “The Government’s Growing Trove of Social Media Data.” Brennan Center for Justice, July 21, 2025.
<https://www.brennancenter.org/our-work/research-reports/governments-growing-trove-social-media-data>.
- Lomas, Natasha, and Roman Dillet. “Terms and Conditions Are the Biggest Lie of Our Industry.” *TechCrunch*, August 21, 2015. <https://techcrunch.com/2015/08/21/agree-to-disagree/>.
- Lorenz-Spreen, Philipp, Lisa Oswald, Stephan Lewandowsky, and Ralph Hertwig. “A Systematic Review of Worldwide Causal and Correlational Evidence on Digital Media and Democracy.” *Nature Human Behaviour* 7, no. 1 (2022): 74-101. <https://doi.org/10.1038/s41562-022-01460-1>.
- Lukito, Josephine, J. Nathan Matias, and Sarah Gilbert. “Enabling Independent Research without Unleashing Ethics Disasters.” *Tech Policy Press*, May 10, 2023.
<https://www.techpolicy.press/enabling-independent-research-without-unleashing-ethics-disasters/>.
- Lurie, Emma, and Frances Schroeder. “TikTok just announced the data it’s willing to share. What’s missing?” Stanford Cyber Policy Center, February 24, 2023.
<https://cyber.fsi.stanford.edu/news/tiktok-just-announced-data-its-willing-share-whats-missing>.
- Lutscher, Philipp M., and Neil Ketchley. “Online repression and tactical evasion: evidence from the 2020 Day of Anger protests in Egypt.” *Democratization* 30, no. 2 (2023): 325-345.
<https://www.tandfonline.com/doi/full/10.1080/13510347.2022.2140798>.
- Malik, Ray. “Are Social Media Settings Intentionally Confusing? Here’s the Truth.” *How-To Geek*, March 22, 2025.
<https://www.howtogeek.com/are-social-media-settings-intentionally-confusing-heres-the-truth/>.
- Marenco, Andre. “Local Government, Brazil - Qualigov.” In *Global Encyclopedia of Public Administration, Public Policy, and Governance*, edited by Ali Farazmand. Springer Cham, June 26, 2018.
- Martin, Kirsten, and Helen Nissenbaum. “Measuring Privacy: An Empirical Test Using Context to Expose Confounding Variables.” *Columbia Science & Technology Law Review* 18 (2016).
<https://journals.library.columbia.edu/index.php/stlr/article/view/4015/1776>.
- Marwick, Alice E., and danah boyd. “Networked privacy: How teenagers negotiate context in social media.” *New Media & Society* 16, no. 7 (2014): 1051-67.
<https://journals.sagepub.com/doi/10.1177/1461444814543995>.
- — —. “I tweet honestly, I tweet passionately: Twitter Users, context collapse, and the imagined audience.” *New Media & Society* 13, no. 1 (2010).
<https://journals.sagepub.com/doi/10.1177/1461444810365313>.

- Massachusetts. “Complaint and Jury Demand.” Suffolk Superior Court, February 3, 2025.
<https://www.mass.gov/doc/tiktok-complaint-unredacted/download>.
- Mathur, Prajvi. “Idaho shooter’s chilling Instagram story goes viral; Here’s what he posted hours before he ambushed firefighters.” *WION*, July 1, 2025.
<https://www.wionews.com/world/idaho-shooter-s-chilling-instagram-story-goes-viral-here-s-what-he-posted-hours-before-he-ambushed-firefighters-1751364637542>.
- Matias, J. Nathan. “Why We Need Industry-Independent Research on Tech & Society.” The Citizens and Technology Lab, January 7, 2020.
<https://citizensandtech.org/2020/01/industry-independent-research/>.
- Matias, J. Nathan, Susan Benesch, Rebekah Tromble, et al. “Manifesto: The Coalition for Independent Technology Research.” Coalition for Independent Technology Research, October 12, 2022.
<https://independenttechresearch.org/manifesto-the-coalition-for-independent-technology-research/>.
- McGrady, Ryan, Kevin Zheng, Rebecca Curran, Jason Baumgartner, and Ethan Zuckerman. “Dialing for Videos: A Random Sample of YouTube.” *Journal of Quantitative Description: Digital Media* 3 (2023). <https://journalqd.org/article/view/4066>.
- Media Authority of North Rhine-Westphalia. “Disclosing Sponsored Influencer Content.” Accessed September 18, 2025.
<https://www.medienanstalt-nrw.de/about-us/topics/sponsored-influencer-content.html>.
- Meltwater. “The Complete Boolean Library.” 2025.
<https://help.meltwater.com/en/articles/4064570-the-complete-boolean-library>.
- — —. “Social Listening & Analytics.” Accessed May 22, 2025.
<https://www.meltwater.com/en/suite/social-listening-analytics>.
- Merz, Nicolas, Sven Regel, and Jirka Lewandowski. “The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis.” *Research & Politics* 3, no. 2 (2016): 2053168016643346. <https://journals.sagepub.com/doi/10.1177/2053168016643346>.
- Meta. “About journalist registration.” Accessed May 22, 2025.
<https://www.facebook.com/business/help/620369758565492?id=1843027572514562>.
- — —. “About Page likes ads.” Accessed August 15, 2025.
<https://www.facebook.com/business/help/209213872548401?id=2223788217707136>.
- — —. “Ad Library.” Accessed May 22, 2025.
https://www.facebook.com/ads/library/?active_status=active&ad_type=political_and_issue_ads&country=US&is_targeted_country=false&media_type=all.
- — —. “Boost a post from your Facebook page.” Accessed August 15, 2025.
<https://www.facebook.com/business/help/347839548598012?id=352109282177656>.
- — —. “Community Standards.” Accessed August 15, 2025.
<https://transparency.meta.com/policies/community-standards/>.
- — —. “Data dictionary.” Accessed May 22, 2025.
<https://developers.facebook.com/docs/content-library-and-api/appendix/data-dictionary>.
- — —. “Facebook Page.” Accessed May 22, 2025.
https://www.facebook.com/business/tools/facebook-pages?content_id=YAvvvDmp3OKQagl.
- — —. “Meta Content Library and API.” Accessed May 22, 2025.
<https://transparency.meta.com/es-la/researchtools/meta-content-library/>.

- — —. “Partner Monetization Policies.” Accessed May 22, 2025.
<https://www.facebook.com/business/help/169845596919485?id=2520940424820218>.
 - — —. “Public Figures.” Accessed May 22, 2025.
<https://about.meta.com/actions/safety/audiences/journalists-activists-public-figures/public-figure>
[S](#).
 - — —. “Public figures safety.” Accessed May 22, 2025.
<https://about.meta.com/actions/safety/audiences/journalists-activists-public-figures/public-figure>
[S](#).
 - — —. “Research partnership to understand Facebook and Instagram’s role in the U.S. 2020 election.” Accessed May 20, 2025. <https://research.facebook.com/2020-election-research/>.
 - — —. “Why search engines might index public Instagram photos and videos.” Accessed August 6, 2025. https://help.instagram.com/147542625391305/?helpref=uf_share.
 - — —. “Widely Viewed Content Report: What People See on Facebook.” 2025.
<https://transparency.meta.com/data/widely-viewed-content-report/#widely-viewed-pages>.
- Metcalf, Jacob, and Kate Crawford. “Where are the human subjects in Big Data research? The emerging ethics divide.” *Big Data & Society* (2016).
<https://journals.sagepub.com/doi/full/10.1177/2053951716650211>.
- Meyer, Michelle N., John Basl, David Choffnes, Christo Wilson, and David MJ Lazer. “Enhancing the ethics of user-sourced online data collection and sharing.” *Nature Computational Science* 3, no. 8 (2023): 660-664. <https://pubmed.ncbi.nlm.nih.gov/38177316/>.
- Microsoft. “Bing Qualified Researcher Program.” Accessed May 23, 2025.
<https://support.microsoft.com/en-us/topic/bing-qualified-researcher-program-b1c5a4b6-3ad6-4c8c-963d-a395c7766c85>.
- Miles, Chris. “From Dashboards to Data Acquisition: Charting the Social Media Monitoring Market.” The George Washington University Institute for Data, Democracy, & Politics, September 2025.
https://iddp.gwu.edu/sites/g/files/zaxdzs5791/files/2025-09/iddp_smm_report_sep_2025.pdf.
- Mimizuka, Kayo, Megan A Brown, Kai-Cheng Yang, Josephine Lukito. “Post-Post-API Age: Studying Digital Platforms in Scant Data Access Times.” May 15, 2025.
<https://www.arxiv.org/abs/2505.09877>.
- Mondal, Mainack, Günce Su Yilmaz, Noah Hirsch, et al. “Moving Beyond Set-It-And-Forget-It Privacy Settings on Social Media.” *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (2019). <https://www.cs.uic.edu/~elena/pubs/mondal-ccs19.pdf>.
- Mosleh, Mohsen, Gordon Pennycook, and David G. Rand. “Field Experiments on Social Media.” *Current Directions in Psychological Science* 31, no. 1 (2021).
<https://doi.org/10.1177/09637214211054761>.
- Mozilla Foundation. “About YouTube Regrets.” Accessed May 23, 2025.
<https://foundation.mozilla.org/en/youtube/regretsreporter/about/>.
- — —. “Full Disclosure: Stress testing tech platforms’ ad repositories.” 2024.
<https://www.mozillafoundation.org/en/research/library/full-disclosure-stress-testing-tech-platforms-ad-repositories/>
- — —. “Open Letter to Meta.” 2024.
<https://www.mozillafoundation.org/en/campaigns/open-letter-to-meta-support-crowdtangle-through-2024-and-maintain-crowdtangle-approach/>.

- Mündges, Stephan, and Kirsty Park. "But Did They Really? Platforms' Compliance with the Code of Practice on Disinformation in review." *Internet Policy Review* 13, no. 3 (2024).
<https://doi.org/10.14763/2024.3.1786>.
- Munger, Kevin, Matthew Hindman, James Bisbee, Omer Yalcin, and Joseph Phillips. "Pressing Play on Politics: Quantitative Description of YouTube." *Journal of Quantitative Description: Digital Media* 5 (2025). <https://journalqd.org/article/view/8674>.
- Myanmar. "Pyidaungsu Hluttaw." Accessed August 6, 2025.
<https://web.archive.org/web/20210204051136/https://pyidaungsu.hluttaw.mm/>.
- Nayak, Chaya. "New privacy-protected Facebook data for independent research on social media's impact on democracy." Meta, February 12, 2020.
<https://research.facebook.com/blog/2020/2/new-privacy-protected-facebook-data-for-independent-research-on-social-medias-impact-on-democracy/>.
- Narayanan, Arvind. "Understanding Social Media Recommendation Algorithms." Knight First Amendment Institute at Columbia University, March 9, 2023.
<https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms>.
- National Telecommunications and Information Administration. "Ethical Research with Online/Pervasive Data." December 17, 2024.
<https://www.ntia.gov/events-and-meetings/ethical-research-online/pervasive-data>.
- Nenadić, Iva, Elda Brogi, Konrad Bleyer-Simon, and Urbano Reviglio. "Structural Indicators of the Code of Practice on Disinformation: The 2nd EDMO report." European Digital Media Observatory March 2024. https://edmo.eu/wp-content/uploads/2024/03/SIs_-_2nd-EDMO-report.pdf.
- Neumann, Arthur, Elizabeth Sims, and Robert Conrad. "An Examination of the Impact of Stylometry, Artificial Intelligence/Machine Learning (AI/ML) on Privacy in Social Media." Carnegie Mellon University, May 11, 2023.
https://kilthub.cmu.edu/articles/report/An_Examination_of_the_Impact_of_Stylometry_Artificial_Intelligence_Machine_Learning_AI_ML_on_Privacy_in_Social_Media/22766711?file=40451135.
- New York Times. "Ethical Journalism." Accessed August 6, 2025.
<https://www.nytimes.com/editorial-standards/ethical-journalism.html>.
- NewsWhip. "NewsWhip, your only CrowdTangle alternative." Accessed August 15, 2025.
<https://www.newswhip.com/crowdtangle-alternative-data-real-time/>.
- — —. "Stay ahead of emerging issues, identify critical moments, and track the spread of misinformation with Newswhip." Accessed May 23, 2025.
<https://www.newswhip.com/nonprofit-solutions/>.
- Nigeria. "Code of Practice for Interactive Computer Service Platforms/Internet Intermediaries." National Information Technology Development Agency, October 2022.
<https://nitda.gov.ng/wp-content/uploads/2022/10/APPROVED-NITDA-CODE-OF-PRACTICE-FOR-INTERACTIVE-COMPUTER-SERVICE-PLATFORMS-INTERNET-INTERMEDIARIES-2022-002.pdf>.
- Nissenbaum, Helen. "Contextual integrity up and down the data food chain." *Theoretical Inquiries in Law* 20, no. 1 (2019): 221-256. <https://philpapers.org/rec/NISCIU>.
- — —. "Privacy as Contextual Integrity." *Washington Law Review* 79, no 1 (2004).
<https://digitalcommons.law.uw.edu/wlr/vol79/iss1/10/>.

- Norris, Donald F., and Christopher G. Reddick. "Local e-government in the United States: Transformation or incremental change?" *Public Administration Review* 73, no. 1 (2013): 165-175. <https://www.jstor.org/stable/23355450>.
- Norton, Sean, and Jacob N. Shapiro. "How to Better Study—and Then Improve—Today's Corrupted Information Environment." *Bulletin of the Atomic Scientists*, March 1, 2024. <https://thebulletin.org/2024/03/how-to-better-study-and-then-improve-todays-corrupted-information-environment/>.
- Nulty, Jake. "Top 5 AI Training Data Providers of 2025." Bright Data, 2025. <https://brightdata.com/blog/ai/best-ai-training-data-providers>.
- Obar, Jonathan A., and Anne Oeldorf-Hirsch. "The Biggest Lie on the Internet: Ignoring the Privacy Policies and Terms of Service Policies of Social Networking Services." *Information, Communication & Society* 23, no. 1 (2020): 128-47. <https://doi.org/10.1080/1369118x.2018.1486870>.
- Observatory on Social Media. "OSoMe Awesome Speakers – Deen Freelon (University of Pennsylvania)." February 27, 2025. <https://www.youtube.com/watch?v=xELvlp90uiQ>.
- Ofcom. "Call for Evidence: Researchers' Access to Information from Regulated Online Services." October 28, 2024. <https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/call-for-evidence-researchers-access-to-information-from-regulated-online-services>.
- Organisation for Economic Co-operation and Development. "AI principles." Accessed August 7, 2025. <https://www.oecd.org/en/topics/sub-issues/ai-principles.html>.
- Pater, Jassic, Casey Fiesler, and Michael Zimmer. "No Humans Here: Ethical Speculation on Public Data, Unintended Consequences, and the Limits of Institutional Review." *Proceedings of the ACM on Human-Computer Interaction* 6, no. 38 (2022): 1-13. <https://dl.acm.org/doi/10.1145/3492857>.
- Paul, Katie, and Anna Tong. "Inside Big Tech's Underground Race to Buy AI Training Data." *Reuters*, April 5, 2024. <https://www.reuters.com/technology/inside-big-techs-underground-race-buy-ai-training-data-2024-04-05/>.
- Perez, Sarah. "Facebook is asking to use Meta AI on photos in your camera roll you haven't yet shared." *TechCrunch*, June 27, 2025. <https://techcrunch.com/2025/06/27/facebook-is-asking-to-use-meta-ai-on-photos-in-your-camera-roll-you-havent-yet-shared/>.
- — —. "Meta drops lawsuit against web-scraping firm Bright Data that sold millions of Instagram records." *TechCrunch*, February 26, 2024. <https://techcrunch.com/2024/02/26/meta-drops-lawsuit-against-web-scraping-firm-bright-data-that-sold-millions-of-instagram-records/>.
- PERVADE. "The Project." Accessed August 6, 2025. <https://pervade.umd.edu/about/general/>.
- Petrovic, Sasa, Miles Osborne, and Victor Lavrenko. "I Wish I Didn't Say That! Analyzing and Predicting Deleted Messages in Twitter." May 13, 2013. <https://arxiv.org/abs/1305.3107>.
- Pierce, David. "'Views' are lies." *The Verge*, April 5, 2025. <https://www.theverge.com/social/639811/view-counts-tiktok-instagram-x-youtube-lies>.
- Pinterest. "Ads Repository." Accessed May 23, 2025. <https://ads.pinterest.com/ads-repository/>.

- — —. “Get a business account.” Accessed August 15, 2025.
<https://help.pinterest.com/en/business/article/get-a-business-account>.
- Potash, Peter J., Eric B. Bell, and Joshua J. Harrison. “Using Topic Modeling and Text Embeddings to Predict Deleted Tweets.” Pacific Northwest National Lab, February 28, 2016.
<https://www.osti.gov/biblio/1334882>.
- Poynter. “Empowering fact-checkers worldwide.” Accessed May 23, 2025.
<https://www.poynter.org/ifcn/>.
- Privacy International. “Large language models and data protection.” August 14, 2024.
<https://privacyinternational.org/explainer/5353/large-language-models-and-data-protection>.
- Public Interest Tech Lab. “About.” Accessed May 23, 2025. <https://techlab.org/about.html>.
- — —. “Fbarchive.” Accessed May 23, 2025. <https://fbarchive.org/about>.
- — —. “Redaction Guidelines.” Accessed August 6, 2025. <https://fbarchive.org/redaction>.
- Quigley, Paul. “Six key features for newsrooms seeking a CrowdTangle replacement.” NewsWhip, May 16, 2024.
<https://www.newswhip.com/2024/05/six-key-features-for-newsrooms-seeking-a-crowdtangle-replacement/>.
- Quin, Federico, Danny Weyns, Matthias Galster, and Camila Costa Silva. “A/B testing: A systematic literature review.” *Journal of Systems and Software* 211 (2024): 112011.
<https://www.sciencedirect.com/science/article/pii/S0164121224000542>.
- Radtke, Tristan. “Disclosure Requirements for Influencer Marketing in the U.S. and Germany.” *New York University Journal of Intellectual Property and Entertainment Law* 12, no. 1 (2023).
<https://jipel.law.nyu.edu/disclosure-requirements-for-influencer-marketing-in-the-u-s-and-germany/>.
- Reddit. “Addressing the community about changes to our API.” 2023.
https://www.reddit.com/r/reddit/comments/145bram/addressing_the_community_about_changes_to_our_api/.
- — —. “Creating a Healthy Ecosystem for Reddit Data and Reddit Data API Access.” *upvoted*, April 18, 2023. <https://redditinc.com/blog/2023apiupdates>.
- Reuters. “Reddit to update web standard to block automated website scraping.” *Reuters*, June 25, 2024.
<https://www.reuters.com/technology/reddit-update-web-standard-block-automated-website-scraping-2024-06-25/>.
- Reynolds, Jen Rosiere. “Findings from a Survey of Researcher Pain Points.” The Accelerator, January 31, 2025. <https://researchaccelerator.org/findings-from-a-survey-of-researcher-pain-points/>.
- Rincón, Daniela Alvarado, Ognjan Denkovski, Salvatore Roman, and Martin Degeling. “Unpacking TikTok’s Data Access Illusion.” *Tech Policy Press*, June 12, 2025.
<https://www.techpolicy.press/unpacking-tiktoks-data-access-illusion/>.
- Roth, Yoel. “The Twitter Moderation Research Consortium is now open to researchers.” Twitter, September 22, 2022.
https://blog.x.com/en_us/topics/company/2022/twitter-moderation-research-consortium-open-researchers.
- Roth, Yoel, and Vijaya Gadde. “Expanding access beyond information operations.” Twitter, December 2, 2021.

https://web.archive.org/web/20231208001243/https://blog.twitter.com/en_us/topics/company/2021/-expanding-access-beyond-information-operations-

Russell, Annelise. *Tweeting Is Leading: How Senators Communicate and Represent in the Age of Twitter*. Oxford University Press, 2021.

Sabha, Rajya. “Members of Parliament.” Accessed May 20, 2025.

<https://www.india.gov.in/my-government/whos-who/members-parliament>.

Sanderson, Zeve, and Lama Mohammed. “A Multi-Stakeholder Approach for Leveraging Data Portability to Support Research on the Digital Information Environment.” *Journal of Online Trust and Safety* 2, no. 4 (September 2024). <https://doi.org/10.54501/jots.v2i4.215>.

Sarkar, Torsha, Gurshabad Grover, Raghav Ahooja, Pallavi Bedi, and Divyank Katira. “On the Legality and Constitutionality of the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021.” The Centre for Internet & Society, 2021.

<https://cis-india.org/internet-governance/legal-constitutionality-il-rules-digital-media-2021>.

Schafer, Joseph S., and Kate Starbird. “Post-Spotlight Posts: The Impact of Sudden Social Media Attention on Account Behavior.” *Conference on Computer Supported Cooperative Work and Social Computing* (2023). <https://dl.acm.org/doi/10.1145/3584931.3606984>.

Schaffner, Brennan, Neha Lingareddy, and Marshini Chetty. “Understanding Account Deletion and Relevant Dark Patterns on Social Media.” *Proceedings of the ACM on Human-Computer Interaction* (2022). <https://dl.acm.org/doi/10.1145/3555142>.

Schmidt, Jan-Hinrik, Lisa Merten, and Felix Victor Münich. “Die Datenbank Öffentlicher Sprecher.” March 28, 2023. <https://osf.io/sk6t5/>.

Scott, Mark. “Survey: New Laws Mandate Access to Social Media Data, but Obstacles Remain.” *Tech Policy Press*, April 8, 2024.

<https://www.techpolicy.press/survey-new-laws-mandate-access-to-social-media-data-but-obstacles-remain/>.

Shiffman, Naomi. “Tools for Platform Research: Lessons from the Medical Research Industry.” *Tech Policy Press*, April 26, 2023.

<https://www.techpolicy.press/tools-for-platform-research-lessons-from-the-medical-research-industry/>.

Shiffman, Naomi, and Brandon Silverman. “The Case for Transparency: How Social Media Platform Data Access Leads to Real-World Change.” The George Washington University Institute for Data, Democracy, & Politics, May 7, 2025.

https://iddp.gwu.edu/sites/g/files/zaxdzs5791/files/2025-05/case_for_transparency_shiffman_silverman.pdf.

— — —. “CrowdTangle Opens Public Application for Academics.” Meta, July 31, 2020.

<https://research.facebook.com/blog/2020/7/crowdtangle-opens-public-application-for-academics/>.

Shilton, Katie, Emanuel Moss, Sarah A. Gilbert, et al. “Excavating awareness and power in data science: A manifesto for trustworthy pervasive data research.” *Big Data & Society* (2021).

<https://journals.sagepub.com/doi/full/10.1177/20539517211040759>.

Silberling, Amanda. “Spotify responds to backlash over public podcast play counts.” *TechCrunch*, May 16, 2025.

<https://techcrunch.com/2025/05/16/spotify-responds-to-creator-backlash-at-public-podcast-play-counts/>.

Silverman, Craig. “How to tell if a social media account is monetized.” Indicator, October 7, 2025.

<https://indicator.media/p/how-to-tell-if-a-social-media-account-is-monetized-meta-tiktok-youtube-x>.

Singapore. “Broadcasting Act 1994: Code of Practice for Online Safety – Social Media Services.”

Infocomm Media Development Authority, December 2021.

<https://www.imda.gov.sg/-/media/imda/files/regulations-and-licensing/regulations/codes-of-practice/codes-of-practice-media/code-of-practice-for-online-safety-social-media-services.pdf>.

Snap. “Ads Gallery.” Accessed June 10, 2025. <https://adsgallery.snap.com/>.

— — —. “How Do I Change My Privacy Settings on Snapchat?” Accessed June 10, 2025.

<https://help.snapchat.com/hc/en-us/articles/7012343074580-How-do-I-change-my-privacy-settings-on-Snapchat>.

— — —. “Researcher Data Access Instructions.” Accessed June 10, 2025.

<https://values.snap.com/privacy/transparency/researcher-access>.

SOMAR. “About.” University of Michigan Institute for Search Research. Accessed August 6, 2025.

<https://socialmediaarchive.org/pages/?page=About&ln=en>.

South Korea. “검색결과.” Accessed June 10, 2025. <https://nl.go.kr/oasis/contents/O3010000.do>.

Spotify. “Spotify for Artists.” Accessed June 10, 2025.

<https://support.spotify.com/us/article/spotify-for-artists/?ref=related>.

— — —. “Spotify for Creators’ Monetization Options.” Accessed June 10, 2025.

<https://support.spotify.com/us/creators/article/spotify-for-creators-monetization-options/>.

Steel, Benjamin, Miriam Schirmer, Derek Ruths, and Juergen Pfeffer. “Just Another Hour on TikTok: Reverse-engineering unique identifiers to obtain a complete slice of TikTok.” April 17, 2025.

<https://arxiv.org/abs/2504.13279>.

Starbird, Kate, Renée DiResta, and Matt DeButts. “Influence and Improvisation: Participatory Disinformation during the 2020 US election.” *Social Media + Society* 9, no. 2 (2023).

<https://journals.sagepub.com/doi/10.1177/20563051231177943>.

Supreme Court of the United States. “Katz v. United States.” 1967.

<https://tile.loc.gov/storage-services/service/ll/usrep/usrep389/usrep389347/usrep389347.pdf>.

— — —. “Lindke v. Freed.” 2023. https://www.supremecourt.gov/opinions/23pdf/22-611_ap6c.pdf.

Švelch, Jan. “Redefining screenshots: Toward critical literacy of screen capture practices.”

Convergence 27, no. 2 (2021): 554-569.

<https://journals.sagepub.com/doi/abs/10.1177/1354856520950184>.

Tan, Chenhao, Lillian Lee, and Bo Pang. “The effect of wording on message propagation: Topic-and author-controlled natural experiments on Twitter.” *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* 1 (2014): 175-185. <https://aclanthology.org/P14-1017/>.

Tavishi, Shobhit S. “Platform Transparency under the EU’s Digital Services Act: Opportunity and Challenges for the Global South.” NLU Delhi Centre for Communication Governance, June 2025.

<https://ccgdelhi.s3.ap-south-1.amazonaws.com/uploads/ccg-nlud-platform-transparency-eu-dsa-gschallenges-738.pdf>.

Thorburn, Luke, Jonathan Stray, and Priyanjana Bengani. “Experiments Are the Best Kind of Transparency.” *Tech Policy Press*, June 3, 2024.

<https://www.techpolicy.press/experiments-are-the-best-kind-of-transparency/>.

TikTok. “Codebook.” Accessed June 11, 2025.

<https://developers.tiktok.com/doc/research-api-codebook>.

— — —. “Community Guidelines.” Accessed August 15, 2025.

<https://www.tiktok.com/community-guidelines/en>.

— — —. “Find Ads on TikTok.” Accessed June 22, 2025. <https://library.tiktok.com/ads/>.

— — —. “Find and Reach Your Target Audience.” Accessed June 11, 2025.

<https://ads.tiktok.com/business/en-US/goals/expand-reach>.

— — —. “Government, Politician, and Political Party Accounts.” Accessed May 1, 2025.

<https://support.tiktok.com/en/using-tiktok/growing-your-audience/government-politician-and-political-party-accounts>.

— — —. “Looking for trends? You’ve come to the right place.” Accessed June 22, 2025.

<https://ads.tiktok.com/business/creativecenter/trends/home/pc/en>.

— — —. “Research Tools.” Accessed October 23, 2025.

<https://developers.tiktok.com/products/research-api/>.

— — —. “TikTok for Developers.” Accessed June 22, 2025. <https://developers.tiktok.com/>.

— — —. “Use Promote to grow your TikTok audience.” Accessed August 15, 2025.

<https://support.tiktok.com/en/using-tiktok/growing-your-audience/use-promote-to-grow-your-tiktok-audience>.

— — —. “Verified accounts on TikTok.” Accessed June 22, 2025.

<https://support.tiktok.com/en/using-tiktok/growing-your-audience/how-to-tell-if-an-account-is-verified-on-tiktok>.

Timberg, Craig. “Facebook made big mistakes in data it provided to researchers, undermining academic work.” *The Washington Post*, September 10, 2021.

<https://www.washingtonpost.com/technology/2021/09/10/facebook-error-data-social-scientists/>.

Tinati, Ramine, Aastha Madaan, and Wendy Hall. “Instacan: Examining Deleted Content on Instagram.” *Proceedings of the 2017 ACM on Web Science Conference* (2017).

<https://dl.acm.org/doi/abs/10.1145/3091478.3091516>.

Trefis Team. “What is Facebook’s Revenue Breakdown?” March 28, 2019.

<https://www.nasdaq.com/articles/what-facebooks-revenue-breakdown-2019-03-28-0>.

Tromble, Rebekah. “Where Have All the Data Gone? A Critical Reflection on Academic Digital Research in the Post-API Age.” *Social Media + Society* 7, no. 1 (2021).

<https://doi.org/10.1177/2056305121988929>.

Trust & Safety Professional Association. “History of Transparency Reports.” Accessed August 6, 2025.

<https://www.tspa.org/curriculum/ts-fundamentals/transparency-report/history-transparency-reports/>.

Tucker, Joshua A., Lama Mohammed, and Sarah Graham, “The Case for Open Data Access to Aid Tech Regulation.” The Brookings Institution, December 17, 2024.

<https://www.brookings.edu/articles/the-case-for-open-data-access-to-aid-tech-regulation/>.

Twitter. “Twitter data for academic research.” Accessed March 7, 2020.

<https://web.archive.org/web/20200307135211/https://developer.twitter.com/en/use-cases/academic-researchers>.

UNESCO. “Guidelines for the Governance of Digital Platforms.” 2023.

<https://www.unesco.org/en/internet-trust/guidelines>.

United Arab Emirates. “حسابات التواصل الاجتماعي لحكومة الإمارات.” Accessed June 22, 2025.

<https://u.ae/ar-ae/about-the-uae/digital-uae/digital-inclusion/uae-government-social-media-accounts>.

United Kingdom. “Data (Use and Access) Act 2025.” Accessed June 22, 2025.

<https://bills.parliament.uk/bills/3825>.

— — —. “Online Safety Act 2023.” Accessed June 22, 2025.

<https://www.legislation.gov.uk/ukpga/2023/50>.

United Nations. “Global Digital Compact.” 2025. <https://www.un.org/global-digital-compact/en>.

United States. “16 CFR § 233.3. - Advertising retail prices which have been established or suggested by manufacturers (or other nonretail distributors).” Legal Information Institute. Accessed August 15, 2025. <https://www.law.cornell.edu/cfr/text/16/233.3#>.

— — —. “Computer Fraud and Abuse Act.” 1986.

[https://uscode.house.gov/view.xhtml?req=\(title:18%20section:1030%20edition:prelim](https://uscode.house.gov/view.xhtml?req=(title:18%20section:1030%20edition:prelim)

— — —. “H.R.6796 - Digital Services Oversight and Safety Act.” February 18, 2022.

<https://www.congress.gov/bill/117th-congress/house-bill/6796/text>.

— — —. “H.R.8152 – American Data Privacy and Protection Act.” June 21, 2022.

<https://www.congress.gov/bill/117th-congress/house-bill/8152>.

— — —. “S.5339 - Platform Accountability and Transparency Act.” December 21, 2022.

<https://www.congress.gov/bill/117th-congress/senate-bill/5339>.

United States Court of Appeals for the Ninth Circuit. “Combined Opinion.” September 4, 2024.

https://www.courtlistener.com/opinion/10104396/x-corp-v-bonta/?type=o&type=o&q=X+v.+Bonta&order_by=score+desc.

— — —. “Opinion.” April 22, 2022.

<https://cdn.ca9.uscourts.gov/datastore/opinions/2022/04/18/17-16783.pdf>.

United States District Court for the Northern District of California. “Demand for Jury Trial.” July 31, 2023.

https://www.pacermonitor.com/view/NV3GMHQ/X_Corp_a_Nevada_Corporation_v_Center_for_Countering_Digital_Candace-23-03836_0001.0.pdf?mcid=tGE3TEOA.

— — —. “Order Denying Meta’s Motion for Partial Summary Judgment; and Granting Bright Data’s Motion for Summary Judgment.” January 23, 2024.

<https://www.courthousenews.com/wp-content/uploads/2024/01/meta-platforms-v-bright-data-ruling-motion-for-summary-judgment.pdf>.

United States Executive Office of the President. “Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.” Federal Register, October 30, 2023.

<https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence>.

- Vertesi, Janet, and J. Nathan Matias. “Divesting from Big Tech: Alternative Possibilities for Research and Futuring in Social Computing.” *Conference on Computer Supported Cooperative Work and Social Computing* (2023). <https://doi.org/10.1145/3584931.3608436>.
- Vogus, Caitlin. “Defending Data: Privacy Protection, Independent Researchers, and Access to Social Media Data in the US and EU.” Center for Democracy and Technology, January 25, 2023. <https://cdt.org/insights/report-defending-data-privacy-protection-independent-researchers-and-access-to-social-media-data-in-the-us-and-eu/>.
- Vitak, Jessica, Yuting Liao, Anouk Mols, et al. “When do data collection and use become a matter of concern? A cross-cultural comparison of US and Dutch privacy attitudes.” *International Journal of Communication* 17 (2022). <https://ijoc.org/index.php/ijoc/article/view/19391>.
- Waem, Heidi. “Who’s who under the DMA, DSA, DGA and Data Act?” DLA Piper, March 12, 2024. <https://www.dlapiper.com/en/insights/publications/2024/03/whos-who-under-the-dma-dsa-dga-and-data-act>.
- Wagner, Michael W. “Independence by permission.” *Science* 381, no. 6656 (2023): 388–91. <https://doi.org/10.1126/science.adi2430>.
- Waldman, Ari Ezra. “Cognitive biases, dark patterns, and the ‘privacy paradox’.” *Current Opinion in Psychology* 31, no. 1 (2020): 105–9. <https://doi.org/10.1016/j.copsyc.2019.08.025>.
- Wang, Yang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. “‘I regretted the minute i pressed share’: a qualitative study of regrets on Facebook.” *Proceedings of the Seventh Symposium on Usable Privacy and Security* (2011). <https://dl.acm.org/doi/10.1145/2078827.2078841>.
- Wanless, Alicia, Samantha Lai, and John Hicks. “Assessing National Information Ecosystems.” Carnegie Endowment for International Peace, February 11, 2025. <https://carnegieendowment.org/research/2025/02/assessing-national-information-ecosystems?lang=en>.
- Welch, Chris. “Facebook may have knowingly inflated its video metrics for over a year.” *The Verge*, October 17, 2018. <https://www.theverge.com/2018/10/17/17989712/facebook-inaccurate-video-metrics-inflation-lawsuit>.
- WhatToFix. “Social Media Monetization 2025.” February 22, 2025. <https://www.whattofix.tech/publications/monetization2025/>.
- The White House. “America’s AI Action Plan.” July 2025. <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf>.
- Whittaker, Meredith. “The Steep Cost of Capture.” *Interactions* 28, no. 6 (2021): 50–55. <https://doi.org/10.1145/3488666>.
- Williams, Matthew L., Pete Burnap, Luke Sloan, Curtis Jessop, and Hayley Lepps. “Users’ Views of Ethics in Social Media Research: Informed Consent, Anonymity, and Harm.” In *The Ethics of Online Research*, edited by Kandy Woodfield and Ron Iphofen. Emerald Publishing Limited, 2017.
- Working Group on Platform-to-Researcher Data Access. “Report of the European Digital Media Observatory’s Working Group on Platform-to-Researcher Data Access.” European Digital Media Observatory, May 31, 2022. <https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf>.

- Worldometer. “Countries in the world by population (2025).” Accessed June 23, 2025.
<https://www.worldometers.info/world-population/population-by-country/>.
- X. “About public-interest exceptions on X.” Accessed June 23, 2025.
<https://help.x.com/en/rules-and-policies/public-interest>.
- — —. “About view counts.” Accessed June 23, 2025. <https://help.x.com/en/using-x/view-counts>.
- — —. “Defining public interest on Twitter.” Last modified October 15, 2019.
https://blog.x.com/en_us/topics/company/2019/publicinterest.
- — —. “Downloading data.” Last modified May 20, 2025.
<https://communitynotes.x.com/guide/en/under-the-hood/download-data>.
- — —. “Professional Accounts.” Accessed June 23, 2025.
<https://business.x.com/en/help/account-setup/professional-accounts#home-faq>.
- — —. “Promoted Tweets.” Accessed August 15, 2025.
<https://developer.x.com/en/docs/x-ads-api/campaign-management/api-reference/promoted-tweets>.
- — —. “X Verification requirements.” Accessed August 15, 2025.
<https://help.x.com/en/managing-your-account/about-x-verified-accounts>.
- — —. “X API v2 data dictionary.” Accessed June 23, 2025.
<https://docs.x.com/x-api/fundamentals/data-dictionary>.
- — —. “The X Rules.” Accessed August 15, 2025. <https://help.x.com/en/rules-and-policies/x-rules>.
- Yang, Yunkang, Ramesh Paudel, Jordan McShan, Matthew Hindman, H. Howie Huang, and David Broniatowski. “Coordinated link sharing on Facebook.” *Scientific Reports* 15 (2025).
<https://pubmed.ncbi.nlm.nih.gov/40325023/>.
- YouTube. “Election information panels.” Accessed May 20, 2025.
<https://support.google.com/youtube/answer/10047535?hl=en>.
- — —. “Implementation Guide.” Accessed June 23, 2025.
<https://developers.google.com/youtube/v3/guides/implementation>.
- — —. “Political content.” Accessed June 23, 2025.
<https://support.google.com/adspolicy/answer/6014595?sjid=6325069900383477347-NA>
- — —. “Privacy.” Accessed June 23, 2025.
<https://www.youtube.com/howyoutubeworks/user-settings/privacy/>.
- — —. “Topical context in information panel.” Accessed June 23, 2025.
<https://support.google.com/youtube/answer/9004474?hl=en>
- — —. “Verification badges on channels.” Accessed September 18, 2025.
<https://support.google.com/youtube/answer/3046484?hl=en>.
- — —. “YouTube Partner Program overview & eligibility.” Accessed June 23, 2025.
<https://support.google.com/youtube/answer/72851?hl=en&co=GENIE.Platform%3DAndroid>.
- — —. “YouTube’s Community Guidelines.” Accessed August 15, 2025.
<https://support.google.com/youtube/answer/9288567?hl=en>.
- Zhang, Yini, Jiyoun Suk, Dongdong Yang, et al. “Center-Left and Right-Wing News, YouTube, and Twitter as Key Connectors in the Social Media System: A Cross-Media and Cross-Platform Analysis of Hyperlinks.” *Journal of Quantitative Description: Digital Media* 5 (2025).
<https://journalqjd.org/article/view/8674/7412>.

Zimmer, Michael. ““But the data is already public’: on the ethics of research in Facebook.” *Ethics and Information Technology* 12 (2010): 313-325.

<https://link.springer.com/article/10.1007/s10676-010-9227-5>.

Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. Profile Books, 2019.