# Beyond Competition: Designing Data Portability to Support Research on the Digital Information Environment *

Zeve Sanderson
znsz02@nyu.edu

February 26, 2024

**Abstract**

In this paper, I aim to situate data portability within the evolving discussions of how to support data access for researchers. More specifically, I explore how, given changes in the digital information environment, data donations enabled by portability requirements provide promising opportunities for facilitating research that is aligned with ethical and legal frameworks. I use generative AI as a case study for how data donations can support urgent research agendas on digital platforms. I then discuss current challenges for using data donations for research and provide recommendations for better aligning portability mechanisms with research. Taken together, I argue that, although portability is often considered through a competition lens, policymakers and companies should understand its potential impact on policy-relevant research efforts and ensure that portability can support research on the impacts of digital platforms and services.
***Keywords:*** *Data Portability; Data Access; Internet Policy; Platform Transparency*

*Thanks to the Data Transfer Initiative for support of this work, as well as the participants of the Data Transfer Summit for their generous feedback.

# Contents

# 1 Introduction

A key concern for policymakers, journalists, civil society organizations, and academics alike is understanding the myriad impacts of digital platforms, which have come to play a central role in social interactions, economic activities, and the dissemination of information. However, a recurring challenge has been that the digital trace data necessary to produce rigorous evidence on platform effects are stored in proprietary databases, often accessible only to the platforms themselves and used for commercial applications (Lazer et al., 2020; Persily & Tucker, 2020). This dynamic enables platforms to act as gatekeepers for both academic research agendas and evidence-based policy evaluations, leaving key questions of societal import unanswered and unanswerable given lack of data (Ausloos & Veale, 2020; de Vreese & Tromble, 2023). Alarmingly, a number of platforms—such Facebook (Freelon, 2018), Twitter (Kharpal, 2023), and Reddit (Gallagher, 2023)—have shut down public APIs in recent years, erecting significant barriers for independent researchers to collect requisite data.

Policymakers have made data access a central concern for efforts to increase platform transparency, oversight, and accountability. In the European context, the Digital Services Act (DSA), which is primarily concerned with platform transparency and user protection, includes provisions to grant access to data from very large online platforms (VLOPs) to vetted researchers (Commission, 2023). In the U.S. context, the Platform Accountability and Transparency Act (PATA) has been introduced, which includes similar mechanisms for requiring independent data access. While promising, these approaches to data access have a number of key limitations, most notably their narrow application to VLOPs. This limitation is especially important given recent developments in the digital information environment, such as the rise of socially important but smaller platforms that do not reach DSA or PATA usage thresholds (Ortiz-Ospina, 2019). The timeline for full DSA implementation, including comprehensive data access for vetted researchers under Article 40, is also not fully known.

Researchers have developed a number of other mechanisms for collecting data, such as web scraping and web tracking (Ohme et al., 2023). A key challenge for collecting data without user or platform consent is that it introduces potential legal risks for researchers and ethical risks for users (Fiesler, Beard, & Keegan, 2020). Within this context, one promising approach is data donations in which users consent to donating digital trace data for research. In addition to establishing user consent, data donations fall within legal data portability provisions, such as those in the EU General Data Protection Regulation (GDPR) and the proposed ACCESS Act in the U.S., and thus provide legal protections for researchers engaging in research on digital platforms. However, data portability, or the right for users to transfer their data from one digital service to themselves and/or to another digital service, has generally been considered through the lens of competition (Castro, 2021; Gulati-Gilbert & Seamans, 2023). This has led to a mismatch between data portability as a mechanism to promote competition in the digital marketplace and a mechanism to collect user data to facilitate research on the digital information environment. On the one hand, poli-

cymakers and platforms have approached the design, implementation, and evaluation of data portability through the lens of competition. On the other, researchers have leveraged data portability provisions for research, but often with challenges due to this misalignment between the needs of competition and research.

In this paper, I aim to situate data portability within the evolving discussions of how to support data access for researchers. More specifically, I explore how, given changes in the digital information environment, data donations enabled by portability requirements provide promising opportunities for facilitating research that is aligned with ethical and legal frameworks. I use generative AI as a case study for how data portability can support both platform competition and transparency. I then discuss current challenges for using data donations for research and provide recommendations for better aligning portability mechanisms with research. Taken together, I argue that, although portability is often considered through a competition lens, policymakers should understand its potential impact on policy-relevant research efforts and ensure that portability can support research on the impacts of digital platforms and services.

## 2 Beyond the Streetlight: Data donations in a multi-platform digital information environment

A challenge for researchers studying the digital information environment is that research agendas have been, to a certain extent, shaped by the data made available to them (Matamoros-Fernández & Farkas, 2021). The clearest impact of data availability is the amount of research undertaken on Twitter: Twitter is over-represented in research not because it is seen by scholars as the most important platform for political or social outcomes, but because its easily accessible API enabled the collection of granular, dynamic, and networked datasets that could support a wide range of research projects (Persily & Tucker, 2020). For example, a stark illustration of the agenda-setting power of Twitter's API is that the number of studies on Twitter in communications journals surpasses studies on YouTube (Lukito et al., 2023), even though YouTube has remained the most popular social media platform among U.S. adults for multiple years (Auxier & Anderson, 2021). The dynamic of data availability impacting research agendas—colloquially referred to as the streetlight effect—has led to significant blind spots in our understanding of the digital information environment (Moritz, 2016).

Scholars have engaged in a number of data collection strategies to facilitate a broader research agenda on digital platforms. Borrowing from Ohme et al. (2023), there are two approaches to collecting platform data. In a *platform-centric approach*, data is collected directly from platforms without the involvement of users. Examples of this approach include the use of Application Programming Interfaces (APIs), both documented and undocumented (Yin, 2023), and webscraping. Within a platform-centric approach, there are a number of specific data collection strategies, each which come with their own trade-offs. APIs, while often providing access to large structured data collections, are subject to deprecation by

platforms (Bruns, 2019; de Vreese & Tromble, 2023; Freelon, 2018) and have potential biases (Allen, Mobius, Rothschild, & Watts, 2021; Ruths & Pfeffer, 2014). Webscraping can be a powerful tool for collecting large scale data, but introduce significant legal and ethical risks (Fiesler et al., 2020; Krotov, Johnson, & Silva, 2020). Collaborations with platforms, though able to support ambitious projects for select researchers (Kupferschmidt, 2023), have introduced issues of researcher independence (Wagner, 2023) and accessibility (Walker, Mercea, & Bastos, 2019). Notably, a platform-centric approach has largely dominated policy discussions around data access (Persily, 2021), with legal mandates through the DSA structured around researchers being able to request data directly from VLOPs (Husovec, 2023). But are there other mechanisms for policymakers to support independent researcher data access?

In a *user-centric approach*, researchers directly involve the user in data collection; two main strategies are browser plug-ins (Haim & Nienierza, 2019) and data donations (Prainsack, 2019). While browser plug-ins (custom software that is able to capture data from a person's browser) can be a powerful tool for data collection, they are technically challenging to build and often tailored for the specific research project (Breuer, Kmetty, Haim, & Stier, 2023). For example, two recent papers on Google Search,[1] both published in *Nature*, developed and used different browser plug-ins to collect search results (Aslett et al., 2023; Robertson et al., 2023). However, a key reason that browser plug-ins are not the focus of this analysis is that they are not well-suited to a policy intervention.[2] While plug-ins collect data directly from a user's browser, data donations require that users be able to download their data from platforms. Data access rights through GDPR grants users the ability to download data from the digital services and platforms they use, as well as mandates that platforms provide the ability to do so (De Hert, Papakonstantinou, Malgieri, Beslay, & Sanchez, 2018; Mondschein & Monda, 2019). In addition to transferring personal data to another online platform or service that someone might use, these data can be donated to researchers for secondary use. Indeed, data donations enabled by GDPR's data access rights have already been used in a number of studies (Boeschoten, Ausloos, Moeller, Araujo, & Oberski, 2020; Halavais, 2019; van Driel et al., 2022).

To be clear, significant trade-offs are present with any approach to data collection based on the particular research question (Ohme et al., 2023; Pfiffner & Friemel, 2023), and data donations are far from a panacea. However, given the platform-centric orientation of policy interventions that aim to increase data access, it is important to note that data donations have a number of characteristics that make this strategy promising for both researchers studying the digital information environment and policymakers working on transparency efforts.

First, data donations allow participants to donate data from multiple platforms in the

[1] I am a co-author on Aslett et al. (2023)

[2] There are no policy proposals, to my knowledge, that would require the development of browser plug-ins, and it seems unlikely that this would become a focus for policymakers or regulators. The one related area where government involvement could be useful is funding shared infrastructure and tooling, such as the recent NSF-funded National Internet Observatory (see `https://nationalinternetobservatory.org/`). However, this falls outside of the scope of this paper.

same study, enabling a richer and more comprehensive view into their online information diets. This capability is especially important given that people increasingly use multiple platforms, (Auxier & Anderson, 2021; Krishnan, 2023), in particular young people (Anderson & Jiang, 2023). It also allows donations from platforms that do not surpass the size threshold to be classified as VLOPs under the DSA, but are nonetheless important for understanding social and political outcomes. These include alt platforms (e.g., Gab or Parler), local platforms (e.g., Nextdoor), video game platforms (e.g., Twitch), and messaging apps (e.g., Telegram).[3]

Second, while some research questions only require digital trace data *per se*, others require researchers be able to collect digital trace data and survey data in order to connect the online and offline—the relationship between online activity and demographic, behavioral, and attitudinal measures (Salganik, 2019). For example, a key area of interest for both scholars and policymakers is the impact of social media on mental health. To study this phenomenon, it is likely that researchers would need to both directly observe a user's social media behavior and collect survey responses to evaluate shifts in mental health outcomes; it is also likely that researchers would need to use both of these methods longitudinally. Similar questions of societal import, such as how online (mis)information impacts support for democratic institutions, would also require pairing of survey and digital trace data. Data donations could serve as a key mechanism for being able to collect digital trace data directly from study participants.

Third, there are a number of online harms that are not common and are not randomly distributed across the population, but instead occur unevenly in sub-populations. Ronald E. Roberston refers to this dynamic as "uncommon yet consequential online harms" (Robertson, 2022). For example, previous research has shown that misinformation consumption (Grinberg, Joseph, Friedland, Swire-Thompson, & Lazer, 2019) and sharing (Guess, Nagler, & Tucker, 2019) are concentrated in small portions of the American public, that hate speech is produced by a small minority of online users (Zannettou, ElSherief, Belding, Nilizadeh, & Stringhini, 2020), and radical content is consumed by a small percentage of online news consumers (Hosseinmardi et al., 2021). Similarly, certain sub-populations may be targeted more by online harms, such as Spanish-language communities in the U.S. (Sanchez & Bennett, 2022). These patterns mean that large data collections through platforms may not capture the so-called "long tails" of distributions where specific harms are concentrated. Welles (2014) reminds us that "Big Data researchers must choose to examine very small subsets of otherwise large datasets." One way of doing so is recruiting study participants who are in the sub-populations of interest and collecting data donations, such as a recent bilingual panel of Latinos in the U.S. that pairs survey data with digital data donations (Abrajano et al., 2022).

Finally, data donations include the explicit consent of users who donate data (Boeschoten

[3]Somewhat ironically, one of the reasons that data sharing mandates in the DSA and PATA are only applied to the largest online platforms is the potential anti-competitive effects of enacting onerous requirements on smaller platforms that may not have the resources for compliance (Keller, 2022). However, portability, which is primarily seen as competition-promoting, has the potential to enable research on these smaller platforms.

et al., 2020; Halavais, 2019; van Driel et al., 2022). Many users see their own digital trace data as potentially sensitive (Hemphill, Schöpke-Gonzalez, & Panda, 2022), are unaware of its use in research (Fiesler & Proferes, 2018), and have different levels of comfort based on the goal of the study (Gilbert, Vitak, & Shilton, 2021). Whereas the data made available through the DSA may not involve the explicit consent of users whose data are included, data donations directly involve the user and require informed consent (Crutzen, Ygram Peters, & Mondschein, 2019).[4] Data donations also fall within the legal regimes that establish user data access rights (Boeschoten et al., 2020; De Hert et al., 2018), thus avoiding a number of the legal risks for researchers that have accompanied methods like webscraping.

### 2.1 Case Study: Generative AI

Generative AI in general and chatbots in particular provide a useful case study for how portability can be leveraged to further both competition and transparency.

Competition among user-facing chatbots was initially focused primarily on model performance, as models did not learn from previous interactions and thus were not tailored to users. In this context, competition was oriented around model quality, and there was not significant friction associated with changing between chatbots (Riley, 2023).[5] However, as memory has to be built into chatbots,[6] chatbots are able to learn user preferences and thus potentially better serve user interests over time. In turn, this change has introduced classic competition dynamics for digital services—namely, that a platform or service becomes more *useful with use*. This dynamic introduces barriers to switching between chatbots, and data portability has been identified as a potential solution to support competition in this new market (Riley, 2023).

Since the introduction of ChatGPT in November 2022, understanding the impacts of user-facing generative AI has become a key question for both policymakers and academics. Thus far, red-teaming and auditing have been the main approaches for identifying potential risks associated with generative AI. Red-teaming is a technical approach that simulates attempts to circumvent a systems rules and identifies the conditions under which the system fails. AI audits, which serve a distinct but complimentary function, are assessments of AI systems to ensure they adhere to established ethical principles, legal standards, and technical guidelines.

While both of these approaches are necessary for understanding the potential risks of AI models, they are abstracted away from actual user behavior, leaving key foundational questions unanswered (Friedler, Singh, Blili-Hamelin, Metcalf, & Chen, 2023; Sanderson & Tucker, 2023). Who uses chatbots? How often are they used and for what tasks? Given

[4]To be clear, data donations may contain information from other users who did not provide consent, and so privacy and ethical considerations are still present. However, this data collection approach at least involves the informed consent of the person donating data, which is not involved in many other approaches.

[5]There were some reasons for staying on a particular chatbot, such as having easy access to input-output history. However, these anti-competitive dynamics were relatively limited compared to competition challenges in the context of other digital platforms and services.

[6]For more information, see `https://openai.com/blog/memory-and-new-controls-for-chatgpt`

that more than 60 global elections will cover roughly half of the world's population in 2024, of particular importance is understanding whether people use chatbots for political information and, if so, the impacts of this behavior. To understand user behavior, data portability could be an important data collection mechanism, as DDPs would provide researchers with the ability to collect user data. It is very unlikely that, in the near term, many chatbots will pass the VLOP threshold that would give researchers access to data under Article 40 of the DSA. As a result, without other mechanisms for data collection, we will be left in the dark about how people are using these systems and to what effects. Relative to social media or messaging services, data portability for chatbots also has more limited privacy risks, as a single user's input-output history does not (yet) include date from other users.[7] While a number of chatbots allow for download of chat history, such as OpenAI,[8] the ability to do so is voluntary and not every chatbot currently has an export feature. As Riley (2023) notes, "Openness and effective portability aren't the same thing." Similarly, portability that serves the purposes of competition and of research aren't the same thing. As regulators work to both increase competition and bring transparency and accountability to the generative AI market, designing portability for both purposes has the potential to create significant public benefits.

## 3 Limitations of Portability for Data Donations

While there are number of scientific, ethical, and legal benefits to using data donations for the study of the digital information environment, key challenges have limited researchers' ability to use data donations. There are three stages to a data donation study. The first is a consideration stage in which potential participants are provided with information about the study, such as the topic of the research and details about participation, and decide whether they will participate. The second is the donation stage in which consenting participants donate their data. And finally, the third is the analysis stage in which researchers are able to use donated data.

The first stage requires users to consent to participate, and previous work has measured the individual-level characteristics associated with willingness to participate in data donation studies (Pfiffner & Friemel, 2023). While the ability for data donation is dependent on the right of access that regulations like GDPR have established, the consideration stage is determined by an individual's willingness to donate data and it is not clear how policymakers could (or should) influence an individual's willingness to participate in research. As a result, this stage does not directly involve new policy questions and so I will focus on the challenges that impact the next two stages, and how policymakers and regulators could potentially better align data portability with the needs of researchers.

[7]That said, there has been reporting that suggests people are using chatbots for tasks with potential privacy concerns for data portability, such as writing performance reviews or creating resumes; e.g., https://www.mckinsey.com/featured-insights/themes/can-generative-ai-help-you-deliver-better-feedback.

[8]For more information, see https://help.openai.com/en/articles/7260999-how-do-i-export-my-chatgpt-history-and-data

The donation stage requires a study participant to request a data download package (DDP) from a digital platform or service. This process involves a high level of digital literacy, potentially impacting the representativeness of the study sample. Indeed, one study actually invited participants to a facility to support them with the data donation process (Kmetty & Németh, 2022). Even for users with requisite digital literacy, complex tasks in a research project may lead to attrition among those who expressed willingness to participate and so require clear instructions and ongoing participant support (which still might not be enough to mitigate attrition) (Breuer, Bishop, & Kinder-Kurlanda, 2020; Ohme, Araujo, de Vreese, & Piotrowski, 2021). Another challenge is that users generally need to download DDPs directly to their device. Given the size of these files, participants likely need to have access to a desktop and high-speed internet. In turn, this may limit the ability for data donations among users who do not have access to a desktop or high-speed internet, leading to within- and between-country variations in the ability to download DDPs. Finally, researchers need to implement a technically secure donation process. While there are some projects that aim to support data donation,[9], researchers often need to create their own implementation of the donation process (Ausloos & Veale, 2020), limiting such study designs to scholars with the technical expertise to do so.

The analysis stage requires that researchers have access to documented, structured data in machine-readable formats (Ohme et al., 2023). Previous research using DDPs have shown that data structures were unclear (e.g., posts showing up multiple times) and metadata categories were not well-documented in DDPs, leading to confusion about how to transform data for analysis and measure key concepts (van Driel et al., 2022). At best, these challenges require significant work from researchers to clean and transform data for analysis; at worst, these challenges make some data impossible to use for research given the lack of clarity.

## 4 Recommendations & Discussion

Data donations are a powerful mechanism for researchers to collect multi-platform digital trace from consenting users, leveraging data access rights established by regulations with portability provisions. While regulations like the DSA have platform-centric data access provisions, data donations are potentially better aligned with a number of compelling research questions of scholarly and public interest. However, significant challenges currently limit the ability for researchers to utilize data donations; in this context, there are a number of ways that regulators and companies should work with researchers to better align data donation processes with the needs of research on the digital information environment.

First, given that regulation is creating incentives for portability, companies are investing in portability systems to transfer data to similarly positioned companies. In their efforts, companies should talk to researchers to ensure that these portability systems can transfer data to researchers as well. One particularly effective mechanism, depicted in Fig-

[9]For example, see the Data Donation Module GitHub Repo: `https://github.com/uzh/ddm`

ure 1, would be direct data donations via transfers from a data host straight to a researcher data store, which avoids the complexities of asking users to download and upload. The ease of use would improve sample quality and decrease attrition; the direct transfer would remove the need for participants to have the device storage or bandwidth necessary for large data downloads; and the common infrastructure would increase the accessibility of engaging in data donation-based research. While this system could be developed and maintained by academics, it could also be mandated and funded through regulations or companies.

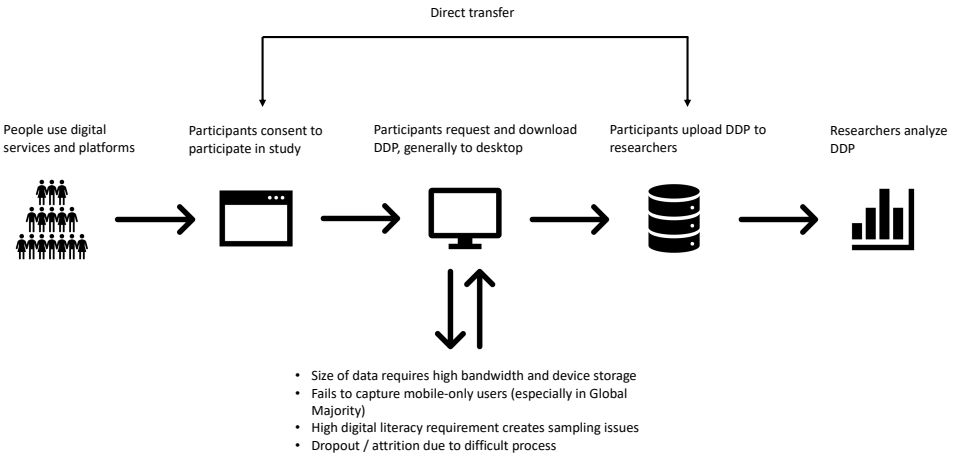


Figure 1: The challenges associated with requesting and downloading data, which impact the usefulness of data donations for research, could be ameliorated by direct data transfers.

Second, there should be investment in intermediary structures that could be effective bridges to reduce the burden of negotiating platform-to-researcher donation mechanisms. For example, this could take the shape of a research consortium that would set up mechanisms for transfers from major platforms, and researchers could interact with that consortium to support their particular projects. There are already models for this type of consortium approach for negotiating and provisioning data access between companies and researchers, such as the Social Media Archive at ICPSR and Social Science One. A similar model could be developed here; like the others, it would need platform buy-in.

Third, while some regulations have already come into effect (e.g., GDPR and the Digital Markets Act), others are still being considered. During the process of designing policy or regulation with data portability provisions, policymakers and regulators could think about how to design portability *for* research, such as by standardizing file formats and requiring clear documentation (Table 1).

Finally, in pushing for portability, regulators need to ensure that, in the pursuit of portability for competition, they do not inadvertently close the door on using portability for re-

Table 1: Aligning DDPs for Analysis

| | Current Challenges | Potential Solutions |
|---|---|---|
| Documentation | Documentation lacks clear explanations of variables | Mandate clear documentation of variables included in DDPs |
| Structure | Platforms do not provide DDPs structured for research | Require standardization of data structures, such as file and variables names |
| Machine readability | Platforms do not always provide files that are machine readable (e.g., HTML) | Ensure DDPs can be downloaded in machine readable formats |

search. A key mechanism for avoiding this unintended consequence is for policymakers to engage directly with researchers. There are successful models for regulator-academic communication and collaboration—such as the European Media Observatory working group, an independent intermediary body with experts across academia, industry, and civil society to support research on digital platforms—that could be adopted for this topic.[10]

[10]For more information, see `https://edmo.eu/about-us/edmoeu/`

# 5 References

## References


Abrajano, M., Garcia, M., Pope, A., Vidigal, R., Kamau, E., & Tucker, J., Joshua A. Nagler. (2022). Social media, information, and politics: Insights on latinos in the u.s.

Allen, J., Mobius, M., Rothschild, D. M., & Watts, D. J. (2021). Research note: Examining potential bias in large-scale censored data. *Harvard Kennedy School Misinformation Review*.

Anderson, M., & Jiang, J. (2023). *Teens, social media and technology 2023.* Pew Research Center: Internet, Science and Tech. https://www. pewresearch ....

Aslett, K., Sanderson, Z., Godel, W., Persily, N., Nagler, J., & Tucker, J. A. (2023). Online searches to evaluate misinformation can increase its perceived veracity. *Nature*, 1–9.

Ausloos, J., & Veale, M. (2020). Researching with data rights. *Amsterdam Law School Research Paper*(2020-30).

Auxier, B., & Anderson, M. (2021, April 7). Social media use in 2021. *Pew Research Center*. Retrieved from `https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/`

Boeschoten, L., Ausloos, J., Moeller, J., Araujo, T., & Oberski, D. L. (2020). Digital trace data collection through data donation. *arXiv preprint arXiv:2011.09851*.

Breuer, J., Bishop, L., & Kinder-Kurlanda, K. (2020). The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships. *New Media & Society*, *22*(11), 2058–2080.

Breuer, J., Kmetty, Z., Haim, M., & Stier, S. (2023). User-centric approaches for collecting facebook data in the 'post-api age': Experiences from two studies and recommendations for future research. *Information, Communication & Society*, *26*(14), 2649–2668.

Bruns, A. (2019). After the 'apicalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566.

Castro, D. (2021). *Improving consumer welfare with data portability* (Tech. Rep.). Information Technology and Innovation Foundation.

Commission, E. (2023). *Commission designates first very large online platforms and search engines under the digital services act.* Retrieved from `https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413`

Crutzen, R., Ygram Peters, G.-J., & Mondschein, C. (2019). Why and how we should care about the general data protection regulation. *Psychology & Health*, *34*(11), 1347–1357.

De Hert, P., Papakonstantinou, V., Malgieri, G., Beslay, L., & Sanchez, I. (2018). The right to data portability in the gdpr: Towards user-centric interoperability of digital services. *Computer law & security review*, *34*(2), 193–203.

de Vreese, C., & Tromble, R. (2023). The data abyss: How lack of data access leaves research and society in the dark. *Political Communication*, 1–5.

Fiesler, C., Beard, N., & Keegan, B. C. (2020). No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. In

*Proceedings of the international aaai conference on web and social media* (Vol. 14, pp. 187–196).

Fiesler, C., & Proferes, N. (2018). "participant" perceptions of twitter research ethics. *Social Media+ Society*, *4*(1), 2056305118763366.

Freelon, D. (2018). Computational research in the post-api age. *Political Communication*, *35*(4), 665–668.

Friedler, S., Singh, R., Blili-Hamelin, B., Metcalf, J., & Chen, B. J. (2023). Ai red-teaming is not a one-stop solution to ai harms. Retrieved from https://datasociety.net/wp-content/uploads/2023/10/Recommendations-for-Using-Red-Teaming-for-AI-Accountability-PolicyBrief.pdf

Gallagher, J. (2023). Reddit will begin charging for access to its api. *TechCrunch*. Retrieved from https://techcrunch.com/2023/04/18/reddit-will-begin-charging-for-access-to-its-api/

Gilbert, S., Vitak, J., & Shilton, K. (2021). Measuring americans' comfort with research uses of their social media data. *Social Media+ Society*, *7*(3), 20563051211033824.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on twitter during the 2016 us presidential election. *Science*, *363*(6425), 374–378.

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances*, *5*(1), eaau4586.

Gulati-Gilbert, S., & Seamans, R. (2023). Data portability and interoperability: A primer on two policy tools for regulation of digitized industries.

Haim, M., & Nienierza, A. (2019). Computational observation: Challenges and opportunities of automated observation within algorithmically curated media environments using a browser plug-in. *Computational Communication Research*, *1*(1), 79–102.

Halavais, A. (2019). Overcoming terms of service: a proposal for ethical distributed research. *Information, Communication & Society*, *22*(11), 1567–1581.

Hemphill, L., Schöpke-Gonzalez, A., & Panda, A. (2022). Comparative sensitivity of social media data and their acceptable use in research. *Scientific Data*, *9*(1), 643.

Hosseinmardi, H., Ghasemian, A., Clauset, A., Mobius, M., Rothschild, D. M., & Watts, D. J. (2021). Examining the consumption of radical content on youtube. *Proceedings of the National Academy of Sciences*, *118*(32), e2101967118.

Husovec, M. (2023). How to facilitate data access under the digital services act. *Available at SSRN 4452940*.

Keller, D. (2022). *Before the united states senate committee on the judiciary, subcommittee on subcommittee on privacy, technology and the law, hearing on platform transparency: Understanding the impact of social media.* Retrieved from https://www.judiciary.senate.gov/imo/media/doc/Keller%20Testimony1.pdf

Kharpal, N. (2023). Twitter announces new api with only free, basic and enterprise levels. *TechCrunch*. Retrieved from https://techcrunch.com/2023/03/29/twitter-announces-new-api-with-only-free-basic-and-enterprise-levels/

Kmetty, Z., & Németh, R. (2022). Which is your favorite music genre? a validity comparison of facebook data and survey data. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *154*(1), 82–104.

Krishnan, S. (2023, July 15). Opinion — threads, twitter, and the future of social media. *The New York Times*. Retrieved from https://www.nytimes.com/2023/07/15/opinion/social-media-threads-twitter-reddit.html

Krotov, V., Johnson, L., & Silva, L. (2020). Tutorial: Legality and ethics of web scraping.

Kupferschmidt, K. (2023). *Does social media polarize voters? unprecedented experiments on facebook users reveal surprises.*

Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., ... others (2020). Computational social science: Obstacles and opportunities. *Science*, *369*(6507), 1060–1062.

Lukito, J., Brown, M. A., Dahlke, R., Suk, J., Yang, Y., Zhang, Y., ... Soorholtz, K. (2023). *The state of digital media data research, 2023.* Retrieved from https://mddatacoop.org/files/2023/State%20of%20Digital%20Media%20Data%20Research%202023.pdf

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, *22*(2), 205-224. doi: 10.1177/1527476420982230

Mondschein, C. F., & Monda, C. (2019). The eu's general data protection regulation (gdpr) in a research context. *Fundamentals of clinical data science*, 55–71.

Moritz, M. (2016). Big data's 'streetlight effect': Where and how we look affects what we see. *The Conversation*, *17*.

Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., Reeves, B. B., & Robinson, T. N. (2023). Digital trace data collection for social media effects research: Apis, data donation, and (screen) tracking. *Communication Methods and Measures*, 1–18.

Ohme, J., Araujo, T., de Vreese, C. H., & Piotrowski, J. T. (2021). Mobile data donations: Assessing self-report accuracy and sample biases with the ios screen time function. *Mobile Media & Communication*, *9*(2), 293–313.

Ortiz-Ospina, E. (2019). The rise of social media. *Our World in Data*. (https://ourworldindata.org/rise-of-social-media)

Persily, N. (2021). A proposal for researcher access to platform data: The platform transparency and accountability act. *Journal of Online Trust and Safety*, *1*(1).

Persily, N., & Tucker, J. A. (2020). Conclusion: The challenges and opportunities for social media research. In N. Persily & J. A. Tucker (Eds.), *Social media and democracy* (p. 313–331). Cambridge University Press.

Pfiffner, N., & Friemel, T. N. (2023). Leveraging data donations for communication research: Exploring drivers behind the willingness to donate. *Communication Methods and Measures*, 1–23.

Prainsack, B. (2019). Data donation: How to resist the ileviathan. *The ethics of medical data donation*, 9–22.

Riley, C. (2023). *The future of generative ai is personal – and portable?* Retrieved from https://www.linkedin.com/pulse/future-generative-ai-personal-portable-chris-riley-rx4dc/?utm_source=rss&utm_campaign=articles_sitemaps&utm_medium=google_news

Robertson, R. E. (2022, Aug.). Uncommon yet consequential online harms. *Journal of Online Trust and Safety*, *1*(3). Retrieved from https://tsjournal.org/index.php/jots/article/view/87 doi: 10.54501/jots.v1i3.87

Robertson, R. E., Green, J., Ruck, D. J., Ognyanova, K., Wilson, C., & Lazer, D. (2023). Users choose to engage with more partisan news than they are exposed to on google search. *Nature*, 1–7.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063–1064.

Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.

Sanchez, G. R., & Bennett, C. (2022). Why spanish-language mis-and disinformation is a huge issue in 2022.

Sanderson, Z., & Tucker, J. A. (2023). *Beyond red teaming: Facilitating user-based data donation to study generative ai.* Tech Policy Press. Retrieved from https://www.techpolicy.press/beyond-red-teaming-facilitating-user-based-data-donation-to-study-generative-ai/

van Driel, I. I., Giachanou, A., Pouwels, J. L., Boeschoten, L., Beyens, I., & Valkenburg, P. M. (2022). Promises and pitfalls of social media data donations. *Communication Methods and Measures*, *16*(4), 266–282.

Wagner, M. W. (2023). Independence by permission. *Science*, *381*(6656), 388–391.

Walker, S., Mercea, D., & Bastos, M. (2019). *The disinformation landscape and the lockdown of social platforms* (Vol. 22) (No. 11). Taylor & Francis.

Welles, B. F. (2014). On minorities and outliers: The case for making big data small. *Big Data & Society*, *1*(1), 2053951714540613.

Yin, L. (2023). *Journalists should be looking for undocumented apis. here's how to start.* Retrieved from https://www.niemanlab.org/2023/03/journalists-should-be-looking-for-undocumented-apis-heres-how-to-start/

Zannettou, S., ElSherief, M., Belding, E., Nilizadeh, S., & Stringhini, G. (2020). Measuring and characterizing hate speech on news websites. In *Proceedings of the 12th acm conference on web science* (pp. 125–134).