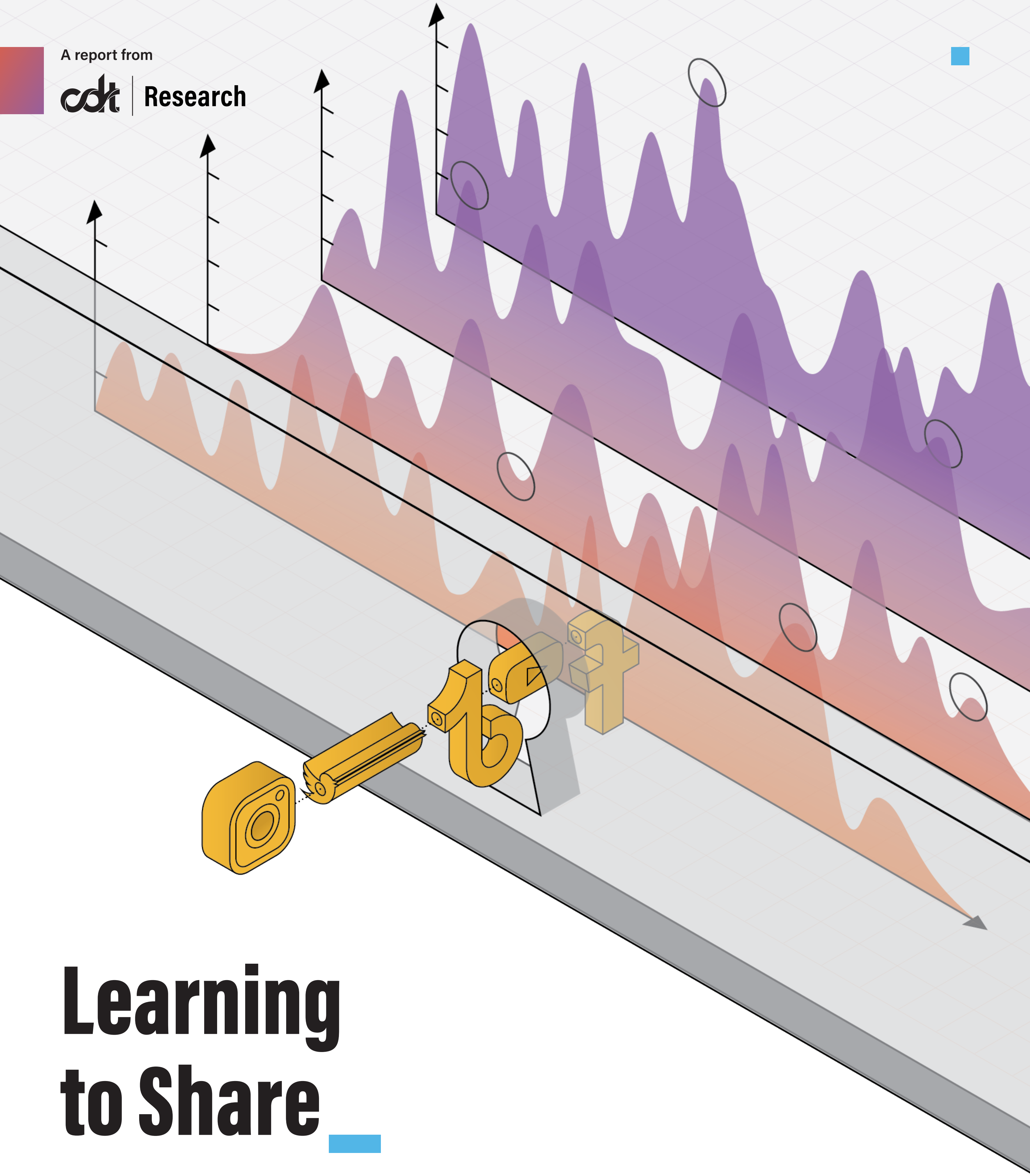
A report from
cdt | Research



Learning to Share

Lessons on Data-Sharing from Beyond Social Media

Gabriel Nicholas
Dhanaraj Thakur

September 2022

CENTER FOR
DEMOCRACY
& TECHNOLOGY

The Center for Democracy & Technology (CDT) is a 25-year-old 501(c)3 nonpartisan nonprofit organization working to promote democratic values by shaping technology policy and architecture. The organisation is headquartered in Washington, D.C. and has a Europe Office in Brussels, Belgium.

GABRIEL NICHOLAS

Research Fellow at the Center for Democracy & Technology.

DHANARAJ THAKUR

Director of Research at the Center for Democracy & Technology.

A report from

cdt | Research

Learning to Share

Lessons on Data-Sharing from Beyond Social Media

Gabriel Nicholas

Dhanaraj Thakur

WITH CONTRIBUTIONS BY

Caitlin Vogus, Emma Llansó, Samir Jain, Aliya Bhatia, Will Adler, and Michal Luria. Ari Goldberg copy-edited the draft and Tim Hoagland led the design and layout of the report.

ACKNOWLEDGEMENTS

We thank Rebecca Williams for the extensive research she provided to support this report and the experts she interviewed for sharing their time. We also thank the external reviewers of this report for their insightful comments and suggestions. They are Christopher Morten, Jessica Wentz, and Michael Murray. All views in this report are those of CDT.

This work is made possible through a grant from the John S. and James L. Knight Foundation.

Suggested Citation: Nicholas, G. and Thakur, D. (2022) Learning to Share: Lessons on Data-Sharing from Beyond Social Media. Center for Democracy & Technology. https://cdt.org/insights/learning-to-share-lessons-on-data-sharing-from-beyond-social-media/

References in this report include original links as well as links archived and shortened by the Perma.cc service. The Perma.cc links also contain information on the date of retrieval and archive.

Contents

Contents

Executive Summary

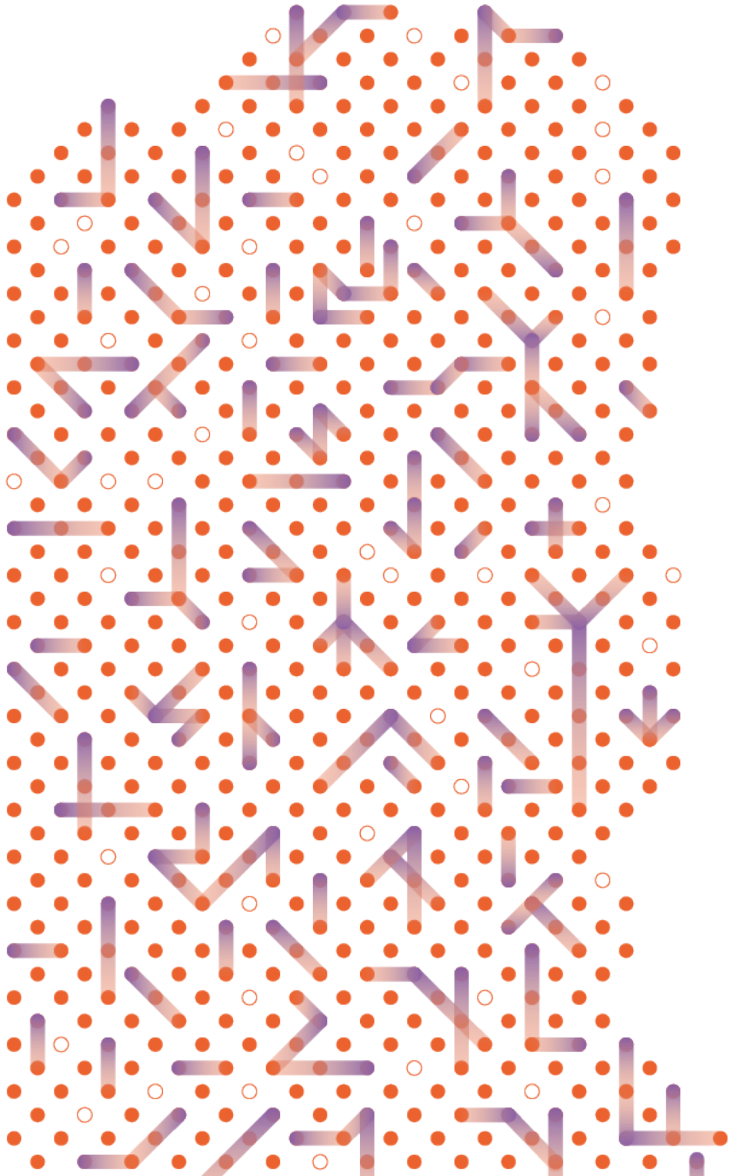
What role has social media played in society? Did it influence the rise of Trumpism in the U.S. and the passage of Brexit in the UK? What about the way authoritarians exercise power in India or China? Has social media undermined teenage mental health? What about its role in building social and community capital, promoting economic development, and so on?

To answer these and other important policy-related questions, researchers such as academics, journalists and others need access to data from social media companies. However, this data is generally not available to researchers outside of social media companies and, where it is available, it is often insufficient, meaning that we are left with incomplete answers.

Governments on both sides of the Atlantic have passed or proposed legislation to address the problem by requiring social media companies to provide certain data to vetted researchers (Vogus, 2022a). Researchers themselves have thought a lot about the problem, including the specific types of data that can further public interest research, how researchers should be vetted, and the mechanisms companies can use to provide data (Vogus, 2022b).

For their part, social media companies have sanctioned some methods to share data to certain types of researchers through APIs (e.g., for researchers with university affiliations) and with certain limitations (such as limits on how much and what types of data are available). In general, these efforts have been insufficient. In part, this is due to legitimate concerns such as the need to protect user privacy or to avoid revealing company trade secrets. But, in some cases, the lack of sharing is due to other factors such as lack of resources or knowledge about how to share data effectively or resistance to independent scrutiny.

The problem is complex but not intractable. In this report, we look to other industries where companies share data with researchers while also addressing privacy and other concerns. In doing so, our analysis contributes to current public and corporate discussions about how to safely and effectively share social media data with researchers. We review experiences based on the governance of clinical trials, electricity smart meters, and environmental impact data.



Clinical Trials[1]

In most cases, the FDA requires companies and research centers to share data about the clinical trials they use to verify the safety and efficacy of a medical product as a condition of bringing that product to market. Group-level summary data and metadata on the studies' methodologies is made publicly available on ClinicalTrials.gov. Voluntary mechanisms such as the Yale Open Data Access Project (YODA) enable those running clinical trials to securely share additional anonymized data with independently vetted researchers for independently approved projects. In general, researchers use clinical trial data to monitor the safety and efficacy of certain drugs, assess the validity of trial methodologies, and at a more meta-level, assess the extent to which companies are complying with their data publishing requirements.

Electricity Smart Meters

Smart meters record electricity usage and report this information back to the utility for billing purposes using radio frequency networks. Researchers also use smart meter data to help evaluate and improve the energy efficiency of buildings (Adams et al., 2021), inform energy demand response strategies (National Council on Electricity Policy, 2008), and improve battery management (Zheng et al., 2019). Some smart meter data sets collated by academics and governments exist, but it is more difficult for researchers to request data from utilities, especially for a specific geographical area. Some states have pathways for researchers to get electricity consumption data directly from utilities, though that data cannot exceed certain aggregation and anonymization thresholds. Often, these thresholds are easy to understand and implement, but in their simplicity can be too conservative in some cases or too liberal in others, needlessly preventing harmless research, or failing to protect some individuals, respectively.

Environmental Impact Statements

Government agencies are required to assess the environmental impact of any project that uses federal land, federal tax dollars, requires federal authorization, or is under the jurisdiction of a federal agency (Middleton, 2021). These assessments come in the form

[1] The analysis of clinical trials in this report is based on a forthcoming law review article by Christopher Morten, Gabriel Nicholas, and Salomé Viljoen, which in much greater depth considers lessons social media can draw from the clinical trial sector's legal and technical approaches towards sharing data with researchers. For a copy of the latest draft of that article, contact the authors at cjm2002@columbia.edu, gnicholas@cdt.org, or sviljoen@umich.edu. The article is cited here as "Morten et al., forthcoming".

of an environmental impact statement (EIS), which the public can then comment on (U.S. EPA, n.d.-b). Researchers use the EIS process itself as a source of political leverage for citizen science, and also use historical data to both assess methods of mitigating environmental harm (Marcot et al., 2001) and evaluate and improve the effectiveness of the environmental review process itself (O'Faircheallaigh, 2010). However, EIS data does not come in a standardized form that can be easily used by researchers. In the United States in particular, EISs are allowed to exclude a lot of information under the protections of trade secrecy (Lamdan, 2017). Alternatively, in the UK, there is a public interest test: authorities "can refuse to provide information only when the public interest in maintaining the exception outweighs the public interest in disclosure" (Information Commissioner's Office, 2022).

Lessons for social media companies from other industries

Using these three cases, we outline ten lessons that policymakers, companies, and others should consider when developing policies to improve researcher access to social media data:

1. *Sharing data with researchers can help make more informed policy decisions.* Clinical trial data, smart meter electricity data, and data underlying environmental impact statements are all governed in a way that lets researchers use the knowledge they gain to help inform the policymaking process. When designing mechanisms to give researchers access to social media data, policymakers should consider designing analogous feedback loops.
2. *Sharing data can let researchers double check otherwise unverifiable corporate claims.* Social media companies often respond to public criticism by making changes to their systems, but there is no way for independent researchers to verify the effectiveness or veracity of these changes. Other sectors show a way forward — clinical trial data is shared in a way that is particularly designed to allow third-party researchers to stress-test and verify whether medical products work. Environmental impact statements further shift the paradigm, allowing the public to identify shortcomings or knock on effects before an intervention is rolled out.
3. *The "denominator problem" can be addressed without compromising privacy.* When an independent researcher establishes some finding based on the limited data they have available, there is no way for them to precisely determine the overall size of the finding relative to the social media platform in question. For example, if they find that 10% of users in a given sample of data share misinformation it's hard to know what that means about the population of all users on the platform. Experience in other industries show that aggregation and anonymization techniques can allow this kind of population related information to be shared without compromising individual privacy.

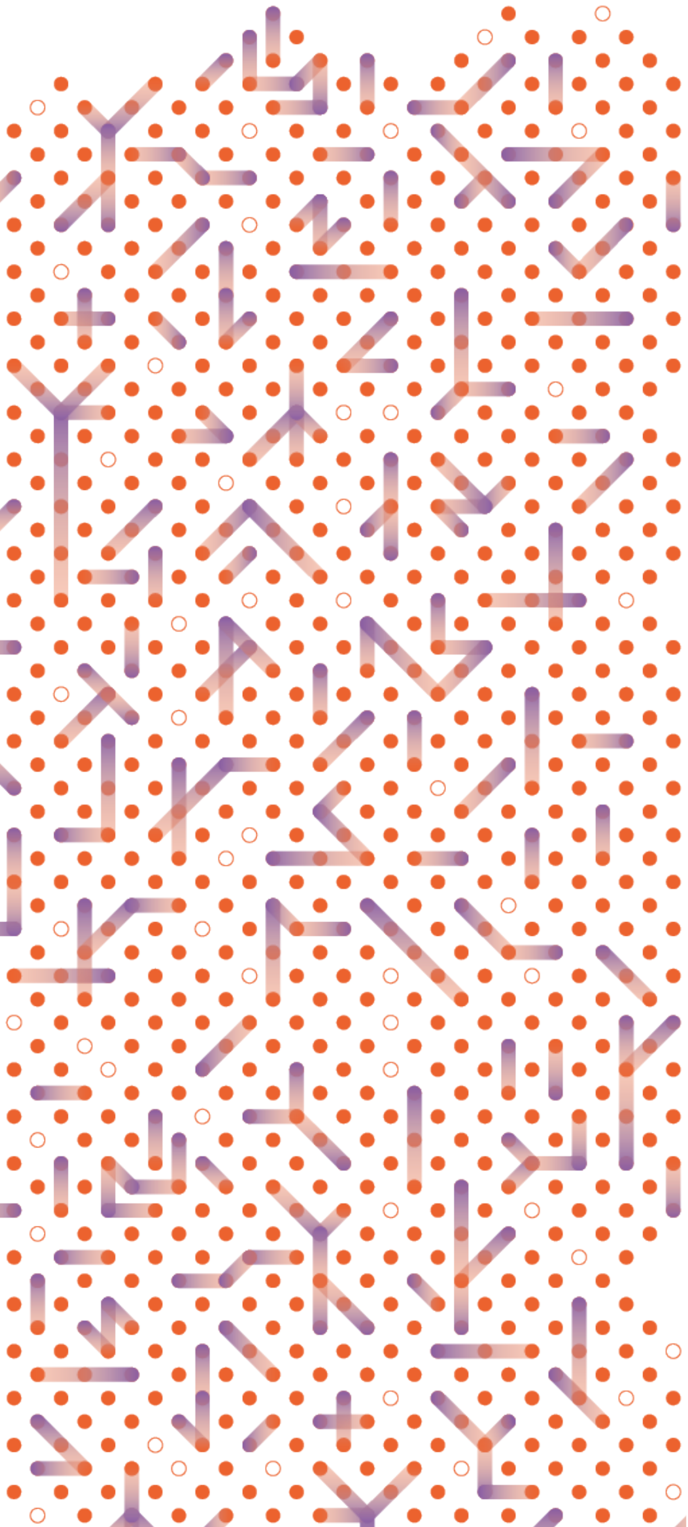
4. *Addressing the "black box" problem will make research more widely applicable.* Researchers struggle to use data sharing tools provided by social media companies because they offer little information on how a given data set was produced. Clinical trials, on YODA, ClinicalTrials.gov, and elsewhere give researchers the context they need by including metadata about how the data was generated, such as trial protocols and statistical analysis methods.
5. *Transparency mechanisms let civil society serve as data sharing watchdogs.* The lack of data available for researchers, particularly those in civil society, undermines attempts at meaningful transparency and accountability for social media. The EIS review process and FDAAA Trials Tracker, which uses ClinicalTrials.gov data to calculate how many covered trials have reported their results, show how sharing even a little data with researchers can contribute meaningfully to oversight.
6. *Standards make shared data usable.* Standards are an important way for researchers to know what data to expect and how they can expect to receive it. Robust standards set by the FDA and NIH have made clinical trial data more useful for researchers. A lack of those standards has made EIS data less systematized and thus less useful. Today, with each platform having its own protocol for sharing data, social media falls closer to the latter camp.
7. *Data sharing should be flexible to accommodate public crises.* The experiences from the three industries show that normative trade-offs can be made when it comes to public crises and sharing data. For example, ClinicalTrials.gov expedited and broadened its data sharing about COVID-19 vaccines, though many in the medical community called for even greater transparency than they actually provided. Social media should support greater access to data when the public interest is particularly important such as in the case of events such as natural disasters and elections.
8. *Ease of understanding is a factor to consider in privacy.* Social media companies tend to be opaque about the methods they use to ensure user privacy. Examples from other industries show how privacy rules for preparing and sharing data can be intuitive and easier to understand. The 15/15 rule with smart meter data (where each geographical unit of data requested must include at least 15 commercial customers, and no customer may make up more than 15% of the total power usage) and the 18 direct identifiers in HIPAA, which cannot be shared in clinical trials, for example, are easy for the public to understand and likely easier to enforce, though they also come with sacrifices in effectiveness.

9. *Data access can be tailored to different use cases.* A tiered access approach is sometimes posited for and used by social media companies when it comes to access to data. In other industries, access is also more specifically tailored for the capabilities and goals of different types of researchers, such as the California Public Utilities Commission's (CPUC) distinction between government and academic researchers. This approach allows for greater flexibility where researchers are more likely to be able to access the most useful type of data for their research.

10. *Diverse data stewards offer new affordances.* In each of the industries we examined, different actors play a role in facilitating the sharing of data, including private, government, academic, and civil society organizations. EIS data, for example, is organized both by government actors, such as the EPA, and academic actors, such as Northwestern. Clinical trial data is also shared by multiple actors, through compulsory and voluntary data sharing mechanisms. This expands the range of options (types of data, requirements, limitations, etc.) available to researchers. Social media does not benefit from this diversity because as of now, private companies are the sole stewards of data.

Introduction

Social media plays an important role in nearly every economic, social, and political institution — what products people buy, which groups people support, how people vote, and so on. It is also an important means for content creators and innovators to share their ideas and work, including through music, fashion, and politics. As a result, social media plays a key role in determining which ideas flourish and which flounder. If we fail to understand social media, we may fail to fully understand how society works.



Despite its relevance to so many facets of human life, researchers have only begun to understand the effects of social media on society at large. Researchers have identified some potential impacts of social media, such as its benefits as a tool for community building and political organizing (Segerberg & Bennett, 2011), its harms as a source of radicalization (Marwick et al., 2022), filter bubbles (Pariser, 2012), and various forms of abusive content (Citron, 2014). But there is a lot we also don't yet understand, such as the efficacy of organizing tools, the relative scale of filter bubbles, and the real life impacts of these phenomena on users. In general, research on the effects of social media is a nascent, evolving field, and much more (and different kinds of) research is needed to fully understand these complex phenomena.

One reason our knowledge is so incomplete is how new social media is. Another is its sheer size: as many as 72% of Americans (Pew Research Center, 2021) and half the world population (Dixon, 2022) use social media; globally, at least 500 hours of content are uploaded to YouTube every minute (YouTube, n.d.-a), and over 1 billion stories are uploaded each day on Facebook (Meta, n.d.). As a result, it is hard to conceptualize social media as a single "thing." However, neither of these factors explains the knowledge gap fully. There are only a handful of major social networks, so that should in some ways make studying them easier than, say, school districts, of which there are thousands in the U.S. alone (National Center for Education Statistics, 2012). And social media platforms are not *that* new — Facebook was started in 2004, YouTube in 2005, and Twitter in 2006.

Another perhaps more significant reason we understand so little about the societal effects of social media is that the companies who run the major platforms have failed to give researchers the data they need to do effective research. Many services currently make little to no data available to researchers, including TikTok, Instagram, LinkedIn, and

Snapchat.[2] When services do make data available (e.g. Facebook, Twitter), the data they provide can be faulty (e.g. Timberg, 2021), missing key contextual information (Tromble, 2021), or insufficient in volume (Edelson & McCoy, 2021). And when researchers build tools to ask users directly for data, social media companies sometimes shut down their efforts (e.g. Hatmaker, 2021; Kayser-Bril, 2021).

Social media companies have a range of possible rationales for withholding data from researchers. Small and medium-sized social media companies may simply lack the resources and expertise necessary to share data with researchers. Larger social media companies have highlighted the need to protect the privacy and security of their users (e.g. Clark, 2021). To lawmakers, they have emphasized the need to protect trade secrecy to ensure innovation (Meta, 2021; Protalinski, 2011). While privacy, security and trade secrets concerns may sometimes be legitimate (Hearing on Platform Transparency: Understanding the Impact of Social Media, 2022), social media companies may also be reluctant to subject themselves to independent scrutiny, perhaps out of fear that research conclusions based on independent analysis (e.g. Imana et al., 2021; Yin & Sankin, 2021) will tarnish their brands' reputations.

A balance must be struck between giving researchers access to data and protecting the interests of users and companies. For most of their history, social media companies have had near-total power to strike that balance as they please. The Digital Services Act in Europe will change that by requiring very large online platforms to make at least some data available to researchers (Digital Services Act, Article 31, 2020). Some legislative proposals in the U.S. also aim to rebalance that power, such as the Platform Accountability and Transparency Act, Social Media Data Act , Digital Services Oversight and Safety Act, and Kids Online Safety Act (Vogus, 2022a).

Platforms have also made voluntary attempts to provide researchers with improved access to data, such as Twitter's academic API (Twitter, n.d.-b) and Moderation Research Consortium (Roth & Gadde, 2021), Facebook Open Research and Transparency initiative (Facebook, n.d.; Isaac, 2022), YouTube's Researcher Program (YouTube, n.d.-b), and TikTok's announcement that it will provide a moderation system API for researchers (Pappas, 2022). However, many researchers feel these efforts still fall short in the quality, quantity, and freedom of use for data (Edelson & McCoy, 2021; Shapiro et al., 2021; Vogus, 2022b). Current practices, indeed the current paradigm of how social media companies could or should share data while balancing their own needs is unsatisfying and in need of new ideas.

[2] As of this writing, YouTube only recently began to offer API access to university affiliated researchers (YouTube, n.d.-b). TikTok announced it would offer an API by the end of 2022 but has not yet provided specific documentation for what it will include (Pappas, 2022).

To help come up with those new ideas, this paper examines how companies in industries outside of social media share data with independent researchers, either as required by law or through voluntary mechanisms. In particular, we reimagine how social media companies could make data available to researchers by drawing lessons from the successes (and failures) of attempts to share three other types of data with researchers: clinical trial data, electricity smart meter data, and data used in environmental impact assessments.

We could have chosen from among many types of data in several industries for this study (e.g. NHTSA automotive data, education data, genomic data, census data. However, we chose clinical trial data, electricity smart meter data, and environmental impact assessments for three reasons. First, these industries are older and have had time to develop a more mature consensus between researchers demanding data and companies holding it, even if those arrangements are imperfect. Second, some core characteristics make all three types of data somewhat comparable to social media data, particularly since there are public benefits of sharing all of this data, and harms that would arise from improper or insufficient data sharing. Third, these three types of data also have very different data governance paradigms from social media data, which can be helpful for breaking existing modes of thought.

This report will examine how the industries that hold these three kinds of data answer three open questions that social media has to answer for making its data available to researchers:

1. How can researchers be ensured access to useful data?
2. How can data be shared in ways that respect individuals' privacy?
3. How can third parties (besides the data holder and data requester) facilitate access to data?

Part 1 of this paper will briefly summarize how social media companies deal with these questions and how they often come up short. Part 2 will look at how governors of data from clinical trials, electricity smart meters, and environmental impact assessments deal with these same problems. Part 3 will summarize new approaches social media companies could adopt based on what has and hasn't worked in other industries.

Researcher Access to Data on Social Media

How can researchers be ensured access to useful data?[3]

Shapiro et al. divides the types of social media data that researchers are interested in into three categories: "(1) *content data*, which refers to the information posted on social platforms by users, by advertisers or by the companies themselves; (2) *moderation data*, which refers to information about content throttling, content labeling, and information about which posts were allowed to be posted on a platform and which were removed; and (3) *distribution data*, which encodes information about who encounters what content posted online" (2021, p. 17). Shapiro et al. also discusses three methods social media researchers have for accessing data: *platform-sanctioned open data*, such as APIs (e.g. Twitter API) or released datasets (e.g. Meta Ad Library); *commercially available platform data*, usually bought through social media listening platforms, such as Brandwatch and Meltwater, that are aimed at marketers (Hayes et al., 2021); and *unsanctioned open media ecosystem data*, such as the unofficial TikTok API (Teather, 2019/2022), or user-permissioned web scrapers, like Mozilla Rally (Mozilla, 2022) or Ad Observer (NYU Cybersecurity for Democracy, n.d.). The methods and data types can be mapped on to a table (see Figure 1).

[3] For those interested in a more in depth discussion of how social media companies address the challenges of researcher access to data see: (Persily & Tucker, 2020; Shapiro et al., 2021; Vogus, 2022b; Vogus & Llansó, 2021).

	Platform-sanctioned open data	Commercially available social platform data	Unsanctioned open media ecosystem data
Content data	Meta Ad Library, Twitter Academic API	TikAPI	Pushshift
Moderation data	Twitter Moderation Research Consortium; transparency reports		Shadowban detection tools
Distribution data	Crowdtangle	Social media listening platforms (e.g. Brandwatch, Meltwater)	Mozilla Rally, NYU Ad Observer

▲ **Figure 1. Examples of data access services for researchers, divided by type of data and access method.**

Each of these methods, when made available, has shortcomings. On platform-sanctioned open data methods, for example, social media companies often sharply limit how much data researchers can access. For example, the academic Twitter API allows researchers to get ten million tweets per month (Twitter, n.d.-a). Academics have noted that this is significantly lower than it once was and insufficient to evaluate given topics (Vogus, 2022b). Platform-sanctioned data can also be biased in ways that platforms don't document. For example, Facebook data released on interactions with political pages was found to be missing users whose political affiliations were not clearly identifiable from those interactions, which amounted to about half of all potential users (Alba, 2021).

In addition, certain types of data are nearly inaccessible through sanctioned means alone. Moderation data is particularly scarce, and largely impossible for researchers to obtain without help from platforms themselves. To address the limitations of platform sanctioned methods, academics and others have built systems to allow users to donate their data for research, which could then be subject to problems of representativeness. Nevertheless, platforms have sometimes responded by shutting these efforts down, such as data donation tools used by Algorithm Watch and NYU Ad Observatory.

How can data be shared in a way that respects individuals' privacy?

Social media data is often sensitive and personally identifiable. There are active debates about how much access researchers, specifically, should have to particular types of data and what kind of privacy and cybersecurity protections should be in place. Social media companies have repeatedly used user privacy as a justification to not share data (e.g. Clark, 2021), and there are genuine risks to individual privacy and safety from certain kinds of data disclosure. However it is not always clear when companies are raising this out of genuine concern for their users and when it is a case of what Van Loo calls a "privacy pretext", i.e. a case of a private company "exploiting privacy to avoid competition and accountability" (Nicholas, 2021; Van Loo, forthcoming; Viljoen, 2022).

Companies have multiple, sometimes competing incentives concerning data sharing, including protecting their users, pleasing their advertisers, and avoiding scandal and bad press. Some researchers have responded to companies' refusals to provide them with data by engaging in unsanctioned data collection, but if not done carefully, unsanctioned data collection could enable unethical practices. As an extreme example, the Cambridge Analytica app that collected so many users' data, thisisyourdigitalife, was nominally designed for academic research (Cadwalladr & Graham-Harrison, 2018). A related problem is data purchased from commercial sources. In such cases, data may be acquired using unethical practices or without the knowledge or consent of the social media user, which pose a similar problem to that of data brokers (Franklin et al., 2021).

How can third parties facilitate access to data?

When negotiating access to data, researchers and social media companies often have interests at odds with one another. Researchers, though they do want to protect the privacy of subjects for ethical and other reasons, are otherwise incentivized to seek access to as much data as they believe they need to do interesting and useful research. Social media companies on the other hand, though perhaps interested in an improved understanding of how their platforms work and the reputational benefits that come from being considered an open data sharer, are otherwise incentivized to restrict data access to researchers, since the possibility of disclosure of private information, legal violations, leaks, bad press, and exposure of business secrets are against their interest.

Other actors in government, civil society, and academia could help intermediate this process, but, as of today, they largely choose not to. The Digital Services Act in Europe will do this, and the Platform Accountability and Transparency Act (PATA) in the U.S. proposes to do this, by compelling platforms to share certain data with researchers. Upon request of the Digital Services Coordinator of establishment, Article 31 of the DSA will require providers of very large online platforms to turn over certain data to certain researchers approved by that body. PATA would similarly create a way for researchers to submit projects to the National Science Foundation and ask them to compel social media companies to share the necessary internal data for them. (Vogus, 2022a). However, researchers may not even know what data they can and should ask for from social media companies, since they do not know what data they hold (Shapiro et al., 2021; Vogus, 2022b).

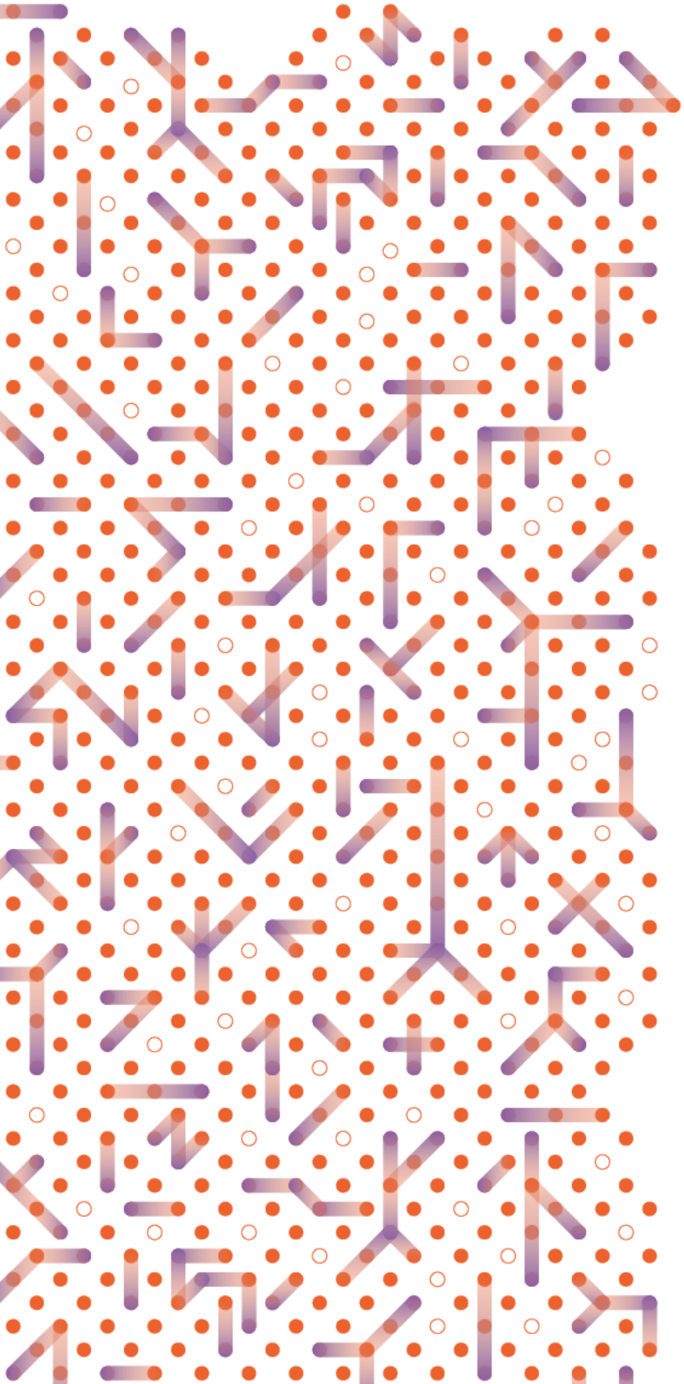
Researcher Access to Data in Other Industries

We now turn to examine how governors of data from clinical trials, electricity smart meters, and environmental impact assessments attempt to deal with the challenges of how data can be shared with researchers in a privacy protective way and the role of third-parties in supporting that kind of access.

Clinical Trials[4]

The Food and Drug Administration (FDA) was created to protect public health by ensuring that health interventions such as drugs, vaccines, and medical devices are safe and efficacious (Food and Drug Administration, 2018). Corporate and academic researchers can only gain FDA approval to bring their health products to market if they validate their safety and efficacy with randomized controlled trials with human subjects called *clinical trials*. Pharmaceutical companies and academics have been required to share the results and methodologies of their clinical trials with the FDA since 1962 (Halperin, 1979).

Notwithstanding these requirements, companies that run clinical trials may be incentivized not to share their underlying data, or to minimize the data they do share: clinical trials can reveal very sensitive and private information about people, datasets can be very large and thus expensive to generate and maintain, and there can be significant economic and competitive value in these data (Kapczynski & Morten, 2021). In the last sixty years, the systems for sharing clinical trial data with the FDA, researchers, and the general public have become far more robust and multifaceted. In this section, we look at how different actors share clinical trial data at different levels of specificity and to different stakeholders.



[4] The analysis of clinical trials in this report is based on a forthcoming law review article by Christopher Morten, Gabriel Nicholas, and Salomé Viljoen, which in much greater depth considers lessons social media can draw from the clinical trial sector's legal and technical approaches towards sharing data with researchers. For a copy of the latest draft of that article, contact the authors at cjm2002@columbia.edu, gnicholas@cdt.org, or sviljoen@umich.edu. The article is cited here as "Morten et al., forthcoming".

Data from medical product clinical trials can be organized into three categories: complete data, summary data, and metadata (Institute of Medicine, 2015; Morten et al., forthcoming):

- *Complete data (or individual patient data)* is the entirety of the raw data collected on individual patients throughout a given trial. It includes information on diagnoses, treatments, interventions, side effects, and troves of other personal health information. This data can be the most useful for some researchers but also the most sensitive. This data can allow researchers to reproduce the results of a clinical trial and run in depth analyses to double check its work.
- *Summary data* is data aggregated across a variety of demographics, typically including age, race, gender, and health conditions. It also includes takeaways from the trials, including conclusions about a drug's safety and/or efficacy or its lack thereof. Categories of summary data are not determined ad-hoc, rather they are standardized by government agencies (e.g. FDA, NIH) and medical organizations (42 CFR § 11.48). There are many categories of summary data, and summary data can sometimes run thousands of pages for a single study (Sharfstein et al., 2017). This data can help third parties understand how medical interventions may affect different demographic groups. It can also help other members of the medical community identify promising avenues for future research.
- *Metadata* is data about how the trial data gets generated. In other words, it is information on the way the study was conducted — the trial protocol, the methods of statistical analysis, what patient outcomes are measured, and other precise methodological questions (Morten et al., forthcoming). This data can help contextualize findings and aid watchdogs in spotting mistakes or fraud.

The FDA does not make complete data from clinical trials available to third parties because it includes extremely sensitive and personal medical information, but it does make summary and metadata available. The FDA itself publicly releases some summary and metadata from trials for the drugs that it approves (Food and Drug Administration, 2019), though the statutory language (21 U.S. Code § 355 (l)) is vague, and since 2020, the FDA has been sharing the bare minimum data required (Herder et al., 2020). Separately, the FDA and NIH require *companies* to publish extensive summary and metadata on ClinicalTrials.gov, which is available to the public. This has made far more data available to researchers — as of this writing, nearly half a million studies are registered on ClinicalTrials.gov with 55,000 providing research results (National Institutes of Health, 2021). Under the Food and Drug Administration Amendments Act of 2007, drug companies are required to publish summary and metadata for most clinical trials, even ones for products that have not or will not make it to market, within one year of completing the trial (42 USC 282(j)).

ClinicalTrials.gov has been an important source of data for scientists doing "secondary research," that is, research on previous studies that may detect overlooked patterns, mistakes, or outright fraud. Academics, watchdog groups, and individual patients all engage in this kind of work. Access to clinical trials data has allowed researchers to identify previously unknown safety problems in drugs, such as the painkiller Bextra (Wolfe, 2004) and the antidepressant Paxil (Doshi, 2015), which was prescribed to over two million children per year before it was found to cause suicidal ideation in young people. ClinicalTrials.gov has allowed even more groups to act as watchdogs, including with the antidiabetes drug rosiglitazone (traded as Avandia), which researchers found was associated with increased cardiovascular risk (Wallach et al., 2020). ClinicalTrials.gov data has also been used to double check COVID-19 vaccine research and vaccine information (Korang et al., 2022).

Though the ClinicalTrials.gov data sharing system has had some successes, it falls short of its potential. About 45% of industry-funded clinical trials are not required to be reported to ClinicalTrials.gov (Anderson et al., 2015). However, ClinicalTrials.gov does not even have data on all the trials it should. Science looked at 4700 clinical trials that were required to upload their results to ClinicalTrials.gov and found that less than 45% had reported their results on time and 31% had not reported at all (Piller, 2020). The federal government can fine companies $11,569 per day if they fail to report data from their trials (45 CFR § 102.3), meaning that in total, the government could have levied fines for over $33 billion. To date, however, it has imposed no fines (DeVito et al., 2020; FDAAA Trials Tracker, n.d.). Companies may also withhold the most important data: some in the medical community have speculated that pharmaceutical companies are hesitant to share bad or controversial results (Harris, 2015).

There are voluntary alternatives to ClinicalTrials.gov that allow researchers to access complete data from clinical trials. One alternative is the Yale Open Data Access Project (YODA), which partners with companies to host de-identified but otherwise complete clinical trial data sets on their secure platform (Yale University Open Data Access Project, n.d.-c). This allows vetted researchers to gain far deeper access to clinical trial data and thus, conduct more in depth research. The YODA Project started as a partnership with Medtronic, but today, 95% of the 421 trials it hosts are from Johnson & Johnson (Ross et al., 2019; Yale University Open Data Access Project, n.d.-c). As of this writing, the YODA Project has received 283 requests and approved 94.5% percent of them, leading to a total of 71 peer reviewed publications (Yale University Open Data Access Project, 2022).

In order to gain access to a dataset hosted on the YODA Project, researchers must submit a project proposal for approval (Yale University Open Data Access Project, n.d.-a). (There are no specific eligibility criteria for researchers, but all applicants must go through YODA's vetting process to demonstrate their capacity to conduct the study.) Proposals include the purpose of the project, the data they are requesting, how they will use that data to inform science and public health, and how that research can reasonably be done with that data (Ross et al., 2019). Proposals are reviewed double blind by two independent reviewers from the scientific community who are unaffiliated with the YODA Project or the company that provided the data (Yale University Open Data Access Project, n.d.-b). Once accepted, researchers gain access to a virtual platform where they can run analyses. Participant level data can only be accessed through the platform, but some of the associated metadata, such as protocols, statistical analysis plans, and data definition specifications, can be downloaded off the virtual platform (Yale University Open Data Access Project, n.d.-b). Researchers pay no fee for applying or for accessing data. Costs are covered by the data providers (Ross et al., 2018).

Data access on the YODA Project comes with some restrictions. Researchers must sign a data use agreement, which requires that they only disseminate their findings through a peer-reviewed publication and not share clinical trial data with others (although other researchers could apply to analyze the same data). The agreement also requires researchers to destroy any data they download after the research is complete. Data cannot be used for litigation or commercial interests and the data provider holds rights to any inventions that come directly out of the data (Yale University Open Data Access Project, n.d.-a). The YODA Project only hosts data from trials for products that have received regulatory approval and were completed more than 18 months in the past (Ross et al., 2018).

Finally, to ensure privacy, YODA Project data providers also de-identify patient data. This entails either removing the 18 direct identifiers listed in HIPAA (e.g. name, address, IP address, account number, specific dates such as birth or discharge date, etc. (Loyola University Chicago, n.d.)), or getting approval from an expert statistician that there is a low risk that the information could be used to identify an individual. However, very early stage clinical trials and trials for drugs to combat rare diseases often cannot be fully de-identified, so they are usually not included in YODA datasets (Yale University Open Data Access Project, n.d.-a). To date, there has been no known mis-use of YODA datasets (e.g., re-identification, unauthorized transfer to data, or sale of data).

Electricity

For decades, electric utilities could only collect data on how much electricity their customers used by sending an employee to manually read a meter attached to their home. But since the 1990s, smart meters have spurred a dramatic shift in the electricity industry: in-home electricity smart meters went from being a rare gadget to the dominant way electric utilities measure and bill customers for how much electricity they use in homes and buildings (Strong, 2017). Today, in part due to a shift in federal energy policy (Energy Independence and Security Act of 2007, 2007) and federal subsidies in the American Recovery and Reinvestment Act for smart meters, over 75% of U.S. homes have smart meters (Jones, 2021), and a total of over 115 million were estimated to be deployed at the end of 2021 (Cooper & Shuster, 2021).

Smart meters record electricity usage every 5-, 15-, 30- or 60 minutes and report this information back to the utility using radio frequency networks for billing purposes. This allows customers to know exactly what their energy bill will be instead of having an estimated bill (Torriti, 2020). It also allows better utilization of distribution infrastructure, which is necessary to manage increased demand for electricity from things like electric cars. Smart meters can potentially make it easier to integrate renewables onto the grid by identifying local areas of congestion on the distribution system, particularly as rooftop solar becomes more widespread.

However, smart meter data also raises serious privacy concerns. Many smart meters collect about 3,000 data points per month, and usage patterns could reveal when people are asleep, at work, or traveling (Lerner, 2008). Certain electrical signatures can even reveal when a television or medical device is in use (Duarte, 2015). This data could be of interest to many, including financial institutions that make mortgage loans, insurers or advertisers targeting ads. It could also be used by law enforcement to detect criminal or simply erratic behavior (Lynch & Tien, 2010). Although there is still public distrust of smart meters, both for privacy reasons (Hooks, 2013; Stop Smart Meters!, 2011) and rumors of excess radiation (American Cancer Society, n.d.), NIST has released extensive guidelines to manage cybersecurity and privacy risk around this data (Pillitteri & Brewer, 2014).

Even before electricity meters were "smart", researchers have been interested in access to electricity usage data in order to test technological and policy interventions aimed at improving energy efficiency and reducing carbon emissions. Researchers can use data about where and when energy is being used to support other energy analysis methods that help evaluate and improve the energy efficiency of buildings (Adams et al., 2021), inform energy demand response strategies (National Council on Electricity

Policy, 2008), and improve battery management (Zheng et al., 2019). Sub-hourly measurements of energy usage are also used to identify faults in heating, ventilation and air conditioning equipment, and, when summed into daily usage figures, the data can help measure the effect of external conditions like weather (e.g. Kang & Reiner, 2022). This can inform policy efforts focused on weather-sensitive loads (insulation, HVAC system efficiencies, etc.) and non-weather-sensitive loads (behavior patterns, appliance efficiencies, etc.). Smart meter data has even been used to highlight the inequitable effects of climate change, showing that in extreme hot or cold temperatures, poor people end up paying disproportionately more in energy bills (Chen et al., 2022).

Individual smart meter data is rarely made available to researchers, but several collated datasets of how individual households consume power have been published (Himeur et al., 2020; Li et al., n.d.). Data sets range in the number of households (dozens, hundreds, or in a few cases, thousands), the length of time data is collected for (weeks, months, or years), and the intervals at which data is collected (seconds or minutes) (Li et al., n.d.). In the European Union one of the most comprehensive datasets comes from the Irish Commission for Energy Regulation's Smart Metering Project, which collected data from over 5,000 homes and businesses (Commission for Energy Regulation, 2012). Larger datasets like this allow researchers to use data analytics to cluster buildings into categories of how they use electricity (Albert & Rajagopal, 2013). Again, this creates some privacy risk, especially if data were to be de-identified, but also allows for improved energy efficiency and carbon emissions mitigation.

However, researchers interested in accessing electricity usage data from a specific geographical area have a more difficult time. The technical standard for individuals sharing their own smarter meter data is called Green Button Connect (Green Button Data Alliance, 2022). This standard is not tailored to the needs and data flows of researchers, though; instead, it is designed to allow customers to download their own data or share it automatically with third party applications, usually energy budgeting apps or other energy providers that can offset customers' energy usage with renewables (Green Button Data Alliance, 2022). Even in this latter context though, many utilities poorly implement Green Button Connect, and their data sharing is filled with delays, incorrect data, unplanned outages, and poor conformance to the Green Button standard (Mission::data, 2019).

Some states, including Colorado, Illinois, and New York, explicitly allow customers to share their Green Button data with whomever they please (Mission::data, 2019, p. 6). The California Public Utilities Commission (CPUC) has gone further, authorizing a specific pathway for people to legally donate their electricity consumption data to public interest research (Decision Establishing Building Decarbonization Pilot Programs, 2020), but so far it has not been put into practice (Best et al., 2021, p. 11).

A handful of states such as California and Illinois have pathways for researchers to get electricity consumption data directly from utilities without direct consent from the consumer (American Council for an Energy-Efficient Economy, 2016). California, for example, has had its Energy Data Request Program since 2014. The program was outlined by the CPUC's Decision 14-05-016, which established data access baselines for various stakeholders. Per the decision, different groups received different volumes and granularity of data (Decision Establishing Building Decarbonization Pilot Programs, 2020). The public, for instance, can access quarterly gas and electricity usage, aggregated by zip code or customer classes (i.e., residential, commercial, industrial, agricultural) (PG&E, n.d.-a). Local governments can access monthly electricity usage aggregated by zip code or census block group (Decision Establishing Building Decarbonization Pilot Programs, 2020; PG&E, n.d.-a). However, the data they receive cannot exceed certain aggregation and anonymization thresholds. For example, for data on commercial electricity usage, each geographical area requested must include at least 15 commercial customers, and no customer may make up more than 15% of the total power usage (PG&E, n.d.-a). Multiple states have the same threshold, and it is colloquially known as the "15/15 rule" (Best et al., 2021). For residential electricity data, the area must include at least 100 homes' energy usage.

In California, researchers have their own path to request data on a case-by-case basis. They are not limited by the 15/15 rule, but they are limited in the volume of data they can collect and by various privacy rules. Researchers must have a faculty or staff affiliation at a nationally accredited non-profit college or university, and the research must be sponsored by a professor. The study must have IRB approval and researchers must sign a strict non-disclosure agreement (PG&E, n.d.-a) if the data provided is more granular than the aggregation thresholds assigned for public disclosure. All applications and decisions are posted publicly online (PG&E, n.d.-b).

Important tools and research have come out of this data access. The largest example is also based on the CPUC's rulemaking: University of California Los Angeles' Energy Atlas project, a map that displays energy consumption across Los Angeles county and the Bay Area by city, neighborhood, building type, age, type of energy, greenhouse emissions, and sociodemographic information. The Atlas is a tool that can be used for research, creating a baseline to compare energy improvements against and helping local governments coordinate climate action plans (UCLA California Center for Sustainable Communities, n.d.-a). After several NDAs and onerous legal fights, UCLA eventually gained access to complete user consumption data in many geographical areas, but it remained limited in what data and insights its tool could expose by the same 15/15 and 100-person aggregation rules that bound utilities (UCLA California Center for Sustainable Communities, n.d.-b). The project revealed, among other things, that buildings are responsible for 40% of greenhouse gas emissions in Los Angeles county.

Aggregation thresholds like the 15/15 rule are contentious. On one hand, they are easy for state regulators to understand and easy to implement. On the other hand, they can be arbitrary and may be too conservative in some cases or too liberal in others, needlessly preventing harmless research, or failing to protect some individuals, respectively. There are other methods that may allow for higher risk research, when the benefits outweigh the risk. For example, differential privacy offers a principled way to balance the competing objectives of having detailed data and protecting individual privacy. The Department of Energy sponsored research using other privacy-preserving techniques, such as differential privacy (Best et al., 2021). These techniques are less intuitive but more effective and flexible than the 15/15 rule.

Environment

The National Environmental Policy Act of 1969 (NEPA) requires the government to assess the environmental impact of any project that uses federal land, federal tax dollars, requires federal authorization, or is under the jurisdiction of a federal agency (Middleton, 2021). NEPA does not prohibit harming the environment. It requires only that the government identify and document potential environmental harm. If a project will have significant environmental impacts, the responsible agency must prepare a comprehensive environmental impact statement (EIS), which the public can then comment on (U.S. EPA, n.d.-b). EISs under NEPA usually contain four sections: a proposed action, a description of its effect on the environment, a range of reasonable alternatives, and an analysis of the environmental impact of each of the proposed actions and alternatives, including cumulative impacts, assessment models used to come up with those predictions, and a description of mitigation measures the project will take (Middleton, 2021; Wentz, 2016). Even projects that do not require full EISs may require a less thorough alternative called an environmental assessment, or an explanation for their exemption, called a Finding of No Significant Impact (U.S. EPA, n.d.-b).

EISs are used in over one hundred other countries (Caldwell, 1998), but scholars and activists vary widely on what policy ends they believe EISs can help realize. Some see EISs as a form of applied analytical science — agencies objectively evaluating a project with the scientific method, subjecting processes to falsification and demarcation between facts and value judgments (Cashmore, 2004; Morrisey, 1993). However, this fails to acknowledge that, unlike with science, there is no way to judge whether a decision to approve a project is "correct" (Willis, 1995).

Others see EISs strictly as a planning tool, meant to inform the public of the environmental impact of a given project (Ortolano & Shepherd, 1995). This view treats EISs as a political process where concerns of scientific input and environmental impact are overpowered by corporate and agency interest. A third, more optimistic way to frame EISs is as an exercise in civic science, a process still based in scientific principles and disprovability but flexible enough to be extended to political decision making processes, usually through public education and participation (Cashmore, 2004). The goal of EISs in this case is not to come up with objectively the best decision but to improve the quality of decisions made (Formby, 1990). Since every agency — and for that matter, every country — has its own way of assessing environmental impact, each of these three purposes may be at play at various levels within each.

In the EIS ecosystem, researchers play the role of critiquing methods and bringing in evidence to validate or falsify various claims. In addition, they often have different roles within each of these conceptualizations of EISs. For example, if the EIS process is an applied science, researchers can look at an EIS' proposed method of mitigating environmental harm and alternative methods, and evaluate the options (Marcot et al., 2001). This can be done while an EIS is being written and used for a proposal, or in hindsight, to inform how future EISs should be considered. Journalists can also play this role of third party fact-checker. ProPublica for example used EPA data (including EISs) and modeling software to find drastic underenforcement of the Clean Air Act, and how EPA under-enforcement turns certain areas into carcinogenic "sacrifice zones" (Younes et al., 2021).

If the EIS process is a planning tool for corporate and agency vested interests, then this type of research will not affect the planning process (Morgan, 2012). However, courts can overturn an EIS and the accompanying project approval if there are analytical errors or erroneous conclusions in an EIS, and this can directly affect the planning process.

At a more meta-level independent researchers can evaluate and improve the effectiveness of the environmental review process itself (O'Faircheallaigh, 2010). Researchers can take an outcomes-oriented approach by evaluating whether an EIS process led to an acceptable environmental impact. NGOs will sometimes use the EIS process to fill in a void of government oversight (Wentz, 2016). In theory, by making EISs available to researchers, what an environmentally beneficial EIS process looks like can be informed by the success and failures of real, historical and current EISs (Antonello & Howkins, 2020; Sadler, 1996).

This more policy-oriented framing often entails exploration and comparison of multiple EISs. Though the Environmental Protection Agency has been collecting EISs from across agencies since the passing of NEPA, those documents have only become available online in a usable way in the last decade. The EPA hosts the Environmental Impact Statement Database, which includes records of all EISs the EPA has received since 1987, all EPA comment letters on EISs since 2001, and PDF versions of all EISs the EPA has received since October 2012 (U.S. EPA, n.d.). However, the federal government is not the only steward of this data. The Northwestern University Transportation Library holds an even larger and older EIS collection — more than 33,000, dating back to 1969. The full text is digitized and searchable, even for the older records (Cole, n.d.). Northwestern's search function also allows for searching within documents, unlike the EPA's. Some states also have their own systems for local projects, such as the NYC City Environmental Quality Act Database, Minnesota Pollution Control Agency Archive, California Environmental Quality Act Database, and Massachusetts EPA Database. These databases vary in the functionality they offer (Wentz, 2016).

Even if searchable, EIS data does not come in a form that researchers can easily use. Documents are often long, filled with extraneous detail, and organized in a way that makes information retrieval a challenge even for a human reader. Agencies each have their own formats for EISs, and since they often contract out the work to third parties, there can be wide variation even within a format (Middleton, 2021). EISs must answer certain questions but do not have to be organized in any standardized way, either in form or in function. As Karrkainen put it, "[B]ecause EISs are produced on a sporadic, ad hoc, and largely project-specific basis, each document is a unique and self-contained universe of information." (2002, p. 23).

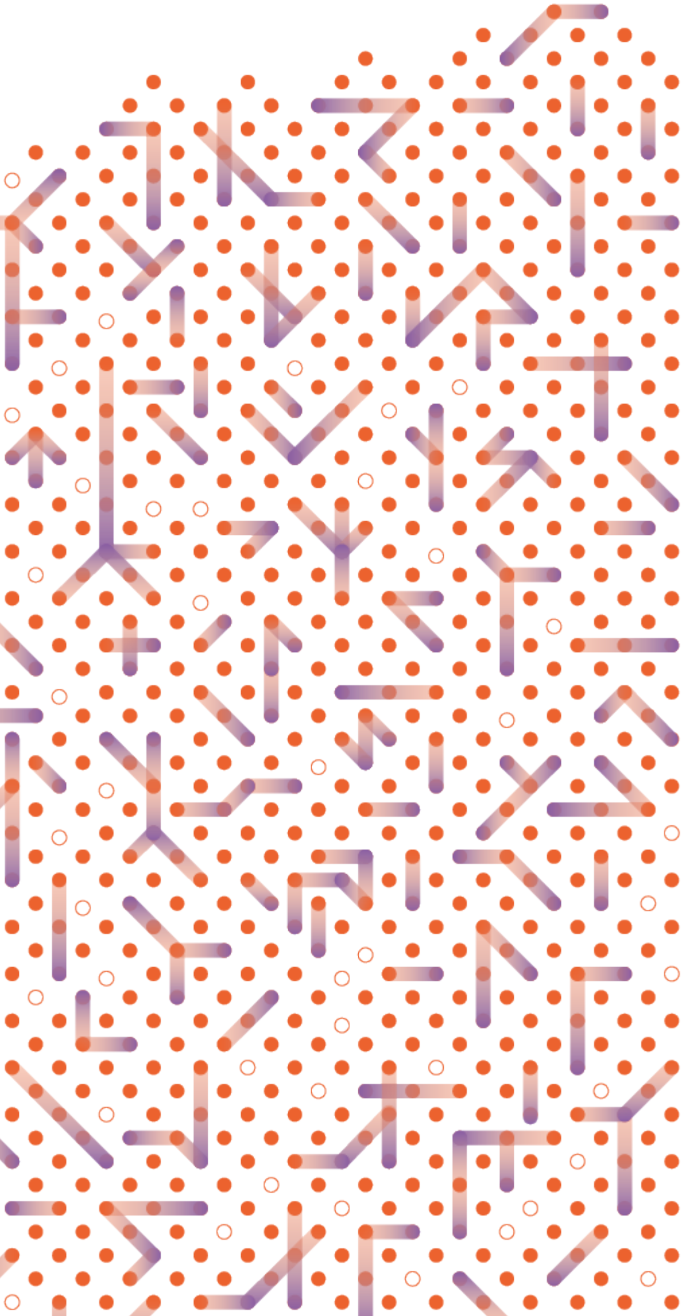
Assessments of environmental impact of federal projects do not risk individual privacy, but they could call on companies to provide information they consider to be trade secrets. In the United States, the government has dealt with this issue by excluding trade secret information from EISs. Courts have interpreted the term "trade secret" to include a lot of information about production processes (Lamdan, 2017). Agencies often have broad discretion to review trade secret information but that information rarely gets shared with the public (Morten, 2022), even if it is necessary for third parties to be able to accurately evaluate an action's impact on the environment. For instance, companies can withhold information such as what chemicals may be used in fracking (Schlanger, 2014) or in pesticides (U.S. EPA, n.d.-a) from their EISs. Many other environmental disclosure laws have broad carve outs for trade secrecy as well, such as the Clean Water Act; the Federal Insecticide, Fungicide, and Rodenticide Act; and the Emergency Planning and Community Right-to-Know Act (Dean, 2021).

European law takes a different approach to balancing the interests of trade secrecy and environmental protection. It treats environmental impact information as particularly important for the public to understand and have input on. In turn, its interpretation for what is protected as a trade secret is relative to its potential health and environmental impact. These principles come from the 1998 Aarhus Convention, an agreement from the United Nations Economic Commission for Europe that guarantees citizens the right to obtain certain environmental information (United Nations Economic Commission for Europe, 1999). The United Kingdom in particular based their environmental disclosure laws on the Aarhus Convention and created a separate framework for environmental transparency that doesn't rely on FOIA (Lamdan, 2017). Instead, it assumes a right to environmental information since it is germane to environmental health and applies a public interest test: authorities "can refuse to provide information only when the public interest in maintaining the exception outweighs the public interest in disclosure" (Information Commissioner's Office, 2022).

Ten Lessons Social Media Can Learn from Other Industries

Governments and private entities involved in clinical trials, electric smart meters, and environmental impact statements have dealt with policy questions about providing researchers with access to data for longer than social media has existed. They have had more time to develop laws, technology, and norms around sharing data. They have undergone the requisite back and forth between civil society, academia, government, and private industry to determine how much data can and should be shared with researchers. No sector manages these trade offs perfectly, and each of the sectors discussed here has glaring issues of its own, but their data sharing mechanisms are more mature and less tumultuous than social media's are now and are thus worth drawing lessons from.

In this section, we describe some of the broader patterns in what has worked well in these sectors' approaches to sharing data with researchers. This is not meant to be a set of precise policy proposals for governments or best practices for companies. Rather, it is a set of principles about what effective data sharing looks like elsewhere and could look like on social media.



1. Sharing data with researchers can help make more informed policy decisions

All three sectors discussed in this paper share data with researchers in a way that, at least to some extent, perpetuates a virtuous cycle of research and policy mutually informing one another. Shared clinical trial data helps researchers do impactful research while also allowing the FDA and other agencies to learn from their methodologies and results, leading to safer and more efficacious health products. Smart meter data helps researchers identify weak points in the grid, name-and-shame electricity-gluttonous actors and technologies, and overall ease the energy transition. And environmental impact statements tightly weave researchers into the political process, giving scientists and members of civil society leverage to push for more health and environmentally informed decisions.

Research on social media however has been more limited to uncovering new harms and understanding the scale of the problem, rather than evaluating interventions and generating new policy. When designing mechanisms to give researchers access to social media data, policymakers should consider designing processes and information pathways that let the results of research more directly inform policy.

2. Sharing data can let researchers double check otherwise unverifiable corporate claims

Social media companies sometimes respond to public and political scrutiny with self-regulation: for example, YouTube claimed it reduced borderline content in its recommendation algorithm after being criticized for sending users down radicalizing rabbit holes (Alexander, 2019); Facebook removed sensitive ad targeting categories after auditors concluded it hadn't done enough to protect people from discriminatory ads (Isaac & Hsu, 2021); and companies have responded to countless calls for privacy, interoperability, and security with suites of new policies and features. However, social media researchers often do not have the ability to validate whether these systems are doing as they claim.

Other sectors show a way forward here — clinical trial data shared in a way that is particularly designed to allow third-party researchers to stress-test and verify whether a medical product works. Environmental impact statements further shift the paradigm, allowing the public to have input before an intervention is rolled out, which can help researchers determine possible shortcomings or knock on effects and address them before they occur. Similarly, social media data could allow third-party researchers to verify or test claims companies make about their services; social media companies should focus on voluntarily sharing data that would allow this kind of testing, and when policymakers are considering requiring data sharing, they should aim to ensure researchers have access to this kind of data.

3. The "denominator problem" can be addressed without compromising privacy

Social media companies often do not provide basic, baseline information about the volume of data on a service in general or related to specific groups, such as the number of monthly active users by age group or the number of daily posts in Spanish on their services. This leaves researchers unable to "compare content frequencies between platforms, or compare frequencies on the same platform over time" (Shapiro et al., 2021, p. 25). Shapiro et al. calls this the "denominator problem," which makes it difficult for researchers to understand the context or effect size of their research as it relates to a given platform in its entirety (2021).

Data sharing practices from other sectors show baseline information meaningfully bolsters the utility of research and that it can be done without compromising individual privacy. When the California Public Utilities Commission shares smart meter data for a given area, for example, it includes all customers in that area, even if it is not at the per-building level of granularity. This allows researchers to draw conclusions about that geographical area and compare areas (e.g. UCLA California Center for Sustainable Communities, n.d.-a). Clinical trial data hosted on ClinicalTrials.gov offers even more specific denominator data, including breakdowns of trial data patient groups by gender, age, race, and a variety of health factors. Yet, at the same time, neither smart meter nor clinical trial data is at significant risk of being de-identified since they have basic aggregation and data minimization mechanisms, like the 15/15 rule and the 18 direct HIPAA identifiers. Social media companies should develop methods, similar to the 15/15 rule and others, that can provide researchers with useful baseline information about numbers of users on their platforms.

4. Addressing the "black box" problem will make research more widely applicable

Researchers argue that when they use APIs or data sets provided by social media companies, they have no sense of how that data got produced and what data may be missing or disproportionately represented (Tromble, 2021). Clinical trials show just how important this data is to researchers. Clinical trials, on YODA, ClinicalTrials.gov, and elsewhere all include metadata about how the data was generated, such as trial protocols and statistical analysis methods. The trial protocols describe in detail the trial's ways of measuring patient outcomes. This data lets researchers evaluate methodologies and understand the context data is produced in so they can better understand when/how to use it, and where that process should be criticized. Providing metadata to researchers about how social media data is generated will make research more accurate and detailed.

5. Transparency mechanisms let civil society serve as data sharing watchdogs

Data sharing requirements can in some cases allow third parties to evaluate how well companies are following those very requirements, which is important for meaningful transparency and accountability. This happens with EISs and the larger process of environmental impact reviews, though a better example is the FDAAA Trials Tracker, which uses ClinicalTrials.gov data to calculate how many covered trials have reported their results, what percent of trials have failed to report, and how many fines the U.S. government could impose, and how many it has imposed (so far, none pointing to the problem of a lack of enforcement) (FDAAA Trials Tracker, n.d.). This in turn has allowed civil society and other groups to criticize non-compliant research entities (e.g. Bioethics International, n.d.). Social media companies should make public information about the datasets that they share with researchers, including who the researchers are, the purpose of the study, what is being shared, results where available, etc., particularly where they are required to share such data by law.

6. Standards make shared data usable

Stable, documented, and widely adopted standards are key to making smart meter data and clinical trial data usable by researchers. Standards help researchers know what data to expect and how they can expect to receive it. For example, the FDA, NIH, and other regulators have helped establish standards around the structure and content of summary data and metadata associated with clinical trials.

By contrast, with social media, each company has its own protocol for sharing data. These have evolved significantly over time, with new versions of APIs making more, less, or different data available to researchers in unpredictable ways. Technical standards between social media sites are incompatible with one another, meaning that each social media company that supports researcher access to data via an API will do so using different methods. There is no standard that governs how such companies should develop an API for researchers in the first place. (The W3C attempted to create such a standard but was largely ignored by industry (Guy, 2017). This makes research that uses multiple platforms difficult if not impossible (Vogus, 2022b).[5] EIS data shows how if untreated, this problem can fester; environmental researchers struggle to compare data between EISs or even find relevant data with an EIS because they are so unstandardized (Karkkainen, 2002).

Standards also eventually make data generation cheaper and more streamlined. For example, when YODA first began, Johnson & Johnson had to pay upfront costs to reformat the data for any older or ongoing clinical trial that they shared data from. However, knowing that they would eventually need to share the data with YODA, Johnson & Johnson researchers quickly began to collect and organize the data in a

[5] A related issue is how social media companies report on activity on their platforms, which can be a source of data for researchers too. However, we *recognize* the challenge in developing shared categories for reporting activity across different social media platforms is made more difficult due to the genuine differences in how these companies count phenomena such as posts, users, actions taken against content, etc. (See for example (Keller, 2021)).

way that was compatible with YODA from the beginning, thus lowering the costs of running the program (Ross et al., 2018). Social media companies should invest in inclusive efforts to develop technical standards for APIs for researchers, which can be beneficial not just for the industry as a whole but the wider public.

7. Data sharing should be flexible enough to respond to public crises

Like social media, energy, health, and the environment are all critically important areas where knowledge needs to grow to improve the future of humanity. Social media data sharing often does not reflect those stakes. Social media companies have responded to some researcher demands for data for example, such as Meta sharing election ad targeting data (Jagadeesh et al., 2021), but researchers have argued that even this data is incomplete; Facebook has taken “several steps” to protect their users’ privacy in this data, but has not outlined what specifically all those steps are.

Though far from perfect, some data sharing governance mechanisms in other sectors are more flexible in their ability to share data when it can help realize a distinct public good. We see this in the United Kingdom’s approach to sharing environmental data — whether data is made available is determined by its relevance to environmental health and the public interest, and other factors like corporate secrecy are weighed less heavily for this important case. Normative tradeoffs can also be made in response to acute crises. ClinicalTrials.gov expedited and broadened its data sharing about COVID-19 vaccines, though many in the medical community called for even greater transparency than they actually provided (Morten et al., 2020). Social media companies should be flexible in how they design their data sharing rules and consider greater access (at the expense of their perceived internal costs) when the public interest is particularly important such as in the case of natural disasters, elections, etc.

8. Ease of understanding is a factor to consider in privacy

As discussed previously, social media companies tend to be opaque about the methods they use to ensure user privacy. When they do share higher level information though, often they point to complex technical solutions, such as differential privacy. Examples from other industries however show how privacy rules for preparing and sharing data can be intuitive, and easier to understand. The 15/15 rule with smart meter data and the 18 direct identifiers in HIPAA which cannot be shared in clinical trials, for example, are easier to implement and enforce.

However, simple privacy rules also have downsides. They may not achieve their desired privacy ends, since simple interventions don't tend to have the flexibility to respond to real life privacy threats. In clinical trials, patients could potentially be able to be identified by combining multiple types and sources of other data. Likewise, depending on other factors, the 15/15 rule may either insufficiently protect energy consumers' identities or overly protect their identities, and get in the way of researchers getting access to useful data. Neither of these is to say that all privacy rules should be maximally intuitive: it is only to say that intuitiveness and ease of understanding have benefits, and social media companies and policymakers should consider them as factors to trade off with other factors in any approach to privacy in data sharing.

9. Data access can be tailored to different researcher needs and risks

Social media companies share data through a small number of tiers of access — often they only offer one universal tier, but at most, they offer a free tier for the general public, a paid tier for enterprise, and an academic tier. Other sectors, however, more narrowly tailor the data they make available to the capabilities and motives of different actors by sharing data at several tiers of access. The CPUC, for example, offers different access to data depending on whether a researcher is from the government or an academic. Governments can request data for any area they have jurisdiction over, but have to follow the 15/15 rule; academics on other hand are not limited by the 15/15 rule, but they are limited in the volume of data they can collect. This reflects the different privacy concerns that relate to government (e.g. sharing data with law enforcement) and researchers. YODA even tailors data access within a system depending on the sensitivity of the data. YODA only lets researchers access individual patient data by logging onto a secure platform, and it does not let them download to their local machines. However, researchers can download metadata such as clinical trial protocols, which is less sensitive and poses no risk of de-identification.

Social media companies and policymakers should keep in mind that a "different strokes for different folks" approach to sharing data gives researchers the flexibility they need to engage in effective work, instead of having them only receive data available to the lowest common denominator.

10. Diverse data stewards offers new affordances

Every governance model for giving researchers access to data offers its own unique affordances, and the range of affordances available can be stretched by allowing more actors to share and hold data, both public and private. Social media does not benefit from this diversity because as of now, private companies are the sole stewards of data. In other sectors though, other organizations can share and organize data as they see fit. EIS data for example is organized both by government actors, such as the EPA, and academic actors, such as Northwestern. Each of these modes of organization allows data to be used in different ways, and researchers may be more interested in some or the other. Clinical trial data is also shared by multiple actors, through its voluntary and compulsory data sharing mechanisms.

Both approaches have benefits and drawbacks. Voluntary mechanisms like YODA give companies a chance to allow third parties to check their work without compromising privacy or trade secrecy. However, they also let companies easily hide data that makes them look bad. Compulsory mechanisms like ClinicalTrials.gov help shed light on processes companies might otherwise keep opaque. Even the knowledge that light might be shed on these processes could be enough to inspire good behavior. However, data providers may be adversarial in how they provide data, following the letter of the law but avoiding making themselves look bad. One example of this is EISs where agencies eager to push a project through may potentially downplay its environmental impact or the effectiveness of costly anti-pollution measures in an EIS.

Policymakers should not try to replace voluntary data sharing mechanisms with compulsory ones. Instead, they should foster a data sharing ecosystem where different approaches — compulsory, voluntary, or third-party research consortia — fill in each other's gaps.

References

Adams, J. N., Bélafi, Z. D., Horváth, M., Kocsis, J. B., & Csoknyai, T. (2021). How Smart Meter Data Analysis Can Support Understanding the Impact of Occupant Behavior on Building Energy Performance: A Comprehensive Review. *Energies, 14*(9), 2502. https://perma.cc/8KN5-A8UF

Alba, D. (2021, September 10). Facebook sent flawed data to misinformation researchers. *The New York Times*. https://perma.cc/54ME-28GP

Albert, A., & Rajagopal, R. (2013). Smart Meter Driven Segmentation: What Your Consumption Says About You. *IEEE Transactions on Power Systems, 28*(4), 4019–4030. https://perma.cc/K2FV-KK9V

Alexander, J. (2019, June 26). YouTube introducing changes to give people more control over recommended videos. The Verge. https://perma.cc/GU6H-HR8T

American Cancer Society. (n.d.). *Smart Meters*. American Cancer Society. Retrieved August 2, 2022, from https://perma.cc/3LW7-8BV5

American Council for an Energy-Efficient Economy. (2016, July). *Data Access Summary: State and Local Policy Database*. https://perma.cc/R572-L5UK

Anderson, M. L., Chiswell, K., Peterson, E. D., Tasneem, A., Topping, J., & Califf, R. M. (2015). Compliance with Results Reporting at ClinicalTrials.gov. *The New England Journal of Medicine, 372*(11), 1031–1039. https://perma.cc/779J-3BBX

Antonello, A., & Howkins, A. (2020). The rise of technocratic environmentalism: The United States, Antarctica, and the globalisation of the environmental impact statement. *Journal of Historical Geography, 68*, 55–64. https://perma.cc/C3WL-M7Y4

Best, C., Teehan, M., & Kim, J. (2021). *Energy Data Access: A Guide to Leveraging Differential Privacy* (Recurve). U.S. Department of Energy and National Renewable Energy Laboratory. https://perma.cc/AU9D-9GYG

Bioethics International. (n.d.). *Good Pharma Scorecard*. Bioethics International. Retrieved August 3, 2022, from https://perma.cc/MWW4-ZUTN/

Cadwalladr, C., & Graham-Harrison, E. (2018, March 17). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*. https://perma.cc/9CX6-8RMQ

Caldwell, L. K. (1998). *The National Environmental Policy Act: An Agenda for the Future. Indiana University Press.* Decision Establishing Building Decarbonization Pilot Programs, Rulemaking 19-01-011 Decision 20-03-027 (2020). https://perma.cc/7E9U-Q9BK

Cashmore, M. (2004). The role of science in environmental impact assessment: Process and procedure versus purpose in the development of theory. *Environmental Impact Assessment Review, 24*(4), 403–426. https://perma.cc/3EKC-Z265

Chen, H., Zhang, B., & Wang, Z. (2022). Hidden inequality in household electricity consumption: Measurement and determinants based on large-scale smart meter data. *China Economic Review, 71*, 101739. https://perma.cc/P7F2-59ZB

Citron, D. K. (2014). *Hate Crimes in Cyberspace*. Harvard University Press.

Clark, M. (2021, August 4). Research Cannot Be the Justification for Compromising People's Privacy. *Meta*. https://perma.cc/PP5E-ZPP3/

Cole, R. (n.d.). *Environmental Impact Statements: Getting Started*. Northwestern University Transportation Library. Retrieved August 3, 2022, https://perma.cc/69HL-ARK4

Commission for Energy Regulation. (2012). *CER Smart Metering Project—Electricity Customer Behaviour Trial, 2009-2010* (1st Edition). Irish Social Science Data Archive. https://perma.cc/3KXR-JQW2/

Cooper, A., & Shuster, M. (2021). *Electric Company Smart Meter Deployments: Foundation for a Smart Grid (2021 Update)*. Institute for Electric Innovation. https://perma.cc/W2B5-8DKM

Dean, M. (2021). An Environmental FOIA: Balancing Trade Secrecy with the Public's Right to Know. *California Law Review, 109*(6). https://perma.cc/GWA3-HWEU/

DeVito, N. J., Bacon, S., & Goldacre, B. (2020). Compliance with legal requirement to report clinical trial results on ClinicalTrials.gov: A cohort study. *The Lancet, 395*(10221), 361–369. https://perma.cc/4N27-4WKW

Digital Services Act, Article 31, COM/2020/825, European Commission, 52020PC0825 (2020). https://perma.cc/K46B-TJ83

Dixon, S. (2022, June). *Number of worldwide social network users 2018-2027*. Statista. https://perma.cc/6XJV-MK9P/

Doshi, P. (2015). No correction, no retraction, no apology, no comment: Paroxetine trial reanalysis raises questions about institutional responsibility. *BMJ, 351*, h4629. https://perma.cc/5CB5-XFUL

Duarte, N. (2015). The Home Out of Context: The Post-Riley Fourth Amendment and Law Enforcement Collection of Smart Meter Data. *North Carolina Law Review, 93*(4), 1140. https://perma.cc/RRP6-2GG4

Edelson, L., & McCoy, D. (2021, September 22). *How Facebook Hinders Misinformation Research*. Scientific American. https://perma.cc/S7KH-RT5R

Energy Independence and Security Act of 2007, 110–140 121 STAT. 1783 § 1301 (2007).

Facebook. (n.d.). *Facebook Open Research and Transparency Home*. Retrieved August 2, 2022, https://perma.cc/358T-6SN2

FDAAA Trials Tracker. (n.d.). *Who's sharing their clinical trial results?* Retrieved August 2, 2022, https://perma.cc/3WLU-SLEN

Food and Drug Administration. (2018, March 28). *What We Do*. FDA. https://perma.cc/5E8U-NKVU

Food and Drug Administration. (2019). *About Drugs@FDA*. FDA. https://perma.cc/9DZX-24GP

Formby, J. (1990). The Politics of Environmental Impact Assessment. *Impact Assessment, 8*(1–2), 191–196. https://perma.cc/J54T-YVGY

Franklin, S. B., Nojeim, G., & Thakur, D. (2021). *Legal Loopholes and Data for Dollars: How Law Enforcement and Intelligence Agencies Are Buying Your Data from Brokers*. Center for Democracy & Technology. https://perma.cc/6LK5-LHPA

Green Button Data Alliance. (2022). *Green Button Data*. Green Button Data. https://perma.cc/ATT8-2L5V

Guy, A. (2017). *Social Web Protocols*. W3C Working Group. https://perma.cc/B5WN-BJML

Halperin, R. (1979). FDA Disclosure of Safety and Effectiveness Data: A Legal and Policy Analysis. *Duke Law Journal, 28*(1), 286–326. https://perma.cc/WUT2-55SG

Harris, R. (2015, March 11). Results Of Many Clinical Trials Not Being Reported. *NPR*. https://perma.cc/22EY-KPUU

Hatmaker, T. (2021, August 4). Facebook cuts off NYU researcher access, prompting rebuke from lawmakers. *TechCrunch*. https://perma.cc/57PG-2HFB/

Hayes, J. L., Britt, B. C., Evans, W., Rush, S. W., Towery, N. A., & Adamson, A. C. (2021). Can Social Media Listening Platforms' Artificial Intelligence Be Trusted? Examining the Accuracy of Crimson Hexagon's (Now Brandwatch Consumer Research's) AI-Driven Analyses. *Journal of Advertising, 50*(1), 81–91. https://perma.cc/DE29-GMPF

Hearing on Platform Transparency: Understanding the Impact of Social Media. Subcommittee on Privacy, Technology, and the Law (2022, May 4). https://perma.cc/RP2W-GL3E

Herder, M., Morten, C. J., & Doshi, P. (2020). Integrated Drug Reviews at the US Food and Drug Administration—Legal Concerns and Knowledge Lost. *JAMA Internal Medicine, 180*(5), 629–630. https://perma.cc/DM9X-MKLR

Himeur, Y., Alsalemi, A., Bensaali, F., & Amira, A. (2020). Building power consumption datasets: Survey, taxonomy and future directions. *Energy and Buildings, 227*, 110404. https://perma.cc/X94F-CQ8W

Hooks, C. (2013, May 19). As Towns Say No, Signs of Rising Resistance to Smart Meters. *The New York Times*. https://perma.cc/C63V-KJHD

Imana, B., Korolova, A., & Heidemann, J. (2021). Auditing for Discrimination in Algorithms Delivering Job Ads. *Proceedings of the Web Conference 2021*, 3767–3778. https://perma.cc/3CEW-S8X5

Information Commissioner's Office. (2022, August 1). *When can we refuse a request for environmental information?* ICO. https://perma.cc/4QB9-5EB8/

Institute of Medicine. (2015). *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*. The National Academies Press. https://perma.cc/4BZG-GXC9

Isaac, M. (2022, May 23). Meta to give researchers more information on political ad targeting. *The New York Times*. https://perma.cc/W9R2-V9TB

Isaac, M., & Hsu, T. (2021, November 9). Meta plans to remove thousands of sensitive ad-targeting categories. *The New York Times*. https://perma.cc/TKU5-XRQG

Jagadeesh, K., Raje, R., Leone, I., Gruen, A., & Hillenbrand, B. (2021, February 10). *Introducing new election-related ad data sets for researchers*. Meta Research. https://perma.cc/3Q7H-46CD/

Jones, J. S. (2021, May 3). 75% of US households have smart meters – report. *Smart Energy International*. https://perma.cc/4MUU-HHSJ/

Kang, J., & Reiner, D. M. (2022). What is the effect of weather on household electricity consumption? Empirical evidence from Ireland. *Energy Economics, 111*, 106023. https://perma.cc/ZQ3A-VXQ6

Kapczynski, A., & Morten, C. J. (2021). The Big Data Regulator, Rebooted: Why and How the FDA Can and Should Disclose Confidential Data on Prescription Drugs and Vaccines. *California Law Review, 109*(2). https://perma.cc/Q344-F4XP/

Karkkainen, B. C. (2002). Toward a Smarter NEPA: Monitoring and Managing Government's Environmental Performance. *Columbia Law Review, 102*(4), 903–972. https://perma.cc/Z39Q-4N2W

Kayser-Bril, N. (2021, August 13). AlgorithmWatch forced to shut down Instagram monitoring project after threats from Facebook. *AlgorithmWatch*. https://perma.cc/BG2W-LCSA/

Keller, D. (2021, March 19). *Some Humility About Transparency*. https://perma.cc/K7CM-QCKY

Korang, S. K., Rohden, E. von, Veroniki, A. A., Ong, G., Ngalamika, O., Siddiqui, F., Juul, S., Nielsen, E. E., Feinberg, J. B., Petersen, J. J., Legart, C., Kokogho, A., Maagaard, M., Klingenberg, S., Thabane, L., Bardach, A., Ciapponi, A., Thomsen, A. R., Jakobsen, J. C., & Gluud, C. (2022). Vaccines to prevent COVID-19: A living systematic review with Trial Sequential Analysis and network meta-analysis of randomized clinical trials. *PLOS ONE, 17*(1), e0260733. https://perma.cc/UJK3-JWE6

Lamdan, S. (2017). Beyond FOIA: Improving access to environmental information in the United States. *Georgetown Environmental Law Review, 29*(3), 481–513. https://perma.cc/F477-D3DH

Lerner, J. (2008). Taking the "Long View" on the Fourth Amendment: Stored Records and the Sanctity of the Home. *Stanford Technology Law Review, 2008*. https://perma.cc/GUP4-9BK7

Li, S., Grijalva, S., & Chau, P. (n.d.). *Smart Meter Data Catalog* [JavaScript]. Georgia Tech. Retrieved August 3, 2022, https://perma.cc/D2WL-GU8V/

Loyola University Chicago. (n.d.). *18 HIPAA Identifiers: Information Technology Services: Loyola University Chicago*. Loyola University: Information Technology Services. Retrieved August 2, 2022, https://perma.cc/F6T8-MXEG/

Lynch, J., & Tien, L. (2010). *Joint Comments of CDT and EFF on Proposed Policies and Findings Pertaining to the Smart Grid*. Center for Democracy & Technology and Electronic Frontier Foundation. https://perma.cc/7DM7-TGK5/

Marcot, B. G., Holthausen, R. S., Raphael, M. G., Rowland, M. M., & Wisdom, M. J. (2001). Using Bayesian belief networks to evaluate fish and wildlife population viability under land management alternatives from an environmental impact statement. *Forest Ecology and Management, 153*(1), 29–42. https://perma.cc/AU85-3WVZ

Marwick, A., Clancy, B., & Furl, K. (2022). Far-Right Online Radicalization: A Review of the Literature. *The Bulletin of Technology & Public Life*. https://perma.cc/9X63-PNTC

Meta. (n.d.). *Stories Ads*. Meta for Business. Retrieved August 1, 2022, from https://perma.cc/Z88D-T8BU

Meta. (2021, June 22). *Facebook's five pillars of Responsible AI*. ML Applications | Research. https://perma.cc/G9EZ-RGZG/

Middleton, T. (2021, March 2). *What is an Environmental Impact Statement?* American Bar Association. https://perma.cc/AL9D-XYZX/

Mission::data. (2019). *Energy Data Portability: Assessing Utility Performance and Preventing "Evil Nudges."* https://perma.cc/AMW3-9VGL

Morgan, R. K. (2012). Environmental impact assessment: The state of the art. *Impact Assessment and Project Appraisal, 30*(1), 5–14. https://perma.cc/83L8-GKGU

Morrisey, D. J. (1993). Environmental impact assessment—A review of its aims and recent developments. *Marine Pollution Bulletin, 26*(10), 540–545. https://perma.cc/54MW-AFGB

Morten, C. (2022). Publicizing Corporate Secrets. *University of Pennsylvania Law Review, 171*. https://perma.cc/JTQ5-BMY2

Morten, C., Kapczynski, A., Krumholz, H. M., & Ross, J. S. (2020, March 26). *To Help Develop The Safest, Most Effective Coronavirus Tests, Treatments, And Vaccines, Ensure Public Access To Clinical Research Data*. *Health Affairs*. https://perma.cc/7Q7R-95CT/

Morten, C., Nicholas, G., & Viljoen, S. (forthcoming). *Researcher Access to Social Media Data: Lessons from the Clinical Trial Data Settlement*. Privacy Law Scholars Conference, New York, NY.

Mozilla. (2022, January 10). *Mozilla partners with The Markup to launch Rally study into Facebook's tracking and data collection practices*. The Mozilla Blog. https://perma.cc/EZ6C-E6CG/

National Center for Education Statistics. (2012). *Digest of Education Statistics, 2012*. Number of Public School Districts and Public and Private Elementary and Secondary Schools: Selected Years, 1869-70 through 2010-11; National Center for Education Statistics. https://perma.cc/3L7M-VJGK

National Council on Electricity Policy. (2008). *Demand Response and Smart Metering Policy Actions Since the Energy Policy Act of 2005: A Summary for State Officials*. U.S. Demand Response Coordinating Committee for The National Council on Electricity Policy. https://perma.cc/NTK2-AS87

National Institutes of Health. (2021, May). *ClinicalTrials.gov Background*. ClinicalTrials.Gov. https://perma.cc/8JJG-WHDQ

Nicholas, G. (2021, April 11). Be wary when Big Tech says it's defending your privacy. *Boston Globe*. https://perma.cc/T6KZ-H6ER/

NYU Cybersecurity for Democracy. (n.d.). *Ad Observer*. Ad Observer. Retrieved August 2, 2022, https://perma.cc/B9WL-SVSQ/

O'Faircheallaigh, C. (2010). Public participation and environmental impact assessment: Purposes, implications, and lessons for public policy making. *Environmental Impact Assessment Review, 30*(1), 19–27. https://perma.cc/M9G5-APEJ

Ortolano, L., & Shepherd, A. (1995). Environmental Impact Assessment: Challenges and Opportunities. *Impact Assessment, 13*(1), 3–30. https://perma.cc/8QHD-DGWY

Pappas, V. (2022, July 27). *Strengthening our commitment to transparency*. Newsroom | TikTok. https://perma.cc/B8Y6-NDDJ

Pariser, E. (2012). *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think* (Reprint edition). Penguin Books.

Persily, N., & Tucker, J. A. (2020). Conclusion: The Challenges and Opportunities for Social Media Research. In *Social Media and Democracy: The State of the Field, Prospects for Reform* (p. 313). Cambridge University Press. https://perma.cc/Z6KK-SND5

Pew Research Center. (2021, April 7). Social Media Fact Sheet. *Pew Research Center: Internet, Science & Tech*. https://perma.cc/C6W5-U4X6/

PG&E. (n.d.-a). *Energy Data Request User Guide*. https://perma.cc/HL78-QDWV

PG&E. (n.d.-b). *User Requests | PG&E's Energy Data Request Portal*. Energy Data Request Program. Retrieved August 3, 2022, from https://perma.cc/9GH2-MQFB

Piller, C. (2020, January 13). FDA and NIH let clinical trial sponsors keep results secret and break the law. *Science*. https://perma.cc/A8UR-VM56

Pillitteri, V. Y., & Brewer, T. L. (2014). Guidelines for Smart Grid Cybersecurity. *NIST*. https://perma.cc/8RQM-G7Q5

Protalinski, E. (2011, October 12). *Facebook: Releasing your personal data reveals our trade secrets*. ZDNet. https://perma.cc/E89K-8HBS/

Ross, J. S., Waldstreicher, J., Bamford, S., Berlin, J. A., Childers, K., Desai, N. R., Gamble, G., Gross, C. P., Kuntz, R., Lehman, R., Lins, P., Morris, S. A., Ritchie, J. D., & Krumholz, H. M. (2018). Overview and experience of the YODA Project with clinical trial data sharing after 5 years. *Scientific Data, 5*(1), 180268. https://perma.cc/WT2M-PFZK

Ross, J. S., Waldstreicher, J., & Krumholz, H. M. (2019, November 18). Sharing clinical trial data: Lessons from the YODA Project. *STAT*. https://perma.cc/V8JM-FZUD/

Roth, Y., & Gadde, V. (2021, December 2). *Expanding access beyond information operations*. https://perma.cc/V98L-W9RW

Sadler, B. (1996). *Environmental Assessment in a Changing World: Evaluating Practice to Improve Performance*. Canadian Environmental Assessment Agency. https://perma.cc/3K8P-B7ZV

Schlanger, Z. (2014, August 13). There's Still a Lot We Don't Know About Fracking Chemicals. *Newsweek*. https://perma.cc/GC82-RH25

Segerberg, A., & Bennett, W. L. (2011). Social Media and the Organization of Collective Action: Using Twitter to Explore the Ecologies of Two Climate Change Protests. *The Communication Review, 14*(3), 197–215. https://perma.cc/UT9F-LRCC

Shapiro, E. H., Sugarman, M., Bermejo, F., & Zuckerman, E. (2021). *New Approaches to Platform Data Research*. NetGain Partnership. https://perma.cc/4GVU-59DQ

Sharfstein, J. M., Miller, J. D., Davis, A. L., Ross, J. S., McCarthy, M. E., Smith, B., Chaudhry, A., Alexander, G. C., & Kesselheim, A. S. (2017). Blueprint for Transparency at the U.S. Food and Drug Administration: Recommendations to Advance the Development of Safe and Effective Medical Products. *The Journal of Law, Medicine & Ethics, 45*(2), 7–23. https://perma.cc/HST6-UCYV

Stop Smart Meters! (2011). CA Local Governments On Board. *Stop Smart Meters!* https://perma.cc/VSU5-43GG/

Strong, D. R. (2017). *The Early Diffusion of Smart Meters in the US Electric Power Industry*. https://perma.cc/LG6J-MKEK

Teather, D. (2022). *TikTokAPI* (5.2.2) [Python]. https://perma.cc/7AFN-47GM (Original work published 2019)

Timberg, C. (2021, September 10). Facebook admits it bungled data it shared with social scientists. *Washington Post*. https://perma.cc/N4H5-NW36/

Torriti, J. (2020). *Appraising the economics of smart meters: Costs and benefits.* Routledge, Taylor & Francis Group.

Tromble, R. (2021). Where Have All the Data Gone? A Critical Reflection on Academic Digital Research in the Post-API Age. *Social Media + Society, 7*(1), 2056305121988929. https://perma.cc/RFT6-V24P

Twitter. (n.d.-a). *Getting Started with the Twitter API*. Retrieved August 2, 2022, from https://perma.cc/9NCX-FMMY

Twitter. (n.d.-b). *Twitter API for Academic Research*. Twitter Developer Platform. Retrieved August 2, 2022, from https://perma.cc/JSZ2-QRXP

UCLA California Center for Sustainable Communities. (n.d.-a). *UCLA Energy Atlas—About*. Retrieved August 3, 2022, https://perma.cc/DH4M-NMCQ

UCLA California Center for Sustainable Communities. (n.d.-b). *UCLA Energy Atlas—Methods*. Retrieved August 3, 2022, https://perma.cc/52AX-7D25

United Nations Economic Commission for Europe. (1999). Convention on Access to Information, Public Participation in Decision-Making and Access to Justice in Environmental Matters. *International Legal Materials, 38*(3), 517–533. https://perma.cc/UP9F-UVPC

U.S. EPA. (n.d.-a). *Limitations on Disclosure of Information under Pesticide Law* [Reports and Assessments]. United States Environmental Protection Agency. Retrieved August 3, 2022, https://perma.cc/9YCH-9CAA

U.S. EPA. (n.d.-b). *National Environmental Policy Act Review Process* [Overviews and Factsheets]. United States Environmental Protection Agency. Retrieved August 3, 2022, https://perma.cc/VMC6-8KM3

U.S. EPA. (n.d.). *Environmental Impact Statement (EIS) Database*. United States Environmental Protection Agency. Retrieved August 3, 2022, https://perma.cc/3DUL-V4CH

Van Loo, R. (forthcoming). Privacy Pretexts. *Cornell Law Review*. https://perma.cc/U4QK-2LPL

Viljoen, S. (2022, April 29). *(Civil) Libertarianism and the Legal Constitution of Social Data*. Freedom of Expression Scholars Conference 10, Yale Law Review. https://perma.cc/H4HV-JYK4

Vogus, C. (2022a, April 21). Independent Researcher Access to Social Media Data: Comparing Legislative Proposals. *Center for Democracy and Technology*. https://perma.cc/3Q7D-UNP8/

Vogus, C. (2022b). *Improving Researcher Access to Digital Data: A Workshop Report*. Center for Democracy & Technology. https://perma.cc/2H28-7277

Vogus, C., & Llansó, E. (2021). *Making Transparency Meaningful: A Framework for Policymakers*. Center for Democracy & Technology. https://perma.cc/T5DX-CLPH

Wallach, J. D., Wang, K., Zhang, A. D., Cheng, D., Nardini, H. K. G., Lin, H., Bracken, M. B., Desai, M., Krumholz, H. M., & Ross, J. S. (2020). Updating insights into rosiglitazone and cardiovascular risk through shared data: Individual patient and summary level meta-analyses. *BMJ, 368*, l7078. https://perma.cc/CFV5-JU76

Wentz, J. (2016). *Using Online Databasing to Unlock the Full Value of Environmental Impact Assessment* (SSRN Scholarly Paper ID 2897050; Issue ID 2897050). Social Science Research Network. https://perma.cc/3CNT-MDXH

Willis, K. G. (1995). Judging Development Control Decisions. *Urban Studies, 32*(7), 1065–1079. https://perma.cc/6DQ3-5WNH

Wolfe, S. (2004, December 17). *Public Citizen to Call on FDA to Ban Celebrex and Bextra*. Public Citizen. https://perma.cc/X2RZ-4PFG/

Yale University Open Data Access Project. (n.d.-a). *Data Use Agreement* [The YODA Project]. Retrieved August 2, 2022, https://perma.cc/4S5Z-YY8M

Yale University Open Data Access Project. (n.d.-b). *Frequently Asked Questions (FAQs)*. The YODA Project. Retrieved August 2, 2022, from https://perma.cc/HB95-ZF2Z

Yale University Open Data Access Project. (n.d.-c). *Welcome to the YODA Project*. The YODA Project. Retrieved August 2, 2022, https://perma.cc/7YM8-C9PH

Yale University Open Data Access Project. (2022, July 1). *Submitted Requests to Use Johnson & Johnson Data*. The YODA Project. https://perma.cc/FNJ5-PZ3B

Yin, L., & Sankin, A. (2021, April 9). *Google Blocks Advertisers from Targeting Black Lives Matter YouTube Videos*. The Markup. https://perma.cc/TF4B-9NB6

Younes, L., Kofman, A., Shaw, A., & Miller, M. (2021, November 2). *Poison in the Air*. ProPublica. https://perma.cc/RPX9-8B5B

YouTube. (n.d.-a). *YouTube for Press* [YouTube Official Blog]. Blog.Youtube. Retrieved August 1, 2022, https://perma.cc/F9AK-A2UP/

YouTube. (n.d.-b). *YouTube Research*. YouTube. Retrieved August 2, 2022, from https://perma.cc/J9M4-LAT5/

Zheng, J., Xu, Z., Zhong, H., Zhou, X., & Xia, Y. (2019). Optimizing Battery Capacity Based on Smart Meter Data in Battery Energy Storage System. *2019 3rd International Symposium on Autonomous Systems (ISAS)*, 457–461. https://perma.cc/ZP2T-WJU8

cdt.org

cdt.org/contact

Center for Democracy & Technology
1401 K Street NW, Suite 200
Washington, D.C. 20005

202-637-9800

@CenDemTech

CENTER FOR
DEMOCRACY
& TECHNOLOGY