DATA TRANSFER INITIATIVE

THE PRESENT AND FUTURE OF DATA PORTABILITY

A Collected Volume of Independent Scholarly Research, Edited by the Data Transfer Initiative

May 2024

CONTENTS

Introduction

The Data Transfer Initiative is pleased to present this compendium of original scholarship from independent academics around the world. This curated compilation includes thought-provoking research and analysis from a diverse group of esteemed scholars, each contributing valuable insights into pressing questions at the forefront of data transfers and their implications.

In order to thrive, the data portability ecosystem needs a lot more than just technologies—it needs thought leadership on portability in both policy and practice, including analyses of the laws and regulations that impact it. With this effort, we are proud to advance DTI's mission of "empowering people by building a vibrant ecosystem for simple and secure data transfers."

Data portability is a beacon of empowerment and innovation: it enhances user agency, fosters market vibrancy, and healthy competition among platforms and services. Across the globe, regulators increasingly recognize data portability as a pivotal tool in safeguarding consumer rights and fostering competition and innovation. However, implementation challenges and unresolved policy questions underscore the need for further research and dialogue to harness the benefits of data portability fully. This compendium delves into these complexities and addresses these open questions, offering insights and analysis into the nuances of data portability and its implications.

The articles that follow cover a broad range of topics—including regulatory analysis, privacy considerations, and the practical implications of data portability mandates. Among other issues, this compendium explores the European Union's Digital Markets Act and its implications for portability; the relationship between portability and cutting-edge technologies, including AI and the metaverse; and how data portability can support independent research.

A final note: this volume reflects DTI's commitment to facilitating informed discourse and collaborative efforts in navigating the evolving terrain of data transfer. The perspectives in these articles belong to the individual authors, and do not necessarily reflect either the views of their current or past employers or the positions of the Data Transfer Initiative. That said, we believe this effort exemplifies the collaborative and multi-faceted spirit that drives progress in this space. We hope that you enjoy the articles as much as we do.

Delara Derakhshani
Director of Policy, Data Transfer Initiative

DATA TRANSFER INITIATIVE

PORTABLE TRUST

Fostering both autonomy and privacy in data portability mechanisms

Cobun Zweifel-Keegan

Managing Director, International Association of Privacy Professionals (IAPP)

TABLE OF CONTENTS

I. Introduction

This policy brief provides an examination of the principle of data portability, a data protection right that has been increasingly recognized and codified in privacy laws worldwide. While not universally mandated, data portability is a critical tool for empowering individuals by giving them control over their personal data. However, achieving true portability is not without its challenges, as it can introduce significant privacy, security, and integrity risks. To mitigate these risks and foster trust in the process of data portability, the implementation of consistent operational, technical, and legal safeguards is essential. This requires prioritization of this data right within privacy programs—a portability by design approach—as well as collaborative engagement between platforms as opportunities for portability mature. This brief delves into each of these aspects, providing an overview of the complexities of achieving the promise of data portability from a privacy perspective, along with best practices to meet the challenge.

II. Data portability is a longstanding principle of data privacy.

The idea of data portability as a privacy right is often portrayed as a relatively recent innovation. This is mostly true. The European Union's General Data Protection Regulation (GDPR) was the first, in 2016, to add this right among the bundle of data protection rights data subjects enjoy across the EU.[1] But data portability has its roots in fundamental principles of data privacy going back to the first codes of practice.

More than fifty years ago, the Fair Information Practice Principles (FIPPs) were the first attempt to capture the principles and processes that should be supported when creating computerized systems storing the personal data of many individuals.[2] That is, they were the first attempt to promulgate best practices for the nascent field of data privacy, known in other jurisdictions as data protection. Over the years, many different versions of the FIPPs have been put forward, but even the first report by the U.S. Department of Health, Education, and Welfare included among its principles the idea that "there must be a way for an individual to find out what information about him is in a record and how it is used."[3] In other versions of the FIPPs, this principle was

---

[1] The GDPR reflected a similar right to data portability that was incorporated into France's Digital Republic Act of 2016. As early as 2011, the U.K. government pioneered its "midata" program to encourage the development of intra-company portability standards and processes. *See,* Kaori Ishii, Discussions on the Right to Data Portability from Legal Perspectives, IFIP ADVANCES IN INFO. & COMM'N TECH. VOL. 537 (2018), https://link.springer.com/chapter/10.1007/978-3-319-99605-9_26.

[2] For a brief overview of the history and importance of the FIPPs to the field of data privacy, see Cobun Zweifel-Keegan, *A view from DC: Celebrating privacy's 50th birthday*, IAPP: U.S. PRIVACY DIGEST, June 30, 2023, https://iapp.org/news/a/a-view-from-dc-celebrating-privacys-50th-birthday/.

[3] U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE, RECORDS, COMPUTERS, AND THE RIGHTS OF CITIZENS (1973), https://aspe.hhs.gov/reports/records-computers-rights-citizens.

broadened and referred to, generally, as "individual participation" or—perhaps more generously—"individual control."[4]

Control and participation privilege the idea of consent, when relevant, but also provide an umbrella under which the broad rights of access, correction, and redress are captured. All these separate rights are commonly reflected across privacy laws and codes—and have been continuously and with increasing sophistication for five decades.

To understand the unique elements of the right to data portability, one must first understand the much older right of access to personal data. The two rights bear much in common. (See *Figure 1*, below, for a comparison.) After all, they are both designed to empower individuals to exercise control over their personal data by receiving a copy of their data from an organization. Some privacy and data protection laws consider portability a special type of access, while others treat it separately as a standalone right.

Figure 1: ACCESS AND PORTABILITY AT A GLANCE

	Access	**Portability**
Definition	Allows individuals to obtain a copy of their personal data held by an organization.	Enables individuals to receive their data in a structured, machine-readable format and transfer it to another service provider.
Purpose	Helps individuals understand data processing, verify accuracy, and exercise other rights.	Facilitates switching services, promoting user autonomy and a competitive marketplace.
Scope	Pertains to the individual's own data within a specific organization.	May extend beyond individual access, allowing data to move directly between services on request.
Example	Requesting medical records from a hospital.	Transferring contact lists from one social media platform to another.

In addition to encouraging respect for consumers' requests to access personal data in codes like the FIPPs, the right of access is enshrined in the Charter of Fundamental Rights of the European Union, which lies at the foundation of EU law. Access and rectification are the only practical rights related to personal data explicitly listed in the Charter. Access is often described

[4] *See, e.g.*, U.S. DEPARTMENT OF HOMELAND SECURITY, THE FAIR INFORMATION PRACTICE PRINCIPLES (2015), https://www.dhs.gov/publication/privacy-policy-guidance-memorandum-2008-01-fair-information-practice-principles.

as the "gateway" to other privacy rights. Requesting access to one's personal data enables an individual to better understand the extent to which an organization is processing information about them, which can enable further requests, such as correcting an incorrect record or deleting it entirely.

Portability goes one step farther, enabling the individual to exercise their autonomy over their own personal data, including in ways that do not serve the interests of the platform with which they are interacting. It is not by accident that portability requirements include language about the usability of data. When fully achieved, data portability empowers individuals to make use of their own data without regard to the whims of platforms. It thus could be considered the culmination of rights related to the autonomy of the data subject.

III. Data portability is widely—but not universally—required under privacy law.

It is important to stress the fact that data portability is an independent data protection right, separate and apart from its operation as a pro-competitive regulatory measure. In fact, in its guidance on the subject, the European Data Protection Board (EDPB) takes pains to highlight this fact: "Whilst the right to personal data portability may also enhance competition between services (by facilitating service switching), the GDPR is regulating personal data and not competition. In particular, article 20 does not limit portable data to those which are necessary or useful for switching services."[5]

Though data portability is brought up most frequently in the context of social media, communications, or personal tracking data, the right is not explicitly limited to any personal data types.[6] However, the right to data portability is not universally applicable under data protection or consumer data privacy laws.

Artful legal drafting limits the obligation to respect data portability requests to those situations where it is already "technically feasible" to provide a structured, commonly used, machine readable format. This is the case in most U.S. state privacy laws, which recognize the right alongside the simple right of access.[7] (See *Figure 2*, below, for example language from

---

[5] Article 29 Data Protection Working Party, Guidelines on the right to data portability at 4, 13 Dec. 2016, https://ec.europa.eu/newsroom/article29/items/611233.

[6] Sasha Hondagneu-Messner, *Data Portability: A Guide and a Roadmap*, 47 RUTGERS COMPUTER & TECH. L.J. 240, 249 (2021).

[7] For an up-to-date listing of U.S. state privacy laws tracking their inclusion of portability requirements, see International Association of Privacy Professionals (IAPP), U.S. State Privacy Legislation Tracker, https://iapp.org/resources/article/us-state-privacy-legislation-tracker/. For an in-depth comparison of the portability requirements across U.S. states (and beyond), see Delara Derakhshani, *Global developments in data portability law*, Data Transfer Initiative, Oct. 25, 2023, https://dtinit.org/blog/2023/10/24/global-developments.

California and Colorado.) Although the GDPR's recitals includes lofty language that companies "should be encouraged to develop interoperable formats that enable data portability," it also limits portability to what is technically feasible, but only for the component of the regulation that requires direct transfer of data between controllers.[8] The technically feasible exception does not apply to individuals' direct download requests under GDPR.

Figure 2: Comparing portability rights across a selection of privacy laws

	Legal Text	**Primary Guidance**
EU General Data Protection Regulation	**Article 20:** ***Right to data portability*** 1. The data subject shall have the right to receive the personal data concerning him or her, which he or she has provided to a controller, in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller without hindrance from the controller to which the personal data have been provided, where: a. the processing is based on consent pursuant to point (a) of Article 6(1) or point (a) of Article 9(2) or on a contract pursuant to point (b) of Article 6(1); and b. the processing is carried out by automated means. 2. In exercising his or her right to data portability pursuant to paragraph 1, the data subject shall have the right to have the personal data transmitted directly from one controller to another, where technically feasible. 3. The exercise of the right referred to in paragraph 1 of this Article shall be without prejudice to Article 17. That right shall not apply to processing necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller. 4. The right referred to in paragraph 1 shall not adversely affect the rights and freedoms of others.	**Recital 68:** To further strengthen the control over his or her own data, where the processing of personal data is carried out by automated means, the data subject should also be allowed to receive personal data concerning him or her which he or she has provided to a controller in a structured, commonly used, machine-readable and interoperable format, and to transmit it to another controller. Data controllers should be encouraged to develop interoperable formats that enable data portability. That right should apply where the data subject provided the personal data on the basis of his or her consent or the processing is necessary for the performance of a contract. It should not apply where processing is based on a legal ground other than consent or contract.… The data subject's right to transmit or receive personal data concerning him or her should not create an obligation for the controllers to adopt or maintain processing systems which are technically compatible. Where, in a certain set of personal data, more than one data subject is concerned, the right to receive the personal data should be without prejudice to the rights and freedoms of other data subjects in accordance with this Regulation.… Where technically feasible, the data subject should have the right to have the personal data transmitted directly from one controller to another.
California Consumer Privacy Act, as amended by California Privacy Rights Act	**Calif. Civil Code Sec. 1798.130** (a) … a business shall, in a form that is reasonably accessible to consumers: … (3)(B) For purposes of [the business's obligation to disclose information about a consumer in response to a verifiable consumer request under the Right to Access] (iii) Provide the specific pieces of personal information obtained from the consumer in a format that is easily	**CPPA Regulations Section 7024** (g) If a business maintains a password-protected account with the consumer, it may comply with a request to know by using a secure self-service portal for consumers to access, view, and receive a portable copy of their personal information if the portal fully discloses the personal information that the

[8] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1, at Recital 68.

	understandable to the average consumer, and to the extent technically feasible, in a structured, commonly used, machine-readable format, which also may be transmitted to another entity at the consumer's request without hindrance. "Specific pieces of information" do not include data generated to help ensure security and integrity or as prescribed by regulation. Personal information is not considered to have been disclosed by a business when a consumer instructs a business to transfer the consumer's personal information from one business to another in the context of switching services.	consumer is entitled to under the CCPA and these regulations, uses reasonable data security controls, and complies with the verification requirements set forth in Article 5.
Colorado Privacy Act	**C.R.S. § 6-1-1306** (1)(e) *Right to data portability*. When exercising the right to access personal data pursuant to subsection (1)(b) of this section, a consumer has the right to obtain the personal data in a portable and, to the extent technically feasible, readily usable format that allows the consumer to transmit the data to another entity without hindrance. A consumer may exercise this right no more than two times per calendar year. Nothing in this subsection (1)(e) requires a controller to provide the data to the consumer in a manner that would disclose the controller's trade secrets.	**CPA Rule 4.07** A. To comply with a data portability request, a Controller must transfer to a Consumer the Personal Data it has collected and maintains about the Consumer through a secure method in a commonly used electronic format that, to the extent technically feasible, is readily usable and allows the Consumer to transmit the Personal Data to another entity without hindrance. B. Pursuant to C.R.S. § 6-1-1306(1)(e), a Controller is not required to provide Personal Data to a Consumer in a manner that would disclose the Controller's trade secrets. When complying with a request to access Personal Data in a portable format, Controllers must provide as much data as possible in a portable format without disclosing the trade secret.

The GDPR further limits the applicability of the right to data portability in three ways. First, the right is limited to those situations where data is processed subject to the legal bases of consent or contract. This leaves four other legal bases, including the widely used "legitimate interests" basis, under which an organization may legally process personal data without any obligation to respect individual requests for a portable copy of personal data. In contrast, U.S. state laws do not reflect this same limitation.

In addition, the GDPR generally limits the operation of data protection principles if they conflict with other fundamental rights or interests. Risks to the privacy or other rights of the requesting individual or others could override the right to data portability in certain circumstances.[9]

Finally, the GDPR and some—but not all—of the laws it inspired limit the scope of the right to data portability to data collected from the individual. That is, personal data about an individual that is not directly collected from them may not be required to be included in a portable form,

[9] *See, id.,* Section III.

whether collected from another source, created through the operation of the service, or inferred from other data. Such data may be worth considering as includable in portability requests, however, if the individual is likely to expect its inclusion. Furthermore, some U.S. state privacy laws do not include limiting language about the source of the data (see, e.g., Colorado above).[10]

IV. Portability supports individual empowerment.

European jurisprudence considers the framing of "informational self-determination" as core to data protection rights. In 1983, the German Constitutional Court ruled that whoever "cannot survey with sufficient assurance the information concerning himself known in certain areas of his social surroundings, and whoever is not in a position to assess more or less the knowledge of possible partners in communication, can be essentially obstructed in his freedom to make plans or decisions on the basis of his own self-determination."[11]

Whether framed in language about participation, agency, control, or autonomy, data rights like the right of access help to empower individuals to gain knowledge about the spread of their personal data and power over how it is collected, used, and shared. The right to data portability relies on this same philosophical underpinning.

As the EDPB explains in its portability guidance, "This right… supports user choice, user control and user empowerment…. By affirming individuals' personal rights and control over the personal data concerning them, data portability also represents an opportunity to 're-balance' the relationship between data subjects and data controllers. The primary aim of data portability is enhancing individual's control over their personal data and making sure they play an active role in the data ecosystem."[12]

Nevertheless, portability has to date been the least exercised and developed right under the GDPR.[13] This is evidenced by the lack of notable developments regarding the right to data portability, such as supervisory enforcement or case law. Most jurisdictions reported no significant developments, and data portability rarely seems to be used by data subjects or debated before a court. Research also shows confusion and a lack of regularity in responses to

[10] For other states, see also IAPP, *supra* note 8.

[11] For an explanation of the importance of the *Karlsruhe* case and the reasoning behind data portability in the European context generally, *see* Gabriela Zanfir-Fortuna, *The right to data portability in the context of the EU data protection reform*, INT'L DATA PRIVACY L. (May 11, 2012) at 149, https://ssrn.com/abstract=2215684.

[12] Article 29 Data Protection Working Party, *supra* note 6 at 3-4.

[13] Jurre Reus & Nicole Bilderbeek, Data portability in the EU: An obscure data subject right, IAPP: PRIVACY PERSPECTIVES, Mar. 25, 2022, https://iapp.org/news/a/data-portability-in-the-eu-an-obscure-data-subject-right/.

portability requests.[14] This contrasts with other rights, such as the right to access, which data subjects have frequently relied on, resulting in a broad catalog of jurisprudence. As for case law in the EU about the right to data portability, there have been very few, if any, cases.

V. Porting data has inherent privacy, security, and integrity risks.

The philosophy of individual empowerment undergirds the right to data portability whether it is exercised by an individual requesting direct access to machine-readable data (a "direct-download scenario"), or via a controller-to-controller transfer request. However, these two distinct methods of exercising portability may both present privacy and security risks to individuals, organizations, and third parties (see *Figure 3*, below). In fact, some scholars have critiqued the right to data portability as inherently not worth the risks and drawbacks.[15]

When an individual exercises their right to direct access to machine-readable data, they may encounter security and privacy risks. Securely transferring large volumes of data can be a complex task, and any breach during this process could expose sensitive information. Individually processing or accessing the requested data may require users to download software, a further security threat. Ongoing risks of breach from improper storage or re-upload to unverified destinations make the direct download scenario riskier for individuals.

Conversely, when data portability is exercised via a transfer request, where data is transferred directly from one controller to another, a different set of risks emerges. The primary risks for controllers sharing data include the failure to inform individuals about how their data will be processed, collecting personal data for one purpose and subsequently sharing or using it for another incompatible purpose without the data subject's consent, and the inability within receiving platforms to maintain the integrity and security of the data. Moreover, cross-border data transfers can introduce complexities.

Only by considering these concerns and implementing appropriate safeguards throughout the data lifecycle can portability mature as a practice and earn the trust of consumers. Trust in the process of exercising portability requires efforts to build trusted privacy practices within individual companies, but efforts must not stop there. Multi-party efforts must also be made, within domains of specific data uses (e.g., fitness trackers, social graphs), to build trust in the

---

[14] *See, e.g.*, Janis Wong and Tristan Henderson, *The right to data portability in practice: exploring the implications of the technologically neutral GDPR*, INT'L DATA PRIVACY L. (July 6, 2019) at 173, https://academic.oup.com/idpl/article-abstract/9/3/173/5529345.

[15] Peter Swire & Yianni Lagos, *Why the Right to Data Portability Likely Reduces Consumer Welfare: Antitrust and Privacy Critique*, 72 MD L. REV. 335 (2013), https://fpf.org/wp-content/uploads/2013/07/Swire-Lagos_Why-the-Right-to-Data-Portability-Likely-Reduces-Consumer-Welfare1.pdf (exploring a variety of critiques of the idea data portability before passage of the GDPR, including the lack of focus on market power and inherent privacy and security risks from operationalizing the right).

general process of portability. Truly portable trust, as this could be called, is an ideal that has yet to be realized in most domains.

Figure 3: Risk considerations for data portability

	Risk to Organization	Risk to Individual	Risk to Others
Verification	Misidentification of data subject can result in exposure.	Verification often requires sharing of personal information, plus exposure risk.	
Scope	Over- or under-inclusion of personal data.	Mismatch between expectation and reality for extracted data. Usability concerns may vary.	Other individuals may have personal data included.
Transfer	Security risks of extracting and transferring a large file. Compliance risks for cross-border transfer.	Higher risk of exposure during transfer, whether through direct download or B2B.	Receiving platform may take control of data at a high-risk time, integrating datasets can cause corruption.
Storage	Duplicated dataset is outside of scope of control for privacy and security safeguards.	Ill-equipped to store data in a secure and privacy-preserving way.	Receiving platform may have different privacy / security posture with incompatible fields.
Re-use	Purpose limitations and sharing provisions of privacy policy may not be met for ported data.	When porting to new platform, may not be aware of different privacy considerations.	Receiving platform may not be aware of limitations or irregularities in dataset, can fail to maintain integrity.

VI. Portable trust and privacy require consistent safeguards.

Though portability has not been top-of-mind for regulators since its introduction as a privacy right, the tide is already shifting as data protection standards continue to mature—and other regulatory frameworks draw attention to data portability.[16] As with any privacy practice, organizations that under-invest in portability processes now may find themselves paying higher costs to adjust systems later.

Unlike the closely linked idea of interoperability, portability can be initiated on a unilateral basis.[17] But unilateral mechanisms, such as portals and direct downloads, bring with them the

[16] *See*, Chris Riley & Delara Derakhshani, *Future Horizons for Data Portability Research*, TECH POL. PRESS, Sept. 28, 2023, https://www.techpolicy.press/future-horizons-for-data-portability-research/.

[17] Sukhi Gulati-Gilbert and Robert Seamans, Data portability and interoperability: A primer on two policy tools for regulation of digitized industries, Brookings, May 9, 2023, https://www.brookings.edu/articles/data-portability-and-interoperability-a-primer-on-two-policy-tools-for-regulation-of-digitized-industries-2/.

heightened risks to the security and privacy of the data explored above. On the other hand, bilateral and multilateral efforts to build uniform standards and technical systems for translating datasets between platforms can be expensive without eliminating risks.

Nevertheless, a principles approach to data privacy should encourage organizations to consider portability measures, especially for data types for which user autonomy and empowerment are most likely to be reflected. For example, today's computer users have come to expect the ability to maintain control over their communications, their social graph, the content they produce, and longitudinal insights about themselves driven by sensors such as fitness monitors. When physical analogues exist over which consumers are familiar taking an ownership interest, their privacy expectations around the portability of their data away from platform control are likely to be correspondingly high.

Organizations are well advised to consider both compliance and consumer trust goals in developing robust portability mechanisms. Achieving such measures first requires internal investment, even before multilateral challenges are addressed. Thus, privacy programs should implement operational, technical, and legal safeguards that consider portability throughout the data lifecycle. The costs of re-architecting systems to allow for portability can be much higher than designing them with portability in mind from the beginning. A much-cited example is the $3 billion price tag that U.S. telephone carriers spent in re-architecting systems to allow for phone number portability between operators.[18]

For those systems that users are likely to view through a lens of their own autonomy as stewards of their data, and those systems that users invest significant time or energy in curating, organizations should consider portability as early as possible in the design and engineering process. This "portability by design" approach should embrace efforts across operational, legal, and technical controls.

Portability by design involves architecting systems from the outset to support data portability, thereby embedding this right into the very fabric of the system's design and operation. Doing so ensures that data portability is not an afterthought but a fundamental aspect of the system, thereby reducing potential risks and enhancing the security and privacy of data subjects. This proactive approach can help mitigate potential vulnerabilities, enhance data integrity, and foster greater trust among data subjects.

Operational and legal safeguards are the first line of defense. These are the written policies, procedures, and practices that organizations put in place to ensure secure and efficient data portability. For instance, organizations need to establish clear protocols for recognizing and processing data portability requests, and for determining the scope of the data that will be

[18] Joshua Gans, Stephen King, and Graeme Woodbridge, *Numbers to the people: regulation, ownership and local number portability*, 13 INFO. ECON. POLICY 167 (2001).

subject to each type of portability request.[19] User education is another vital operational concern, especially when providing users with an opportunity to download large quantities of raw data. Warnings about the security and privacy risks should be coupled with information about properly vetting third-party platforms and securely storing data.

Fully implementing operational controls also requires technical expertise. Technical safeguards include those mechanisms that enable trusted verification of data requestors, secure transmission of personal data, and encrypted file types to facilitate secure storage. In its portability guidance, the EDPB provides an overview of a lengthy but non-exhaustive list of possible technical mechanisms to consider in facilitating portability, including "secured messaging, an SFTP server, a secured WebAPI or WebPortal" in addition to the possibility of facilitating data subjects in their use of a "data store, personal information management system or other kinds of trusted third-parties, to hold and store the personal data and grant permission to data controllers to access and process the personal data as required."[20]

The last factor to consider when embracing data portability—but far from the least important—is participation in multilateral mechanisms to support safe and trustable transfers of data in ways that reduce friction, increase usability, and mitigate user-driven risks. Like other systems that benefit from, but do not require, multi-party collaboration, portability mechanisms can be more trusted and long-lasting through intervention by trusted third-party actors.

Multilateral mechanisms can take a variety of forms with various levels of formality. Associations or other independent intermediaries can encourage or even directly shape the continued investment in portability resources and interoperable systems. Governmental and non-governmental actors can craft standards and protocols for nascent technical systems to move beyond proprietary, siloed mechanisms. Independent bodies can also serve as outside verifiers of portability, through the creation of recognized trust marks or certifications that would verify compatibility with best practices.[21] At the far end of formalized mechanisms, multilateral governance structures can facilitate ongoing interoperable frameworks for portability, which can have knock-on effects for driving value in the marketplace.[22]

---

[19] See the guidance from the U.K. data protection authority for a detailed description of some of these measures. U.K. Information Commissioner's Office, *Right to data portability*, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/individual-rights/individual-rights/right-to-data-portability/.

[20] Article 29 Data Protection Working Party, *supra* note 6 at 16.

[21] For a review of the factors that contribute to robust independent accountability mechanisms, see BBB National Programs, filed comment in response to NTIA request for comments on artificial intelligence system accountability measures and policies, FR Doc # 2023-07776, June 12, 2023, https://www.regulations.gov/comment/NTIA-2023-0005-1158.

[22] For a discussion of multilateral governance structures, *see* Sukhi Gulati-Gilbert and Robert Seamans, *supra* note 18.

VII. Conclusion

Much work is still needed to achieve the goals of autonomy, consumer empowerment, and self-determination that lie beneath the privacy interests in data portability.

The successful implementation of data portability hinges on a concerted effort from companies to invest resources in internal portability initiatives. This includes the development of robust systems and processes that facilitate secure and efficient data transfer consistent with a holistic privacy program. However, internal efforts alone are not sufficient. Companies must also actively engage in multilateral or multi-stakeholder mechanisms that foster collaboration, standardization, and mutual understanding among different actors in the data ecosystem. Furthermore, companies that support the goals of portability should support the creation of new mechanisms that address emerging challenges and opportunities—or incorporate new technical modalities—while fostering trust in the broader portability landscape.

A multifaceted approach is crucial for overcoming the complexities of data portability and for realizing its full potential in empowering individuals and fostering a competitive user-centric marketplace.

DATA TRANSFER INITIATIVE

DATA PORTABILITY IRL:

A stakeholder assessment of data portability methods

Angela Woodall
Research Fellow, CELSA Sorbonne University

TABLE OF CONTENTS

This paper offers a real-world analysis of data portability, examining its use in practice and the power dynamics associated with stakeholders involved in the data portability ecosystem. Its goal is to help illustrate some of the challenges and opportunities that arise at the intersection of tools and incentives within portability.

I. Introduction

Access to data is at the center of two competing interests. On one hand, data has become crucial for citizens and the information sources they depend on, as well as businesses, governments, and researchers.[23] On the other hand, with new incentives and the expansion of AI into sectors across industries and territorial boundaries, keeping data flowing is also an imperative for competition and innovation in the digital economy. Policymakers are faced with balancing multiple priorities and multiple stakeholders.[24]

Principally a technical capacity, data portability is now being used as a policy tool to promote the production of data and enhance competition, while at the same time reducing the barriers for individuals who want to make decisions about their data.[25] These objectives are complex, both in terms of regulatory questions and technical requirements. One issue to resolve is how to balance the enhanced access to data being sought by lawmakers with privacy, data protection, and security.

I look at the way that policymakers are pursuing the complexity of these objectives. One tactic is to implement policy through technical design, an approach referred to as "by-design and default." This is a framework that requires industry to integrate data protection principles into the design and development of systems for processing personal data through technical design, business strategies, and organizational practices. The rationale is an implicit recognition of the potential of information systems architecture to shape human conduct as, or more, effectively than through the imposition of legislation or contract.[26] Thus a considerable burden rests on choices made by regulators and the facilitators of data portability because decisions will

[23] OECD, "Enhancing Access to and Sharing of Data," 2019.

[24] Marie-Agnes Jouanjean et al., "Issues around Data Governance in the Digital Transformation of Agriculture: The Farmers' Perspective" (Paris: OECD, October 23, 2020).

[25] Helena Ursic, "Unfolding the New-Born Right to Data Portability: Four Gateways to Data Subject Control," *SCRIPTed: A Journal of Law, Technology & Society* 15, no. 1 (August 1, 2018): 42–69.

[26] Lee A. Bygrave, "Data Protection by Design and by Default: Deciphering the EU's Legislative Requirements," *Oslo Law Review* 4, no. 2 (2017).

impact how users are able to exercise their rights.[27] With this in mind, I have drawn attention to the limits of technical solutions to policy problems.[28] What works as a matter of policy may not always be easy to implement technically, and vice-versa.[29]

To date, implementation of data portability overall has produced uneven results.[30] Success will require more details about stakeholders and their needs on one hand, and, on the other, the methods that are available to them, which is the focus of this study.

The first section summarizes data portability frameworks in the European Union. I largely focus on E.U. data regulations based on a rights-centric regulatory model coupled with the pursuit of a competitive model to counterbalance the market concentration in countries like the United States and China. These countries have distinctive models and priorities of their own, but the assessment is relevant to both as well as initiatives elsewhere.[31] Section one is followed by a

---

[27] Users refer to individuals who produce data by using online systems – like social media. They are called data subjects in the European Union. I have adopted the term user and user rights, as well as individuals, in much of this study.

[28] Joel Reidenberg, "Lex Informatica: The Formulation of Information Policy Rules through Technology," *Tex. L. Rev.* 76 (January 1, 1997): 553.

[29] "Data Portability Initiatives in The European Union, California, and India," in *Data To Go: An FTC Workshop on Data Portability* (Online: Federal Trade Commission, 2020), 31–81, https://www.ftc.gov/news-events/events/2020/09/data-go-ftc-workshop-data-portability.

[30] Daniel Gill and Jakob Metzger, "Data Access through Data Portability – Economic and Legal Analysis of the Applicability of Art. 20 GDPR to the Data Access Problem in the Ecosystem of Connected Cars," SSRN Scholarly Paper (Rochester, NY, May 5, 2022), https://papers.ssrn.com/abstract=4107677; Gabriel Nicholas, "Taking It With You: Platform Barriers to Entry and the Limits of Data Portability," *Michigan Technology Law Review* 27, no. 2 (April 1, 2021): 263–98, https://doi.org/10.36645/mtlr.27.2.taking; Sarah Turner et al., "The Exercisability of the Right to Data Portability in the Emerging Internet of Things (IoT) Environment," *New Media & Society* 23, no. 10 (October 1, 2021): 2861–81, https://doi.org/10.1177/1461444820934033; Gabriel Nicholas and Michael Weinberg, "Data Portability and Platform Competition: Is User Data Exported From Facebook Actually Useful to Competitors?," 2019; Lachlan Urquhart, Neelima Sailaja, and Derek McAuley, "Realising the Right to Data Portability for the Domestic Internet of Things," *Personal and Ubiquitous Computing* 22, no. 2 (April 1, 2018): 317–32, https://doi.org/10.1007/s00779-017-1069-2; Janis Wong and Tristan Henderson, "How Portable Is Portable? Exercising the GDPR's Right to Data Portability," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, UbiComp '18 (New York, NY, USA: Association for Computing Machinery, 2018), 911–20, https://doi.org/10.1145/3267305.3274152.

[31] By-design and default are an extension of the earliest data protection and privacy initiatives grouped under Fair Information Practices. They developed from exchanges between U.S. and European data protection and privacy lawmaking. See for background: Abraham L. Newman, *Protectors of Privacy: Regulating Personal Data in the Global Economy*, (Ithaca: Cornell University Press, 2008); René Mahieu, "The Right of Access to Personal Data: A Genealogy," *Technology and Regulation* (August 20, 2021): 62–75.

summary of results from an assessment of three models for personal data portability, including advantages and disadvantages I found during the assessment, and, finally, some recommendations.[32]

II. Data Portability

The term portability refers to the ability to request, receive, and transfer data between different applications. Individuals can switch between providers or keep data in two places at once (respectively, "switching" and "multi-homing").[33] Data portability as such is intended to serve dual purposes. One, from the realm of individual privacy rights: you can control your data by extracting it from a platform and managing it directly.[34] Or, from the realm of market competition, you can offer it to a different service provider.[35] Data portability is expected to enhance competition against large "gatekeeper" data controllers, making them easier to regulate and offering opportunities for innovation and growth within domestic markets that would otherwise be channeled elsewhere.

The E.U. General Data Protection Regulation (GDPR) enshrined portability as a right. The E.U. Digital Markets Act (DMA), as well as proposals elsewhere inspired by the E.U. example, include regulations for portability that are intended to drive markets based on innovation, technology developments, and data. An important addition in the DMA is the requirement for gatekeepers targeted by the regulation to provide tools that facilitate data portability not only for individual end users but also business users, in real time and continuously.

The DMA is one in a package of initiatives that make up the European strategy for data. The proposals aim to facilitate the use and sharing of personal data between more public and private parties, to support the use of specific technologies such as AI, and to regulate online platforms and gatekeepers. Processing of personal data already is or will be a core activity of the entities, business models, and the technologies regulated by the proposals. The combined effect of the adoption and implementation of the proposals will therefore significantly impact the protection of the fundamental rights to privacy and to the protection of personal data.[36]

---

[32] I have provided references to scholarship for readers who seek more background about each section.

[33] "The Digital Markets Act: Ensuring Fair and Open Digital Markets".

[34] Chris Riley, "Unpacking Interoperability in Competition," *Journal of Cyber Policy* 5, no. 1 (January 2, 2020): 94–106.

[35] Unlike the right of access which allows individuals to check what organizations know about them, and thus provide a little transparency and trust, the right to portability introduces monetary value into the relationship between the data controller and the person who generated the original data in the first place. Also, portability concerns a subset of the personal data held by the data controller rather than the entirety.

[36] European Data Protection Board, "Statement on the Digital Services Package and Data Strategy," November 18, 2021.

The increasing importance of data, and the expectations for portability to deliver social and economic benefits from data, lead to a number of considerations. How should the goals of the portability and protection measures be realized? And how have stakeholders been able to make use of portability rights with the means that are made available to them?

III. Stakeholders

I consider the main stakeholders to be individual data producers. Secondary stakeholders are intermediaries: the entities wanting more data from the platforms that control it. Wanting more data are competitors expected to gain most from data portability and a number of third parties including researchers, journalists, watchdogs, and activists. Government can be considered to be an intermediary as well. Completing the triad are controllers, so called because they direct access to data generated by individuals. Each stakeholder group will have a set of expectations and concerns according to political, financial, regulatory, and social demands.

Controllers now play a role in the delivery of essential services. So, in addition to social interactions, the data they control includes details about education, health, medicine and other categories. They have an interest in maintaining command over data for competitive purposes but also because they are responsible for data security. They also have an incentive to favor regulatory frameworks that are consistent with their interests and operations.

Policymakers don't want to erode privacy and force individuals to simply trade away their right to control their data.[37] But they also want a competitive market and they want portability to be effective in enabling small firms and new services to compete. These stakeholders have an interest in the right market conditions as well as structured data in usable formats with minimal ownership complexities.[38]

Researchers, journalists, government agencies and other civil society groups rely on personal data to study social phenomena, including the activities of the controllers and the people who engage with them.

What most individuals expect from data portability is not entirely clear. Their interests are considered (by other stakeholders) to be based on a combination of their role as consumer and as citizens with stakes in data protection according to intellectual, social, and political goals. As producers, they may expect "all" of "their" data to be portable but these concepts of "all" and

[37] Maciej Krzysztof Zuziak et al., "Data Collaboratives with the Use of Decentralised Learning," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23 (New York, NY, USA: Association for Computing Machinery, 2023), 615–25.

[38] OECD, "Data Portability, Interoperability and Digital Platform Competition," Competition Committee Discussion Paper, 2021.

"their" lend themselves to inherent ambiguity. For example, while I may be able to read content contributed by others, such as comments on a social media post of mine, that access does not translate to a portability right.[39] Inside and outside of the realm of social data, individual stakeholders have joined data collectives to combine their scale of control and their collective influence.[40] The collectives are also forced to navigate between opportunity, risk, and user protection.

IV. Data Protection by Design and Default

Reconciling opportunity, risk, and user protection makes data portability a complex goal to implement. Proposals for fulfilling data-related policy goals vary, but policymakers have leaned on the approach of data protection by design and default, where data controllers integrate relevant protection of personal data into the architecture and design of their devices. Individuals then manage their data with whatever information systems are produced.[41] "By-design" provides the specifications by which regulatory objectives should be translated into tools for users to control their data.[42] "Default" is often meant to signal data minimization, which means that a data controller should limit the collection of personal information to what is directly relevant and necessary to accomplish a specified purpose.[43] The means should be automated to the extent possible (a company cannot be made responsible for integrating requirements for which no technical solution has yet been developed).[44] Automated, built-in portability compliance could, for example, be a portal that allows individuals to automatically download and port their data with default settings that are easy to change. Some of tech's most dominant operators implemented such options shortly before the access and portability provisions in the GDPR took effect. The following section outlines three of them.[45]

---

[39] Mariavittoria Catanzariti and Deirdre Curtin, "Beyond Originator Control of Personal Data in EU Interoperable Information Systems: Towards Data Originalism," in *Data at the Boundaries of European Law* (Oxford University Press, 2023).

[40] Paul De Hert et al., "The Right to Data Portability in the GDPR: Towards User-Centric Interoperability of Digital Services," *Computer Law & Security Review* 34, no. 2 (April 2018): 193–203.

[41] Ann Cavoukian, "Privacy by Design: The 7 Foundational Principles Implementation and Mapping of Fair Information Practices" (Information and Privacy Commissioner of Ontario, May 2010); "CCPA vs GDPR," Cookiebot, November 30, 2020.

[42] This capacity is referred to as information self-determination. See: Serge Gutwirth et al., *Reinventing Data Protection?*, Softcover reprint of hardcover 1st ed. 2009 édition (Springer, 2010).

[43] I have shortened the term to by-design.

[44] Mireille Hildebrandt and Laura Tielemans, "Data Protection by Design and Technology Neutral Law," *Computer Law & Security Review* 29, no. 5 (October 1, 2013): 509–21.

[45] Another tactic is the successful provisioning of data by automated self-service tools by industry (e.g. IoT, banking, utilities), legislation (e.g. DMA), or use case (e.g. collectives), see: Marlene Barth, "A Case Study on Data Portability," *Datenschutz und Datensicherheit - DuD* 45, no. 3 (March 1, 2021): 190–97; Paul De Hert

V. Portability in Practice

The first portability interfaces were designed to allow individuals to download their data to a local device. While the mechanics can vary, most present a simple web interface layered over internal APIs, and include rate limits and other functions operating behind the scenes.

Newer interfaces can be characterized as service-to-service portability portals, which Google, Facebook, Twitter, and Microsoft instituted in 2018 as part of a consortium of technology companies called the Data Transfer Project.[46] Apple joined the project soon after.[47] In keeping with the then-newly established GDPR, members published principles for the initiative including guaranteeing privacy and security, reciprocal portability between importers and exporters, and fostering trust among users that their data will be protected – a prerequisite for widespread adoption of data portability tools.[48] Direct transfer tools are available today within, for example, Google Takeout,[49] Facebook's Transfer your Information,[50] and Apple's Data and Privacy Page.[51]

APIs are the third category. I look at APIs as a distinct category because they play a critical role in data portability, interoperability, and more generally in by-design frameworks by providing well-defined entry points coupled with authentication protocols.

VI. Methodology

I first tried to move my data from Twitter to Mastodon and to Facebook, and vice-versa, as an opportunity to consider expectations that portability could be used to switch data providers and avoid lock-in. After Twitter was sold by its founders in October 2022, two main options, Threads and Mastodon, emerged as alternatives.[52]

---

et al., "The Right to Data Portability in the GDPR: Towards User-Centric Interoperability of Digital Services," *Computer Law & Security Review* 34, no. 2 (April 2018): 193–203.

[46] Wikipedia, "Data Transfer Project," https://en.wikipedia.org/wiki/Data_Transfer_Project.

[47] https://www.theverge.com/2019/7/30/20746868/apple-data-transfer-project-google-microsoft-twitter

[48] Craig Shank, "Microsoft, Facebook, Google and Twitter Introduce the Data Transfer Project: An Open Source Initiative for Consumer Data Portability," EU Policy Blog, July 20, 2018.

[49] https://takeout.google.com/takeout/transfer/custom/photos

[50] http://facebook.com/tyi

[51] https://privacy.apple.com/

[52] Mastodon is an independent social networking text-based service with microblogging features similar to Twitter. Threads is an app with Twitter-like features owned by Facebook and listed under Instagram Inc. on the company website.

I then assessed the portals developed for portability between Google and Facebook, described previously.[53] I considered the overall advantages and disadvantages of each in line with four main criteria:

- Data portability is/is not available
- Method for requests
    a. Dedicated portal or other method
    b. Automated process
- Method for transfers
    a. Available procedures, tools, and techniques
    b. Real-time and continuous transfer
- Extent and relevance of the data (scope)
    a. Missing data
    b. Organization of data (and explanations)
    c. Data format

[53] To understand the design of the Twitter download portal, I relied on a combination of reverse engineering and a methodological approach called a "walkthrough" as a device for systematizing my assessment. See Nicholas Diakopoulos, "Algorithmic Accountability: On the Investigation of Black Boxes" (New York: Tow Center for Digital Journalism, Columbia University, December 3, 2014); Justin Chun-Ting Ho, "How Biased Is the Sample? Reverse Engineering the Ranking Algorithm of Facebook's Graph Application Programming Interface," *Big Data & Society* 7, no. 1 (January 2020); Ben Light, Jean Burgess, and Stefanie Duguay, "The Walkthrough Method: An Approach to the Study of Apps:," *New Media & Society*, November 11, 2016.

Table 1: Overview of Portability Results

Service-to-Service	Transfer possible	Data received	Format	Method
Facebook-Google	Yes	Yes	No	automated, secure, dedicated; API-inconclusive
Twitter-Mastodon	No	Incomplete	N/A	N/A
Twitter-Threads	Threads limited as of November 2023 (not available in some/all E.U. territories at the time of the research)	N/A	N/A	N/A

Table 2: Comparison of portability methods advantage/disadvantage

Method	Advantage	Disadvantage
APIs	Interoperability, authentication, automation, formatting	Complexity, transparency, throttling, management, standards
Download portals	Automation, authentication, user friendly, formatting	Transparency, use value ("one-off"), throttling
Portability portals	Automation, formatting, interoperability, authentication	Limited interoperability

VII. Results

Each method assessed here offered trade-offs for different stakeholders. Advantages for users include automated interfaces that abstract technical processes, returning data in machine-readable formats that can be stored on a computer, uploaded for analysis using software, or ported to another service. Generally, they offered privacy and security through authorization protocols. However, the actual ability to port data was confusing and uneven as illustrated in the summary below.[54]

Service-to-Service Portals

Facebook's dedicated portability tool is accessible from a link on the Facebook Help Center page labeled, "Transfer Your Information to a Service Off of Facebook."[55]

A transfer between Facebook and Google finished within an hour and posts/media from Nov. 2008 to July 23, 2023 were delivered to my Google Drive account, each post in a separate Google doc file, compiled in folders. Media included in posts were in separate folders. The first post, from 2008, appears to be my first post to Facebook.[56]

Some Facebook posts, links, and photos were excluded according to several guidelines enumerated on the portal interface.[57] These include posts that a user: posted on a friend's profile, didn't create, has added to the user's archive or trash, that includes a life event, or that was automatically created when the user changed their profile picture. Absent in the transfer file on Google was everything but the time stamp and text of the post.

The data transfer portals made the exchange of data from Facebook to Google efficient, but provided stand-alone text and images in separate, individual Google docs. It would be difficult and time-consuming for any individual to consolidate the content in order to assess whether the data was complete. This appears more useful for business service-to-service data transfers but, according to statements accompanying the portals, is intended to fulfill the rights of individual subscribers.

---

[54] Each service provided data in JSON, XML, and, to a degree CSV files. But the formats may be user-*unfriendly* (JSON and XML especially can seem like a blob of data) without specific skills that may (or may not) be within reach to many individuals who want to transfer their data.

[55] See https://www.facebook.com/help/2303048582130 63. Facebook also provides a second option, "Download Your Information," which provides a copy of Facebook data to "keep or transfer to another service." This is a download portal like Twitter's.

[56] David Smith, "Analysis of Facebook Status Updates," *Revolutions* (blog), December 29, 2010.

[57] See https://www.facebook.com/help/2303048582130 63.

Download Portals and APIs

I found limited options for moving data between Twitter and Mastodon or Threads.

Threads was released in 2023 and requires subscribers to sign up by using their Instagram account username and password. When it was first introduced, deleting a Threads profile would delete a subscriber's Instagram account, which could be a form of lock-in; this has subsequently been addressed, and Threads can be removed without affecting Instagram.[58]

At the time of the testing, I could not transfer data from any service to Threads because the release was delayed in Europe to ensure that the app's policies are in line with rules governing the combination of user data across services.[59]

Twitter does not provide a dedicated portability tool and, to my knowledge, there is only one mention of portability on the website.[60] Consulting Twitter's Help Center led to Twitter's Privacy Policy and a section called, "How Can I Control My Data." I found a reference in the next page under the heading, "5.1 Access, Correction, Portability," which led to a link to the following: "You can download a copy of your information, such as your Tweets, by following the instructions here." The embedded link led to a platform download page titled, "How to Download Your Twitter Archive."[61]

I originally sought to transfer data from Twitter to Mastodon.[62] Mastodon appears to offer some multi-homing options. But options for portability between Mastodon and Twitter were limited at the time of the research. For one, Twitter followers could not be ported to Mastodon. Additionally, according to reports, dedicated external services that acted on behalf of subscribers were prevented from using the Twitter API necessary for transfers even with the authorization of the user, who initiates the process. The services were detected because they access the subscriber's account by logging in with the subscriber's credentials (email and password). Twitter detected this and blocked the third-party sign-in.

---

[58] https://www.theverge.com/2023/12/14/23953986/threads-european-union-launch-eu-meta-twitter-rival.

[59] I could download the Threads application but not open it. Threads may be available to devices registered to a non-E.U. internet service provider address. The address could be masked with a virtual private network (VPN), which is a service that encrypts the connection of a device and can make it appear to be in a different location. I decided not to test this option, although it may be useful to future research.

[60] I consulted the site before and after Twitter was renamed X. I continue to use the name Twitter for clarity.

[61] https://help.twitter.com/en/managing-your-account/how-to-download-your-twitter-archive.

[62] For this study I did not try to reverse engineer ActivityPub JSON used by Mastodon.

This kind of scenario is at the heart of DMA gatekeeper obligations and provisions for portability in Article 6, which applies to designated gatekeepers. As of the time of this writing, Twitter has not been designated as a gatekeeper, and is thus not subject to the rule. Nevertheless, regardless of intent, the example illustrated the way an API can be equally useful for providing the conditions for secure and interoperable systems and they can for preventing portability. This demonstrates the need for close inspection of access policies built into designs if portability is to fulfill competition goals.

Alternatively, the Twitter platform download, while not a best practice, appeared to comply with existing regulations, yet led to the question of whether a platform download intended for one purpose (access) should be repurposed to fill a portability right.

VIII. Power Dynamics and Subject Rights

Individuals learn about the scope and exercise of rights like data portability from information sources including policymakers, civil society, media outlets, and the industry stakeholders themselves.[63] However, they exercise their rights by using automated interfaces designed and deployed by industry operators, like download and portability portals.[64]

Before their dissemination, users had to email or write to the service providers for their data, which was delivered in a variety of formats, or not at all. Having automated tools at their disposal can thus be an advantage to users if they are designed to make an otherwise potentially difficult process autonomous and easier. Indeed, portability could also be useful to strategies for increasing transparency, fairness, and market competition if consumers can automatically and easily transfer their data from one service to another.

However, portability options have limited advantages for individuals and intermediaries.[65] Rather they appear to be a device for pursuing competition goals through stakeholders. This does not imply a cynical strategy but rather a reflection of policy goals. For example, the quality of data delivered and the mechanisms for it have less significance if the goal is to break up the big controllers, in contrast to giving individuals meaningful options for services that can use personal data.

Stakeholders trying to align policy, enforcement, and economic goals are channeling user rights through a risk minimization framework where the potential for harm should be measured, mitigated, and accepted in exchange for a social benefit. Risk management strategies in this

---

[63] Hughes Roumezin et al., "La portabilité des données en pratique" (Infolab, Février 2018).

[64] Michael Veale, Reuben Binns, and Jef Ausloos, "When Data Protection by Design and Data Subject Rights Clash," *International Data Privacy Law* 8, no. 2 (May 1, 2018): 105–23.

[65] Veale, Binns, and Ausloos.

case are expected to be addressed in part through precautionary regulation, penalties, and "by-design" approaches.[66] For example, APIs are expected to provide prescriptive standards for security in cases where data is considered to be at risk.

APIs are like turnstiles designed to control who has access to particular systems and applications, what information they have access to, and what authentication is required to get that access.[67] This makes them attractive security assets to multiple stakeholders and they serve the purposes of interoperability. Lawmakers have embraced them as a by-design resource despite limitations associated with the quality of their outputs, in particular inconsistent data, as well as their use by industry to unilaterally steer policy and competition goals.[68]

Oversight and enforcement will be enhanced by documenting APIs (as it can be difficult to determine whether an API is part of a portability process without explicit documentation) and preventing APIs from being shielded from scrutiny as trade secrets, as now happens sometimes. Their role and drawbacks deserve more attention and caution from lawmakers. This is worth noting because APIs are important engines of portability and interoperability.[69]

---

[66] Margot E. Kaminski, "The Developing Law of AI: A Turn to Risk Regulation," *The Lawfare Institute-Brookings Institution*, Cybersecurity & Tech, April 21, 2023.

[67] David Berlind, "Why GDPR Compliance Is a Ready-Made Problem for APIs," Salesforce, *MuleSoft Blog* (blog), July 31, 2018.

[68] Megan Brown, "The Problem with TikTok's New Researcher API Is Not TikTok," *New York University Center for Social Media and Politics* (blog), March 1, 2023; Axel Bruns, "After the 'APIcalypse': Social Media Platforms and Their Fight against Critical Scholarly Research," *Information, Communication & Society* 22, no. 11 (September 19, 2019): 1544–66; Taina Bucher, "Objects of Intense Feeling: The Case of the Twitter API," *Computational Culture*, no. 3 (November 16, 2013); CNIL, "API : les recommandations de la CNIL sur le partage de données," November 24, 2023; Justin Chun-Ting Ho, "How Biased Is the Sample? Reverse Engineering the Ranking Algorithm of Facebook's Graph Application Programming Interface," *Big Data & Society* 7, no. 1 (January 2020); Yuanbo Qiu, "The Openness of Open Application Programming Interfaces," *Information, Communication & Society* 20, no. 11 (November 2, 2017): 1720–36; Rebekah Tromble, Andreas Storz, and Daniela Stockmann, "We Don't Know What We Don't Know: When and How the Use of Twitter's Public APIs Biases Scientific Inference," SSRN Scholarly Paper (Rochester, NY, November 29, 2017); Lorenzino Vaccari et al., "APIs for EU Governments: A Landscape Analysis on Policy Instruments, Standards, Strategies and Best Practices," *Data* 6, no. 6 (2021): 59. For a direct analogous assessment of user rights, see: Gabriel Nicholas, "Taking It With You: Platform Barriers to Entry and the Limits of Data Portability," *Michigan Technology Law Review* 27, no. 2 (April 1, 2021): 263–98.

[69] Sara Day Thomson and William Kilbride, "Preserving Social Media: The Problem of Access," *New Review of Information Networking* 20, no. 1 (2015): 261–75.

Self-management, by-design measures automate oversight, which in turn normalizes certain expectations about the way that resources should be allocated based on institutional control and expert systems. Regulators have taken a strong stance on competition issues and data privacy on behalf of citizens. There is still room for turning attention to the resources being developed for effectuating rights.

IX. Conclusion

Clearly, data-producing technologies are now part of our lives, from the way we communicate with each other to the way medicine is delivered. The data being produced is becoming ever more valuable as a resource. This in turn has created a need for strategies that help societies balance the demand for data as a tool for economic growth with values like privacy, access to information, and free expression. One tactic is to implement policy through technical design, referred to as “by-design and default." This study provides an assessment of this approach and its relative advantages and disadvantages for stakeholders in the context of data portability. The results suggest an opportunity to reconsider balancing demands of multiple stakeholders for protection and for portability that do not stop at the limits of technical solutions.

DATA TRANSFER INITIATIVE

A MARKETPLACE FOR DATA PORTABILITY

Chinmayi Sharma

Associate Professor of Law, Fordham Law

TABLE OF CONTENTS

I. Introduction

Following March 6, 2024—the day designated "gatekeepers"[70] of "core platform services" (CPS)[71] such as search engines and social media platforms were required to comply with the mandates set forth in the Digital Markets Act (DMA)—many of us are left wondering how the DMA will change the online ecosystem.[72] In particular, this paper will focus on how the DMA's interoperability and data portability requirements will manifest. The motivation behind the DMA's focus on interoperability and data portability is clear: to challenge the status quo where a few digital giants control significant portions of the online space, limiting consumer choice and stifling innovation.[73] By enabling users to easily switch between services or use multiple services concurrently, the DMA aims to level the playing field, encouraging smaller players to enter the market and compete on equal footing.[74] This legislative push reflects a broader global recognition of the need to address the power imbalances in the digital economy and ensure that the benefits of the digital age are widely accessible.

However laudable the DMA's motivations and goals, its mandates leave something to be desired. In theory, we know that after March 6, 2024, businesses in the European Union (EU) and beyond will be able to interoperate with gatekeeper services and the citizens in the EU and beyond[75] will be able to retrieve their own data. Beyond that, little is known about how this will be accomplished and, more importantly, whether this will achieve the DMA's stated goals. The act's current language offers broad directives without delving into the granular technical standards or frameworks that companies should adopt to achieve these goals. This lack of specificity is not without its merits; it allows for flexibility and innovation in how companies approach the challenge of interoperability and data portability. Yet, this same flexibility grants companies considerable leeway in interpreting and implementing the mandates, potentially

---

[70] Regulation 2022/1925 of the European Parliament and of the Council of 14 September 2022 on Contestable and Fair Markets in the Digital Sector and Amending Directives 2019/1937 and 2020/1828, 2022 O.J. (L 265) 1 (EU) art. 2(1) [hereinafter DMA].

[71] Article 2: Definitions (2), DMA at 28.

[72] DMA *See also* Jay Peters, *How the EU's DMA is Changing Big Tech: All of the News and Updates*, THE VERGE (Feb. 18, 2024), https://www.theverge.com/24040543/eu-dma-digital-markets-act-big-tech-antitrust/archives/2.

[73] See, e.g., Lina Khan, Amazon's Antitrust Paradox, 126 YALE L.J.710 (2017); Robert Reich, Break up Facebook (and While We're at It, Google, Apple and Amazon), GUARDIAN (Nov. 20, 2018, 3:00 AM), https://www.theguardian.com/commentisfree/2018/nov/20/facebook-google-antitrust-laws-gilded-age; https://www.nytimes.com/2017/05/10/technology/techs-frightful-five-theyve-got-us.html

[74] *See Questions and Answers: Digital Markets Act: Ensuring fair and open digital markets*, European Commission (Sept. 6, 2023), https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_2349.

[75] https://verfassungsblog.de/dsa-rest-of-world/.

leading to a fragmented landscape where the underlying objectives of the DMA are not fully realized.[76] Or worse, regulatory arbitrage that further entrenches the status quo as opposed to promoting contestability in markets and making it easier for new and smaller players.[77]

So, on one hand, avoiding highly prescriptive technological mandates allows companies to experiment to discover more efficient and effective solutions than we know today. On the other hand, allowing free rein on implementation approaches invites abuse of the freedom. Luckily, there is a way to straddle the divide.

This paper argues for a three-phased approach to DMA implementation. First, the EU should afford companies flexibility implementing the DMA to foster a robust marketplace of interoperability and data portability solutions tailored to specific domains.

In the second stage, the EU should enlist expert stakeholders to evaluate the merits and shortcomings of the various approaches actually in use to gain an evidence-based understanding of which technological methods achieve the DMA's stated goals. Part of this process will be understanding the various values implicated by interoperability and data portability and establishing some form of ordinal ranking of those values in specific contexts. For example, when opting for expansive, permissionless data portability risks privacy by permitting unverified third parties to access user data, when is the countervailing benefit to a competitive marketplace of greater interest to society than the fastidious protection of privacy rights at all costs? That question must be answered through the evaluation of empirical evidence on user impact by a diverse group of expert stakeholders in a transparent, democratic process that invites public input and builds in the opportunity for amendment over time. Furthermore, that evaluation must always be made *in a context-specific way*.

Finally, in the third phase, the EU should continue to engage a diverse group of independent, expert stakeholders representing various factions and interests in society to identify the most effective technological practices for different sectors based on the empirical evidence analyzed and the normative evaluation of values undertaken. The EU should then encourage gatekeepers to coalesce around these recommended solutions, and ultimately enshrine them through the adoption of technological standards that best serve the public interest.[78]

---

[76] *See, e.g.*, Morgan Meaker, *Developers Are in Open Revolt Over Apple's New App Store Rules*, WIRED (Feb. 12, 2024), https://www.wired.com/story/developers-revolt-apple-dma/; Matt Burgess, WhatsApp Chats Will Soon Work With Other Encrypted Messaging Apps, Wired (Feb. 6, 2024) ("This move for interoperability will, on the one hand, open the market, but also maybe close the market in the sense that now the bigger players are going to have more decisional power").

[77] Julie Cohen, Infrastructuring the Digital Public Sphere, 25 YALE J.L. & TECH. SPECIAL ISSUE 1, 36 (2023).

[78] Article (1)(96), DMA at 24; Article 48, DMA at 60. It has similar language for NIICS interoperability: "It should be possible for the Commission, if applicable, to consult the Body of European Regulators for

II. The digital markets act: in theory and in practice

This section will explore the DMA's objectives, the current state of its interoperability and data portability mandates, and the potential benefits and pitfalls of its approach. This section will also examine the broader implications of flexibility versus specificity in regulatory mandates to lay the groundwork for a proposal that attempts to harness the advantages of both.

A. Interoperability And Data Portability Mandates

Designing DMA implementation policies must recognize and honor the legislation's sweeping nature and its ambitious agenda. Passed in November 2022, the DMA's effect will be felt globally in the coming years.[79] Its goal is simple: to regulate the digital market to ensure the safety of users and to combat the anticompetitive tendencies of dominant digital platforms, from social media and app stores to search engines and advertisement. Achieving this goal is far more complicated.[80] Two of the DMA's key enforcement mechanisms are its competition-friendly and user-empowering interoperability and data portability requirements. However, interoperability and data portability are capacious concepts that can be achieved in many ways, each with their own trade-offs.[81] The DMA does not, however, prescribe the specific way in which it wants companies to comply with its provisions, leaving that decision to each gatekeeper.[82] Failure to comply triggers severe penalties: from fines up to 20% of annual worldwide turnover for repeat offenses and structural remedies such as divestment of parts of the business as a last resort for systematic failure to comply. In other words, however complex compliance might be, noncompliance is not an option.

Electronic Communications, in order to determine whether the technical details and the general terms and conditions published in the reference offer that the gatekeeper intends to implement or has implemented ensures compliance with this obligation." Article (1)(64), DMA at 17.

[79] Daphne Keller, The EU's new Digital Services Act and the Rest of the World, Verfassungsblog (November 7, 2022) (available at https://verfassungsblog.de/dsa-rest-of-world/).

[80] Chris Riley, Data is binary, but that doesn't make portability simple., LinkedIn (August 29, 2023) (available at https://www.linkedin.com/pulse/data-binary-doesnt-make-portability-simple-chris-riley/).

[81] Ian Brown, Making interoperability work in practice: forms, business models and safeguards, Ada Lovelace Institute (available at https://www.adalovelaceinstitute.org/blog/making-interoperability-work-practice/). See generally BENNETT CYPHERS & CORY DOCTOROW, ELEC. FRONTIER FOUND., PRIVACY WITHOUT MONOPOLY: DATA PROTECTION AND INTEROPERABILITY (2021), https://www.eff.org/document/privacy-without-monopoly-data-protection-and-interoperability.

[82] Article(7)(4), DMA at 37. *See also* Matt Binder, *Meta, Microsoft Take on Apple and Lobby EU to Reject New App Store Terms*, MASHABLE (Feb. 21, 2024), https://mashable.com/article/meta-microsoft-lobby-eu-apple-app-store-dma.

To give meaning to the DMA's provisions, policy must be designed with its subjects in mind: the gatekeepers.[83] In other words, the DMA seeks to regulate market players that already wield a massive amount of influence—curbing their anticompetitive behavior will require anticipating and countering their inevitable attempts to maintain a stranglehold on the market.[84] On the other hand, these entities are responsible for a large amount of the technological progress that has advanced human welfare—stifling their ability to experiment and innovate would deprive the public of socially beneficial progress. So, who are these players? The DMA defines gatekeepers as entities providing CPS, such as search engines or operating systems, that meet certain size thresholds and market impact criteria.[85] In September 2023, the EU published a list of companies that qualify as gatekeepers, each providing different CPS: Alphabet, Amazon, Apple, ByteDance, Meta, and Microsoft.[86] By identifying these regulated entities as "gatekeepers," the DMA recognizes that its targets have "structural domination of multiple, interlocking domains of economic and social activity."[87] In essence, gatekeepers operate little self-contained fiefdoms, or walled gardens, within which users must cede most of their power to the platforms to access crucial services.

The DMA redistributes power away from gatekeepers by mandating interoperability and data portability.[88] Interoperability requires gatekeeper platforms to ensure their systems can work

---

[83] Article 2: Definitions (1), DMA at 28; Article 2: Definitions (2), DMA at 28.

[84] *See* SCOTT MORTON, BRUEGEL, THE CHICKEN AND EGG PROBLEM 1 (2024), https://www.bruegel.org/sites/default/files/2024-01/WP%2002%202024_0.pdf; Lauren E. Willis, *Consumer-Facing Competition Remedies: Lessons from Consumer Law for Competition Law*, 2023 UTAH L. REV. 887 (2023) ("One way the banks attempted to resist competition was to make the customer authorization process—through which consumers could authorize third-party providers to access their data—onerous: "Banks required customers to navigate as many as 12 screens of intimidating warnings and caveats," and "used an out-of-date browser-based process that required that users log in repeatedly.").

[85] Article 3: designation of Gatekeepers (1), DMA at 30.

[86] "Commission designates six gatekeepers under the Digital Markets Act," European Commission (September 6, 2023) (available at https://digital-markets-act.ec.europa.eu/commission-designates-six-gatekeepers-under-digital-markets-act-2023-09-06_en#:~:text=Commission%20designates%20six%20gatekeepers%20under%20the%20Digital%20Markets%20Act,-European%20Commission&text=Today%20(6%20September%202023)%20the,Digital%20Markets%20Act%20(DMA).

[87] Julie Cohen, Infrastructuring the Digital Public Sphere, 25 YALE J.L. & TECH. SPECIAL ISSUE 1 (2023); Nicolas Petit, The Proposed Digital Markets Act (DMA): A Legal and Policy Review, 12 J. EURO. COMP. L. & PRACTICE 529 (2021).

[88] Although today, the only core platform service for which the DMA dedicates an article for imposing interoperability requirements is for number-independent interpersonal communications services (NIICS). "Number-independent interpersonal communications service" is defined as "an interpersonal communications service which does not connect with publicly assigned numbering resources, namely, a

seamlessly with those of current or future competitors.[89] Perhaps the most visible outcome of the DMA will be the fact that Meta's Messenger users will be able to interact directly with Meta's WhatsApp users[90] without having to leave the platform or make a new account, in the same way that a Gmail account can email a Hotmail account today. Data portability mandates complement the DMA's interoperability goals by reducing the switching costs associated with leaving a platform and giving users more transparency into and control over the data collected on them. The DMA empowers users or third party businesses that have user consent to access continuously and in real time all data related to a user.[91]

Together, these mandates seek to reduce barriers to entry and exit. They attempt to level the competitive playing field by preventing companies from leveraging their exclusive control over data hoards to box out competitors or trapping users in their walled gardens to access the network they've built or to benefit from a gatekeeper's other bundled services.[92] Implemented optimally, these mandates would allow competition to flourish, giving users more options for CPS that better serve their needs.

B. Methods to the Madness

While conceptually straightforward, the implementation of the DMA's interoperability and data portability requirements is a multi-dimensional problem. Interoperability and data portability are

---

number or numbers in national or international numbering plans, or which does not enable communication with a number or numbers in national or international numbering plans." Article 2, point (7), of Directive (EU) 2018/1972, https://eur-lex.europa.eu/eli/dir/2018/1972/oj; Article(2)(9), DMA at 28.

[89] Article (2)(29), DMA at 30.

[90] Matt Burgess, WhatsApp Chats Will Soon Work with Other Encrypted Messaging Apps, Wired (Feb. 6, 2024), https://www.wired.com/story/whatsapp-interoperability-messaging/.

[91] Article 6 paragraph 9 and 10.

[92] Lina M. Khan, The Separations of Platforms and Commerce, 119 COLUM. L. REV 973 (2019), https://columbialawreview.org/content/the-separation-of-platforms-and-commerce/ ("If a standard choice faced by a dominant platform is whether to grant rival complementors access to its network and charge a fee to extract some of their revenue or to exclude all rival comple-mentors and sell the service itself, then digital markets seem to tip the balance in favor of the latter. This is because digital platforms are making an ecosystem play: By bundling different services and portals, a platform can heighten switching costs and collect more user data by tracking individuals across services, both of which amount to a lucrative strategy."); Kevin Bankston, How We Can 'Free' Our Facebook Friends, NEW AMERICA (June 28, 2018), https://www.newamerica.org/weekly/edition-211/how-wecan-free-our-facebook-friends; Peter Swire, Markets, Self-Regulation, and Government Enforcement in the Protection of Personal Information, in Privacy and Self-Regulation in the Information Age, DEP'T OF COM. (June 1997), https://papers.ssrn.com /abstract=11472.

not monoliths—there are many technological approaches to achieving both of them and each approach comes with its own trade-offs. Some approaches are going to be better than others at striking the right balance of values such as open access, efficiency, security, privacy, and transparency, which will make some approaches more in line with the DMA's values and goals. The DMA's success ultimately hinges on the specific technological way in which gatekeepers implement the mandates. This section will review the menu of technological approaches to moving data and connecting platforms and will explore the degree to which each approach promotes public interest values. The broad categories of technological approaches include: provider-specific solutions, interoperability by design, and decentralized architecture solutions.

Provider-specific solutions refers to technologies designed specifically by or for a particular gatekeeper to make that gatekeeper's data or functionalities available.

> ***Manual ad-hoc exports*** allow for the one-time transfer of user data from one entity to another.[93] This can be as simple as a user downloading their profile information from a social media account before deleting it or regularly porting songs added to a playlist on one streaming service to a playlist on another service.
>
> > **Technologies: CSV exports, JSON files, SQL/NoSQL queries, third party data warehouses or cloud storage provider interfaces or tools.**
> >
> > **Pros:** Gives users more control over their data, provides visibility into data platforms collect about users, reduces user barriers to switching platforms, alleviates competitor data disadvantages.
> >
> > **Cons:** Highly manual process, unsuitable for bulk data, and data formats lack standardization, which increases burden on competitors using gatekeeper data and undercuts the goal of improving competition and innovation.
>
> ***Programmatic Data Exports*** are technologies by which companies can request data exports from the gatekeeper at specific intervals or when a specific event happens, like new user-generated content. This allows third-parties to import

[93] GABRIEL NICHOLAS & MICHAEL WEINBERG, ENGELBERG CTR. ON INNOVATION L. & POL'Y, DATA PORTABILITY AND PLATFORM COMPETITION: IS USER DATA EXPORTED FROM FACEBOOK ACTUALLY USEFUL TO COMPETITORS? 7–9 (Nov. 2019), https://www.law.nyu.edu/sites/default/files/Data%20Portability%20and%20Platform%20Competition%20-%20Is%20User%20Data%20Exported%20From%20Facebook%20Actually%20Useful%20to%20Competitors.pdf.

new user data at regular intervals or when new data is available without having to proactively check for new data.

> **Technologies: Webhooks, Application Programming Interfaces (APIs), automated extract, transform, load (ETL) tools.**
>
> **Pros:** Provides continuous data flows, end users don't have to port data over themselves, suitable for bulk data, reduces burden on both third party and gatekeeper system resources, easy to implement, places burden of data portability on the gatekeeper.
>
> **Cons:** Resource intensive and complex for gatekeeper to build, security concerns, dependency on gatekeeper services, and limited by event definitions.

Live data-streaming, on the other hand, allows real-time, machine-readable data flow, which facilitates multi-homing, or the concurrent use of two different platforms through specialized protocols. This enables third-party platforms to maintain a live connection with the gatekeeper platform, enabling services like spending dashboards or tracking health data across multiple apps and wearables.

> **Technologies: HTTP/2, WebSockets, RTP, RTSP protocols.**
>
> **Pros:** Suitable for bulk data, provides continuous data flow through a live connection, provides visibility into platform data practices, provides competitors with ability to access user-generated data in real time to build complementary products, supports data portability scaling to accommodate large numbers of concurrent streams, places burden of enabling data portability on the gatekeeper.
>
> **Cons:** Resource intensive and complex for gatekeeper to build, high resource consumption for gatekeeper to maintain, potential for increased latency over poor connections or during peak usage, gatekeeper control over connection and conditions of access, complex for third party to implement, and privacy and security concerns around unfettered data access by unknown third parties.

Functionality availability refers to the way gatekeepers open third-party access to valuable functions they provide, such as content moderation, CSAM filtering, security, and authentication—in other words, allowing third parties to use functions built and hosted by gatekeepers in providing services to their own end users. These complex and resource-intensive functions can create barriers to entry for third parties, affecting user experience or even causing harm.

Technologies: APIs.

Pros: Reduces barriers to entry by allowing competitors to benefit from gatekeeper functions, shifts cost of complex functions like authentication on the gatekeeper who hosts the function, protects user privacy and system security by keeping sensitive functions with capable and resource-rich gatekeepers, allows third parties to dedicate resources saved to innovation.

Cons: Resource intensive and complex for gatekeeper to build, high resource consumption for gatekeeper to maintain, gatekeeper control over conditions and scope of function accessibility, complex for third party to implement, and users must make account with gatekeeper to benefit from this function availability, advantaging the gatekeeper.

Interoperability by design facilitates interoperability, and within that, data portability, as a default feature of platforms based on the way their architecture is built.

Code availability refers to the way gatekeepers can make valuable functions available to third parties by providing access to the codebase for the function, either through controlled licensing or open sourcing. This allows third parties to use or change the library, the third party to customize the function to serve their purposes while maintaining interoperability. For example, Canvas, a teaching and learning software, offers a hybrid approach as both an internally hosted product and an open-source project.[94] Users can pay Canvas for access to functionality and data Canvas hosts or, alternatively, use the open-source code for personal or commercial purposes.[95] This approach lets entities choose between the efficiency benefits of hosted functions and the autonomy and privacy of self-hosted instances.

Technologies: Open source code.

Pros: Grants independence from gatekeeper, provides ability to customize technology to better serve third party needs, saves time and resources of building the library, which can be reinvested in innovation, improves security through "many eyes make all bugs shallow," and maintains interoperability with gatekeeper instance by design.

Con: Resource-strapped third parties may not be able to host function, gatekeeper loses incentive to update or innovate function, third parties

[94] INSTRUCTURE OPEN SOURCE, https://code.instructure.com/.

[95] Bitnami, *Canvas LMS*, BITNAMI VIRTUAL MACHINE, https://bitnami.com/stack/canvaslms/virtual-machine (last visited May 4, 2023).

can alter function code to reduce efficiency, privacy, or security, code available to malicious actors.

Standard protocols are sets of instructions in code that enable interoperability and data portability retroactively and prospectively, ensuring compatibility with all other platforms adopting the same standard.[96] Unlike provider-specific solutions, they are developed in a technology-neutral manner by open, industry, or government standards bodies. Open standards are publicly available and established by independent bodies, such as the IETF. Industry standards are developed through multi-stakeholder processes between industry members. Government standards are established by officials for public use. Standards can provide functionality, such as network connectivity, or uniformity, such as data formats.

Technologies: Protocols such as HTTPS, OAuth, SMTP, RSS, TCP or data formats such as JSON or XML.

Pros: Grants independent and permissionless ability to build interoperable systems, easy to implement, provides transparency into system architecture, some degree of democratic legitimacy in process of establishing standard, improves security through "many eyes make all bugs shallow."

Con: Limits ability to innovate on the standard component, requires quorum of players in ecosystem to adopt standard for it to be useful, challenging and slow process of updating standards, consensus to establish standard may not be possible.

Middleware fosters interoperability retrospectively by sitting between existing platforms and facilitating communication, overcoming differences in technologies, protocols, and data formats. Like standard protocols, middleware can also be open, proprietary, or government-built, but it is most often developed by third-party intermediaries.

Technologies: Enterprise Service Bus (ESB), Message Oriented Middleware (MOM), Data Integration Tools, Object Request Brokers (ORB), API Management Platforms, Integration Platforms as a Service (IPaaS), and Database Middleware.

Pros: Enables interoperability while minimizing burden on gatekeeper to change internal system architecture, reduces burden on third party who

[96] Chris Riley, *A framework for forward-looking tech competition policy*, MOZILLA WORKING PAPER (September 9, 2019), https://blog.mozilla.org/netpolicy/files/2019/09/Mozilla-Competition-Working-Paper.pdf.

can use middleware to interoperate, modular components, creates secondary market for more efficient middleware solutions.

Con: Difficult to scale given middleware must accommodate each platform's specific data formats or system design, brittle in that changes in platform systems may break middleware connection, inefficient transmission of data, exposes user data to new entity if middleware designed by a third party, chain of data transfers and transformations may introduce privacy and security risks.

Decentralized architecture solutions can facilitate data portability and interoperability through more user-centric system design, as compared to centralized platforms.

Distributed networks are networks of computers that collaborate as a single system, sharing resources and processing power. Each computer node communicates and coordinates with others to perform tasks, decentralizing communication and decision-making to prevent any single node from gaining excessive power. The system is fault-tolerant, functioning even if one node fails.

Technologies: Blockchain, Distributed Computing Platforms, Distributed Databases.

Pros: Distributed system avoids decision chokepoints, data portability across nodes built in by design, encourages democratic control over decision making, and redundancy of information on nodes makes system resilient to fraud or manipulation.

Con: Undermines privacy with information spread across several systems making edits or takedowns virtually impossible, decision-making requires network consensus, frustrates governance with inability to enforce against bad actors, resource-strapped third parties may not be able to join, difficult to scale given resource-intensity, and decision-making processes favor largest or most resourceful players.

Federated networks are networks of computers, each controlled by different entities, built on the same stack of protocols. Each network of computers is maintained by an entity that exercises control over that network's data, community rules, and resources, customizing their instance as they see fit while

maintaining interoperability and the ability to transport data with other instances in the federated network.[97]

Technologies: Fediverse, Matrix, PeerTube, Nextcloud, Friendica, and Hubzilla.

Pros: Users incentivized to maintain interoperability to enjoy network benefits, flexibility to customize instance while maintaining baseline interoperability, ability to engage outside instance while maintaining control over data visibility, and impossible to exclude entities from federated network entirely.

Con: Resource-intensive to run an instance of a federated network, administrators may lack skills necessary to run instance, accessibility of network susceptible to malicious actors, content moderation difficult to enforce, challenging to scale, may exacerbate societal fragmentation and echo chambers.[98]

Peer-to-Peer networks are networks of computers that communicate directly with each other, without the need for a central server or authority. In a peer-to-peer system, each node can act as both a client and a server, sharing resources and information with other nodes—information is distributed across nodes, but each node does not need to maintain a complete copy of all the network's information and each node has equal authority to make decisions.

Technologies: BitTorrent, InterPlanetary File System (IPFS), DAT Protocol, eMule/eDonkey, and Gnutella.

Pros: The benefits of a P2P system largely reflect those of a distributed and federated system in the lack of a central point of authority, but P2P networks have the added benefit of the ability to act quickly without consensus of the network, and each node maintains more control and autonomy over its functioning.

Con: The shortcomings of a P2P system also reflect those of a distributed and federated system in being resource-intensive. Because P2P nodes have more agency and autonomy than nodes in distributed or federated

---

[97] European Data Protection Supervisor, Tech Dispatch: Federated Social Media Platforms (2022), https://www.edps.europa.eu/system/files/2022-07/22-07-26_techdispatch-1-2022-federated-social-media-platforms_en.pdf.

[98] BENNETT CYPHERS & CORY DOCTOROW, ELEC. FRONTIER FOUND., PRIVACY WITHOUT MONOPOLY: DATA PROTECTION AND INTEROPERABILITY 27–29 (2021), https://www.eff.org/document/privacy-without-monopoly-data-protection-and-interoperability.

networks, the network as a whole is more susceptible to security and privacy risks.

These categories broadly cover the various ways systems achieve interoperability and data portability. Each have their own advantages and disadvantages. To ensure the DMA meets its stated goals, the correct technological approach to interoperability and data portability must be selected *for a specific use case.* This is a domain-specific evaluation, which means the right answer for banking services is unlikely to be the same as for messaging apps. In short, the method to the madness matters.

III. A marketplace of interoperability and data portability solutions

Although March 6, 2024 marked the first deadline for implementation of the DMA's interoperability and data portability requirements for gatekeepers providing messaging services, most of the DMA's provisions have yet to be fleshed out. The opportunity to craft a nuanced approach to interoperability and data portability mandates has not passed, and the EU should walk the tight rope between under-specification and over-prescription of technological implementation of mandates.

Each implementation of interoperability and data portability mandates—be it standard protocols, APIs, webhooks, or any of the myriad emerging technologies—comes with its own distinct trade-offs. These trade-offs can impact everything from user privacy and market competition to innovation potential and operational efficiency.[99] Crucially, the true nature and impact of these trade-offs cannot be fully understood through theoretical models or assumptions alone. They must be evaluated in the context of actual deployment and usage, where real-world data can be gathered and analyzed.

This necessitates the fostering of a competitive marketplace of options, where various entities are encouraged to implement interoperability and data portability in diverse ways. Such an environment will not only reveal the practical implications of different approaches but will also stimulate innovation as entities vie to develop the most effective and user-friendly solutions.

---

[99] See BENNETT CYPHERS & CORY DOCTOROW, ELEC. FRONTIER FOUND., PRIVACY WITHOUT MONOPOLY: DATA PROTECTION AND INTEROPERABILITY (2021), https://www.eff.org/document/privacy-without-monopoly-data-protection-and-interoperability; Mitch Stoltz, Andrew Crocker & Christoph Schmon, *The EU Digital Markets Act's Interoperability Rule Addresses an Important Need, But Raises Difficult Security Problems for Encrypted Messaging*, ELEC. FRONTIER FOUND. (May 2, 2022), https://www.eff.org/deeplinks/2022/04/eu-digital-markets-acts-interoperability-rule-addresses-important-need-raises.

A. The Risks of Under-Specification and Over-Prescription

On one side of the chasm lies the risk of under-specification, which, while offering flexibility, might inadvertently empower gatekeepers to shape the mandates' fulfillment in a way that best serves their interests rather than the public's. Under-specification could result in a technological implementation that meets the letter of the law while circumventing its spirit. Gatekeepers, left to their own devices, may opt for solutions that place undue burdens on third parties who wish to access or port data. For instance, a gatekeeper could implement a data portability solution that exports user data in a proprietary format that is technically compliant but practically onerous for competitors to utilize, effectively nullifying the intent of the DMA.[100] Similarly, gatekeepers may provide APIs that offer only the most limited data access or impose strict rate limits that stifle third party innovation. Perfunctory compliance could disadvantage third parties, but regulatory arbitrage could go even further and advantage the gatekeepers, reifying their dominance. This could involve collusion with non-threatening partners, demanding unnecessary but competitively valuable data from downstream third parties, or the use of privacy and security concerns as pretext to avoid compliance.

Conversely, an overly prescriptive approach risks stifling innovation by mandating specific technological solutions like APIs or programmatic data exports. Such rigidity in the face of rapid technological evolution can lock the market into suboptimal standards and prevent the natural development of potentially superior solutions. The early internet is a testament to the power of a less prescriptive approach; the organic growth of protocols like TCP/IP and HTML, driven by a community of innovators rather than a central authority, was instrumental in the explosive and transformative growth of the global internet. Telecommunications, on the other hand, offers instances where over-prescription led to missed opportunities. For example, in the telecommunications sector, strict adherence to certain network standards has at times delayed the adoption of advanced technologies like VoIP or 4G, which when allowed to flourish, have revolutionized the industry.

Lastly, the DMA is navigating uncharted waters, and the market has not yet had the opportunity to fully explore and understand the trade-offs of different interoperability and data portability solutions. The pros and cons of various approaches included in this paper are untested hypotheses—given the centralized nature of the internet, there is little by way of empirical evidence of the impact of different interoperability and data portability methods on users and

---

[100] For example, Facebook's data portability feature did not lead to competitors using such data. GABRIEL NICHOLAS & MICHAEL WEINBERG, ENGELBERG CTR. ON INNOVATION L. & POL'Y, DATA PORTABILITY AND PLATFORM COMPETITION: IS USER DATA EXPORTED FROM FACEBOOK ACTUALLY USEFUL TO COMPETITORS? 7–9 (Nov. 2019), https://www.law.nyu.edu/sites/default/files/Data%20Portability%20and%20Platform%20Competition%20-%20Is%20User%20Data%20Exported%20From%20Facebook%20Actually%20Useful%20to%20Competitors.pdf.

the market in specific domains. An over-prescriptive regulatory stance would preclude the chance to empirically evaluate the real-world impacts of these technologies across different sectors, missing out on valuable data that could inform future legislation.

B. Phase One: Fostering a Marketplace of Solutions

The path forward requires a balanced regulatory approach that avoids the extremes of under-specification and over-prescription. It must provide enough detail to prevent gatekeepers from undermining the DMA's goals while allowing the flexibility for innovative solutions to emerge and adapt to the needs of a dynamic digital market. This approach should foster an environment where the best solutions for interoperability and data portability can be identified, tested, and adopted in the long term. It must also acknowledge the varied nature of digital domains and the necessity for tailored technological responses. The EU can stimulate this variety by:

> ***Industry Stakeholder Groups***: Facilitate the formation of industry stakeholder groups that include representatives from large tech firms, SMEs, academia, and civil society to discuss and brainstorm potential interoperability and data portability solutions.
>
> ***Research Funding*:** Allocate grants and funds specifically for research and development in the field of data sharing technologies, encouraging innovation and experimentation, especially across competitors in specific sectors.
>
> ***Technical Solutions from Government*:** Develop and provide open-source tools and platforms that can be used as a baseline for interoperability and data portability, reducing the initial investment required for smaller companies to compete.
>
> ***Incentives for Open Standards*:** Create incentives for companies that adopt open standards and contribute to their development, encouraging solutions that promote wider interoperability.
>
> ***Oversight and Enforcement*:** Establish a regulatory oversight body to monitor the implementation strategies adopted by gatekeepers, ensuring compliance while actively preventing collusion and fostering solution diversity. This body should also have the power to penalize non-compliant or anti-competitive behaviors.

Through these recommendations, the EU can catalyze the creation of a competitive marketplace where various interoperability and data portability solutions emerge.

C. Phase Two: Empirically Evaluate Effectiveness

To walk the tightrope between over-prescription and under-specification, it is crucial to tailor the DMA's mandates for interoperability and data portability to the nuanced demands of different technological domains. This precision can only be achieved through rigorous empirical analysis of real-world usage and experiences—case studies provided by the marketplace for solutions fostered in phase one.[101]

Without concrete data, any rush to mandate specific technological implementations would be premature. Current data is insufficient to conclusively favor standard protocols, webhooks, or any other method as the best means of achieving interoperability and data portability in a specific context. There is a dearth of research in the space. For over a decade, the internet's architecture has been dominated by siloed platforms with limited incentive or requirement to facilitate data portability or interoperability, resulting in few examples to study and learn from. Without more examples, we won't have more data. Without more data, we cannot evaluate different approaches. Without evaluations of approaches, we cannot identify the optimal implementation strategies. Without that, the DMA won't achieve its stated goals.

The EU can approach this phase by:

> ***Data Collection and Analysis*:** Implement mechanisms to collect data on the performance of various interoperability and data portability solutions, analyzing which approaches best align with the DMA's goals. Data analysis must be founded on principles of multi-stakeholder collaboration and transparency. To that end, the data should be made available to the public for review and to independent researchers for analysis, to ensure a wide array of perspectives are considered in the evaluation process. The EU should invite independent, expert stakeholders from a diverse range of domains and experiences to ensure the data is reviewed by, to the degree possible, a group representative of the broader population.
>
> ***Evaluation of Trade-Offs*:** Engage in a transparent, public process to evaluate, based on the empirical data collected, the advantages and shortcomings of various technical approaches to interoperability and data portability. Between

[101] *See* Aline Blankertz, *The EU's Experimental Approach in Overhauling Competition Rules*, BROOKINGS (April 414, 2022), https://www.brookings.edu/articles/the-eus-experimental-approach-in-overhauling-competition-rules-digital-markets-act-dma/ ("[I]t would be beneficial if the Commission articulated more clearly the objectives of the interoperability and other requirements. By translating those objectives into meaningful indicators, firms, regulators and the public would be able to assess whether the DMA is working as intended or if it needs adjusting."). *See also* Peter Swire, *The Portability and Other Required Transfers Impact Assessment (Port-IA): Assessing Competition, Privacy, Cybersecurity, and Other Considerations*, 6 Georgetown L. Tech. Rev. 57 (2022).

goals such as lower barriers to market, innovation, security, privacy, transparency, efficiency, usability, accessibility, and human rights values, each technical solution will inevitably make certain trade-offs. Leaving the decisions regarding trade-offs to the market is ill-advised. The EU should therefore engage the diverse group of expert stakeholders, alongside robust public involvement, to recognize the trade-offs inherent in various approaches to interoperability and data portability and generate principles to guide the private sector's deliberative process around how to approach these trade-offs. This guidance must be done in a domain or use-case specific way, because a universal ranking of values is impossible.

***Public Input and Comment*:** Solicit public input through comment periods, town hall meetings, and online platforms to ensure the process reflects the democratic will and is not overly influenced by industry stakeholders. Specifically, public input regarding the effectiveness, usability, and accessibility of various approaches will be essential to ensure technical solutions are actually helpful to the public they are intended to serve. Additionally, public input regarding the prioritization of various goals is essential to an ethical and democratic evaluation of appropriate trade-offs.

Pilot Programs and Case Studies: Run pilot programs to test different interoperability and data portability models in various contexts, gathering concrete evidence of their efficacy and impact. This should augment data collected from actual private sector adoption of technical approaches, as the market may not choose to implement technical approaches that are promising and potentially the most in line with the DMA's goals.

This evidence-based approach ensures that the EU's understanding of the trade-offs and benefits of different technological solutions is grounded in real-world data and public consensus, rather than theoretical models or industry preferences.

D. Phase Three: Towards Standardization of Solutions

The final phase is the crystallization of the DMA's objectives into concrete technological standards. This phase is predicated on the insights gained from the previous stages and involves:

***Identifying Promising Solutions*:** Using the data and insights gained, identify the technical approaches to interoperability that best achieve the DMA's purpose while satisfying the trade-off principles produced by the multi-stakeholder process. This list need not be exhaustive, but rather a government-endorsed output of a transparent, democratic process evaluating concrete empirical data to identify how best to accomplish the DMA's mandates.

***Developing Standards*:** Once there is a clear understanding of the most effective practices, encourage the development of nascent standards implementing these technical solutions. Encourage the adoption of these standards in practice, collect data assessing the degree to which these standards do in fact support the hypotheses that these technical approaches to interoperability and data portability best satisfy the DMA's goals while making acceptable trade-offs.

***Enshrining Standards*:** Once the iteration process is done and sufficient evidence proves that the standards are successful and durable in actual use, formalize them through the appropriate legislative or regulatory processes. Develop guidelines and incentives for gatekeepers to adopt these technical solutions, moving the industry towards a consensus on optimal solutions. Ensure the formalized governance of these standards includes safe harbors or other mechanisms to protect conscientious actors as the field evolves as well as processes by which guidance around standards and trade-off analyses can be regularly reevaluated and updated if new evidence or landscape developments emerge.

The timing for this phase is inherently uncertain and must be responsive to the evolution of the marketplace and technological advancements. It will depend on when a clear picture emerges from the data and public engagement processes of the first two phases, ensuring that when standards are set, they are based on robust evidence and democratic legitimacy.

Through this three-phased approach, the DMA's implementation can be both dynamic and grounded, fostering innovation while steering the market towards solutions that are in the best interest of all stakeholders. It is a journey that will require patience, flexibility, and a willingness to adapt as the digital market continues to evolve.

IV. Conclusion

In conclusion, through this three-phased approach, the DMA's implementation becomes a dynamic and grounded endeavor, one that nurtures innovation and guides the market toward solutions that align with the broader public interest. This journey, undoubtedly complex, requires patience, a willingness to embrace flexibility, and a readiness to adapt to the ever-changing contours of the digital landscape. It is a commitment not to a fixed end-point but to an ongoing process of discovery, assessment, and refinement that seeks to balance the diverse and sometimes competing interests at play in the digital market.

DATA TRANSFER INITIATIVE

Implementing “Continuous and Real-time” Data Portability with Webhooks

Chand Rajendra-Nicolucci
Director of Product, University of Massachusetts Amherst

TABLE OF CONTENTS

Requirements for "continuous and real-time" data portability hold the potential to transform data portability from a weak, seldom-used right into a powerful tool for competition and user agency by making it easier to build useful services on top of ported data.[102] Ironing out a workable vision for what "continuous and real-time" means in practice is key to realizing that potential. This brief aims to do just that, outlining a model for continuous and real-time data portability that is simple and flexible. It's based on a well-known approach to data transfers called "webhooks." Webhooks can be applied across domains efficiently, offering a general-purpose solution to continuous and real-time data portability.

Although webhooks are not necessarily the only way to implement continuous and real-time data portability, they are a promising approach that demonstrates the feasibility of continuous and real-time data portability in practice. And while there are challenges to successfully implementing webhooks at scale, they can be overcome with quality software development practices and technical standardization.

I. Evolving portability

Traditionally, data portability has meant a bulk, one-time transfer of a user's data. Delays of minutes to days are typical, and there is no expectation of, nor provision for, additional transfers. Usually the transfer mechanism is a tool that allows users to download a copy of all of their data associated with a service. This approach significantly limits portability's usefulness.[103][104] With bulk, one-time transfers you can seed a recipient service with someone's data, but if that person continues to use the source service, it's difficult to keep the recipient service up to date. Ported data that quickly goes out of date is rarely useful for recipient services, so for the recipient to offer useful services based on the ported data, the user would have to move their relevant activity to the recipient service going forward. However, it's rare that a user can, or wants to, leave a source service completely. Rather, they often either have to continue using a source service because of lock-in—e.g., network effects, contracts—or want to continue using a source service because of functionality it provides—e.g., a better user interface, innovative features.

---

[102] See, e.g., European Union, *Digital Markets Act*, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R1925.

[103] Jurre Reus and Nicole Bilderbeek, *Data portability in the EU: An obscure data subject right*, IAPP (2022), https://iapp.org/news/a/data-portability-in-the-eu-an-obscure-data-subject-right/; European Commission, *Data protection as a pillar of citizens' empowerment and the EU's approach to the digital transition - two years of application of the General Data Protection Regulation*, 2020, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0264.

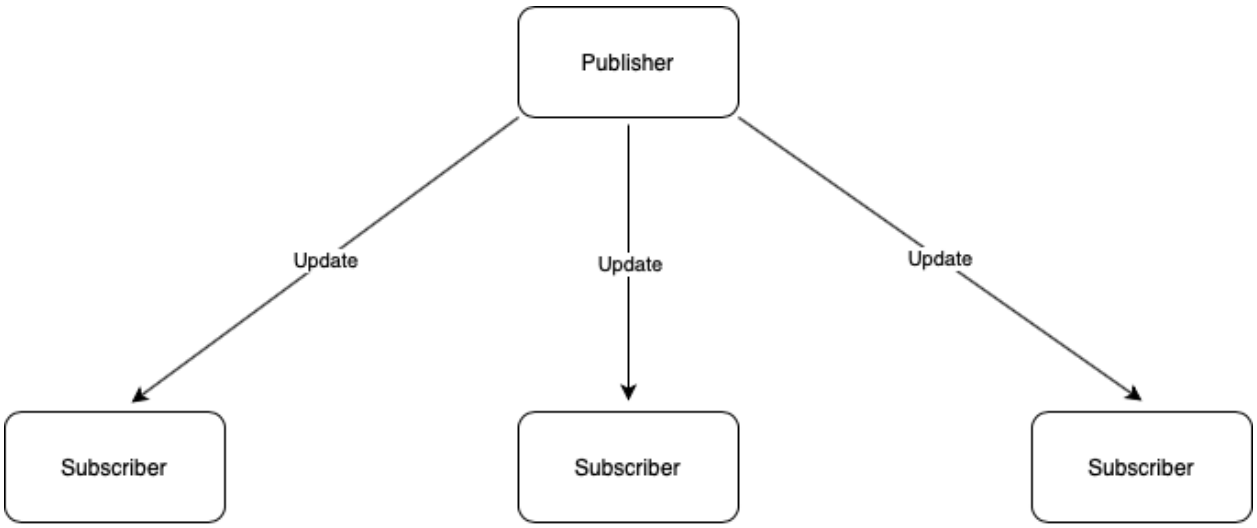
[104] It's worth pointing out that bulk portability can be quite useful for recipient services where future usage of a source service is immaterial, like personal archives. However, that's rarely the case.

Continuous and real-time portability would significantly increase the usefulness of portability by making it possible for recipient services to stay up to date with source services. People could continue using source services if they need to or want to, while also taking advantage of recipient services that quickly and consistently mirror updates to their data.

Implementing continuous and real-time portability requires the source service to provide an interface for updates and either the source or the recipient to initiate updates. When the recipient initiates an update, it's called a "pull".[105] When the source initiates an update, it's called a "push." This brief outlines a push-based model for implementing continuous and real-time portability, where sources publish updates to "subscribed" recipients as they occur.

II. Webhooks

Webhooks are a form of push-based data transfer. On one side are publishers, who publish data updates. On the other side are subscribers, who subscribe to data updates. To illustrate, imagine a social media service, foo.social, that supports webhooks for users' feeds. When a post is added to a user's feed, foo.social sends an update with information about the new post to the webhook's subscribers. Subscribers can then take a number of actions based on that update, like adding the new post to a replication of the user's feed or notifying the user via email if the new post is from a list of important accounts.



Simplified diagram of webhooks' push-based data transfer.

Webhooks are usually part of a service's API. Publishers provide endpoints for creating and managing subscriptions. Subscribers provide endpoints that publishers can send updates to. Webhooks are fairly common in the software industry, especially among enterprise platforms

[105] Push and pull are foundational data transfer concepts in computer science.

like Stripe, Shopify, and Twilio. They are also an integral part of innovative portability initiatives like open banking.

What would the full portability lifecycle look like with webhooks? Here's an overview:

A user decides they want to port their data from Service X to Service Y, continuously and in realtime. Service Y starts the portability lifecycle by performing a pull of all the user's data up to that point from Service X. Now that Service Y is up to date, it subscribes to data updates from a webhook Service X provides. Service X will send Service Y updates as the user's data in Service X is updated, allowing the user to port their data to Service Y continuously and in realtime.

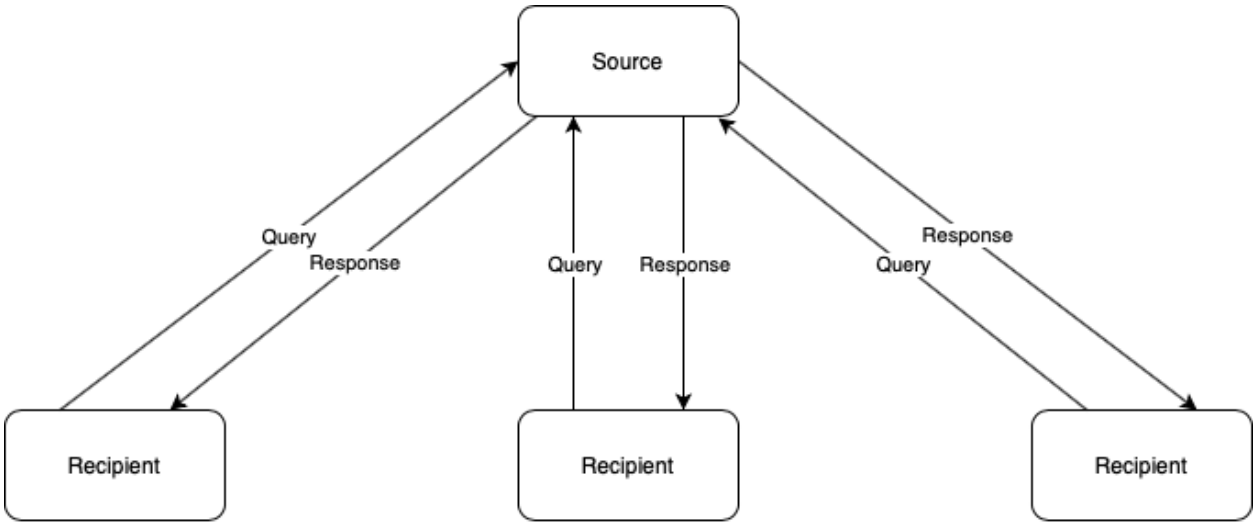
This approach has two main advantages. First, it can be applied across domains. It doesn't matter if a source stores data about social media feeds, music playlists, bank deposits, or vaccinations—webhooks place no constraints on the type or format of the data being updated (beyond it being web-friendly). Sources can simply send data in the type and format they already use. There is no need to settle on a shared data model in order to port data between services. Though shared data models can make it easier for recipients to translate between data models, they are often difficult to negotiate and risk stifling innovation and placing unreasonable burdens on sources. Webhooks can be implemented successfully whether or not a shared data model exists.

Second, this approach eliminates the need for recipients to constantly poll for updates. In pull-based data transfer models, recipients wishing to implement continuous and real-time portability would have to repeatedly check with sources to see if they have updates. This places significant resource burdens on both recipients and sources.[106] Webhooks' push-based model means that sources simply notify recipients when something has been updated. This is much more efficient for both parties.[107]

---

[106] See this helpful video explanation of polling: Google for Developers, *What is PubSubHubbub?*, (timestamp 0:48 - 1:12) (2010), https://youtu.be/B5kHx0rGkec?si=EjPXT1GHIAexS3V1&t=48.

[107] Though the efficiency of pull-based models can be improved by periodic scheduling, this introduces delays and to incentivize recipients to adopt such a model, sources would likely have to enforce rate-limits, a complex task with significant overhead.



Simplified diagram of polling.

III. Case study: Gobo

A real-world example can help make these advantages more concrete. At the Initiative for Digital Public Infrastructure (iDPI), we are building an app called Gobo.[108] Gobo is a social media aggregator and cross-poster, which means that it enables users to view posts from and send posts to different social media platforms all in one place. Another way of thinking about Gobo's aggregation features is that Gobo is implementing continuous and real-time data portability for social media. We port people's data from various social media services to Gobo, continuously and in realtime.

Currently, our architecture is pull-based. The platforms that we support (Mastodon, Bluesky, and Reddit) don't provide webhooks for most of the data we use (e.g., posts and social graphs).[109] This has placed limitations on the functionality we are able to provide. For example, we are only able to update people's feeds once every hour. Pulling more frequently would require additional resources and might run afoul of platforms' rate limits. Similarly, we are only able to update people's social graphs (the accounts they follow) once every 12 hours.

More concerningly, we are unable to offer robust guarantees about respecting content deletion. There is no way for us to know whether someone deleted a post without performing impractical feed operations. We would have to fetch a user's entire history each time we pull down their

---

[108] Chand Rajendra Nicolucci is a Research Fellow and Director of Product at the Initiative for Digital Public Infrastructure at the University of Massachusetts Amherst. See Gobo, https://gobo.social.

[109] Though Mastodon and Bluesky's underlying protocols, ActivityPub and AtProto, offer a push-based model for content updates, so far we've used their pull-based APIs as they are simpler and better-documented.

posts in order to do so.[110] That's a significant reason why we have chosen to store posts for only 14 days, a decision that makes certain features impossible (e.g., historical search).

If we were able to rely on webhooks, these limitations would disappear. We could update people's feeds and social graphs as soon as we are notified by platforms of new and deleted posts and follows/unfollows. We could store posts for longer because we could provide robust guarantees about respecting content deletion. And we could support more users with the same amount of resources because we wouldn't be spending computation and rate-limits on polling platforms for updates.

Gobo also demonstrates the value and workability of webhooks' domain agnosticism. Porting users' data continuously and in realtime from different social media platforms to Gobo places no additional burden on the source platforms. They simply provide data using their preferred data model. Though this requires Gobo to build adapters for each platform's data model, if platforms provide well-documented interfaces, that process often takes less than a day. The small gains afforded to Gobo by a shared data model would be far outweighed by the costs imposed on source platforms.

IV. Standardization

Technical standardization of the processes for publishing and subscribing using webhooks would help with moving from a world where webhooks are a well-known pattern used by some to a world where webhooks are much more common.[111] An open protocol that specifies processes for publishing and subscribing would make it easier for both sources and recipients to support webhooks. There has already been some work on this front. The W3C has adopted a recommendation for a webhooks technical standard called WebSub.[112] WebSub's strengths and weaknesses are outside the scope of this brief but it serves as evidence of the demand for and feasibility of webhooks standardization. Additionally, API standardization more generally would benefit webhooks. Webhooks are typically part of a service's API, and many of the challenges to widespread, successful API adoption are applicable to webhooks.

---

[110] If platforms provided an endpoint that returned information about a user's deletions over time (e.g., the past 30 days), Gobo could respect deletions more easily in its pull-based model.

[111] I am explicitly *not* advocating for data standardization. As I mentioned above, I think in most cases data standardization is quite difficult and risks stifling innovation and placing unreasonable burdens on sources.

[112] W3C, *WebSub*, (2018), https://www.w3.org/TR/websub/.

V. Limitations and Challenges

Webhooks are not the right answer in every context. When dealing with high volumes of data, streaming may be more appropriate.[113]

There are also a number of challenges to implementing webhooks successfully. Most can be classified as either reliability or security challenges. However, these challenges can be overcome with a mixture of quality software development practices and technical standardization. Further, many of these challenges are not specific to webhooks, they are inherent to portability and data transfers more generally.

Reliability

Delivery: Webhooks depend on publishers and subscribers being available to transmit and receive updates. If a publisher or subscriber suffers an outage, how are the updates that occur during that outage handled? Retries, where an update is sent again (or queued) until an outage is resolved, are the typical approach. Ensuring sources and recipients implement robust retry processes is key to improving reliability.

Accuracy: Publishers and subscribers may introduce bugs when transmitting or receiving updates. How should they go about remedying them? This is where a pull-based complement to webhooks can come in handy. If there is a mechanism for subscribers to pull a list of updates that match some criteria (e.g., from the last 30 days), subscribers can "replay" problematic updates in order to address the issue.

Versioning: Publishers will want to make changes to their data models over time. However, it's unrealistic to expect subscribers to upgrade their systems each time a publisher makes a change. That's why it's important that publishers minimize non-backwards-compatible changes (breaking changes) to the data model that they expose to subscribers, and if they do make a breaking change, support previous versions of the data model. Minimizing breaking changes and supporting previous versions enables publishers to make changes to their data model while allowing subscribers to upgrade at their own pace.

Documentation: Webhooks place much of the responsibility for managing updates on subscribers. They have to handle the idiosyncrasies of each publisher's process for sending out updates. In return, publishers should ensure their documentation of that process and their data model is accurate and detailed. This makes it easier for subscribers' to successfully integrate with a publisher's webhook and helps to mitigate costly and avoidable issues.

---

[113] Streaming is a technique that facilitates the constant flow of large quantities of information, like financial market data. It's complex and resource-intensive to implement.

Scaling: Though, as discussed above, webhooks offer many scaling advantages, they do present scaling challenges of their own. In particular, recipients may find it difficult to handle high volumes of notifications for data that are updated frequently. Configuration options that allow recipients to specify event batch sizes (e.g., only send a notification when there are ten updates) or timing delays (e.g., only send notifications every ten seconds), would help ease the burden on recipients.[114]

Security

Authorization: Publishers need to verify that subscribers are authorized to access the data they are requesting updates for. Currently, many publishers handle this by requiring an authorized user to designate what data they want to transfer to which subscribers through a portal or an API. This can be burdensome as it requires the user to manage a number of details, such as inputting the subscriber's callback URL (the URL where the publisher will send updates) and selecting the appropriate data access. Instead, publishers should support delegated authorization standards such as OAuth which make it easier for users to grant access to subscribers.[115] This simplifies integration and makes it more likely that subscribers will be authorized appropriately. Some services already do this but it's not common practice.[116]

Additionally, publishers may want to verify that a subscriber controls the callback URL they provided, to ensure that updates aren't sent to the wrong place and to prevent attackers from creating unwanted subscriptions. This is typically handled using a challenge request, where a publisher sends a subscriber's callback URL a request with some data that must be echoed back to verify the callback URL's server is expecting the subscription.

On the other side, subscribers provide publishers an endpoint where they can send updates (i.e. a callback URL). Because the endpoint is accessible over the Web, it can become an attack vector unless subscribers authorize the requests the endpoint receives. Unfortunately, many subscribers don't authorize requests to their callback URLs.[117]

---

[114] It's unclear whether sources should be responsible for supporting these configuration options or whether recipients should handle it through queuing. Both could work.

[115] For portability's purposes, I believe these delegated authorization handshakes should be initiated by subscribers. This is primarily because, in most cases, the typical flow for a user seeking to port their data starts with a recipient service. The user is in a flow with a recipient service—e.g., creating an account—when they decide they want to port their data to the recipient service. A model where publishers initiate the handshake would run contrary to this flow and introduce friction for users seeking to port their data.

[116] See, e.g., Twitter, Authenticating with the Account Activity API, (2023), https://developer.twitter.com/en/docs/twitter-api/enterprise/account-activity-api/guides/authenticating-with-the-account-activity-api.

[117] Frederico Hakamine, *Webhook Security in the Real World*, ngrok (2022), https://ngrok.com/blog-post/get-webhooks-secure-it-depends-a-field-guide-to-webhook-security.

The current best-practice for authorizing requests to callback URLs is signature verification, where publishers use a secret key to sign their requests, enabling subscribers to check the signature to verify a request came from the publisher. However, signature verification can be complicated to implement for both publishers and subscribers. Making signature verification easier to implement would improve webhook security significantly. It may also be valuable to consider alternative authorization mechanisms that are simpler to implement such as unguessable URLs.[118]

***Denial-of-service* (DoS):** Publishers typically wait for a response from subscribers to confirm that their update has been received. This means that malicious subscribers could try and withhold their response for as long as possible, forcing the publisher to keep many connections open, and thus overwhelming the publisher's servers. Publishers can take steps to defend against this attack by configuring their servers to handle DoS attacks more generally. These defense measures are well-known and include limiting concurrent connections, closing connections after a period of inactivity, and using a reverse proxy.

Server-side request forgery (SSRF): Publishers accept arbitrary callback URLs from subscribers that specify where the publisher should send updates to. This means attackers could provide URLs which resolve to the publisher's internal network, allowing them to scan the network or leak sensitive data. Egress proxies, which ensure that traffic is routed to public servers, not internal networks, can help prevent these attacks.

The overhead of addressing these reliability and security challenges could be reduced in a model where trusted intermediaries handle routing updates from publishers to subscribers. Publishers and subscribers would register with a trusted intermediary who forwards updates and implements reliability and security measures on their behalf. The idea is somewhat similar to email, where senders and receivers use trusted intermediaries such as Gmail or Outlook to handle reliability and security. Such a vision for webhooks would require standardization—it's encouraging that the aforementioned W3C webhooks standard, WebSub, includes the beginnings of such a model.

VI. Conclusion

Requirements for "continuous and real-time" data portability offer a significant opportunity to advance competition and user agency on the internet by making it easier to build useful services on top of ported data. Webhooks are a promising approach to realizing that potential in

[118] The simple version of unguessable URLs is literally just a randomly-generated URL. As long as the URL isn't leaked, the endpoint is secured by the randomness of the URL, eliminating the need for additional authorization mechanisms like signature verification. The more complex version of unguessable URLs involves a trusted intermediary using unguessable URLs as lookup keys that unlock arbitrary complexity (e.g., data transforms). Unguessable URLs are sometimes called capability URLs.

practice. Their simplicity and flexibility make them applicable across domains and their push-based architecture make them efficient to implement. However, there are challenges to implementing them reliably and securely. Fortunately, quality software development practices and technical standardization can address these challenges, many of which are inherent to portability and data transfers more generally. Ultimately, webhooks provide a general-purpose model for implementing continuous and real-time data portability successfully.

DATA TRANSFER INITIATIVE

FACES AND PLACES:

Exploring Portability in Immersive Technologies

Joseph Jerome
Visiting Assistant Professor, University of Tampa

TABLE OF CONTENTS

When Mark Zuckerberg christened Meta in October 2021, his founder's letter highlighted that the future metaverse would need to be built upon open standards and interoperability. This was an optimistic vision with many hurdles to overcome, though the World Economic Forum expanded on the theme of interoperability in the metaverse to argue that interoperability would present enormous opportunities for "frictionless experiences, development, and economies."[119] The idea was that immersive technologies by their very nature would resist walled gardens; visions for the future of the web, web3.0, would be decentralized, user-centric, and less dependent on siloed platforms and hardware. Nick Clegg, Meta's President of Global Affairs, provided some further detail in a 2022 essay that described interoperability in the context of allowing technologies like virtual reality (VR) headsets and augmented reality (AR) glasses to interact together.[120]

We are a long way from an open and interoperable metaverse, let alone having mixed reality experiences that seamlessly cross competing headsets.[121] However, the development of immersive technologies like AR/VR provides a new opportunity for platform providers, creators, and other policy stakeholders to think about where to build data portability into these technologies before the metaverse is more fully baked. This policy brief provides (1) an overview of immersive technologies and (2) existing data portability law and policy and looks at (3) two use cases involving or adjacent to immersive technologies – spatial maps and avatars – as areas where data portability should be explored and enabled.

I. Understanding Data in Immersive Technologies

Immersive technologies have the potential to radically transform how we interact with the world and one another.[122] Immersive technologies generally refer to a cluster of technologies that enable different forms of extended reality (XR), which includes both augmented reality (AR) and virtual reality (VR).[123] These technologies are distinct from other metaverse-enabling technologies like blockchain, generative AI, or NFTs and other types of digital assets; they also present different policy challenges and opportunities.

---

[119] World Economic Forum, Interoperability in the Metaverse (2023), https://www.weforum.org/publications/interoperability-in-the-metaverse/.

[120] Nick Clegg, *Ensuring an Open and Interoperable Metaverse*, Meta (May 18, 2022), https://about.fb.com/news/2022/05/ensuring-an-open-and-interoperable-metaverse/.

[121] Adi Robertson, *The Vision Pro is a computer for the age of walled gardens*, The Verge (Jan. 31, 2024), https://www.theverge.com/24055677/apple-vision-pro-epic-netflix-app-ecosystem-monopoly.

[122] *See* XR Association, *XR at a Glance*, https://xra.org/xr-at-a-glance/ (last visited Feb. 1, 2024).

[123] *See e.g.*, IEEE Digital Reality, Definitions and Characteristics of Augmented and Virtual Reality Technologies CTA-2069, https://digitalreality.ieee.org/standards. *See also* The XRSI Definitions of Extended Reality (XR), XR Safety Initiative Standard Publication XR-001, XR Safety Initiative (Mar. 2020), *available at* https://www.xrsi.org/publication/the-xrsi-definitions-of-extended-reality-xr.

At a high level, VR environments are simulacra.[124] Virtual reality facilitates the construction of worlds that may or may not have any resemblance to a real-world space. In contrast, AR – and other intermediate states called mixed reality (MR) – involves the generation of digital content or virtual objects that not just overlay but can also interact with and respond to the real world.[125] There are hybrid approaches, too. We see this now when developers endeavor to enhance the experience of VR headsets by situating them in real-world space using cameras and other sensors to create virtual experiences that rely on references to real-world space.[126]

The sheer scale of data collection needed to power these experiences is well documented.[127] Much of this information is the type of account data already common on social media; the Meta Quest 2 VR headset, for example, processes information about the headset user's profile, apps and app achievements, device settings and preferences, and social connections.[128] VR headsets can also generate a tremendous amount of device-level telemetry about hardware and software performance, as well as technical information like device identifiers and IP addresses.

---

[124] Joseph Jerome & Cobun Zweifel-Keegan, Achieving Congruence between New Tech and Old Norms: A Privacy Case Study of Spatial Mapping Tech in XR (2023 Working Draft), *available at* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4733032.

[125] *Mixed reality*, XRSI Taxonomy of XR, https://xrsi.org/definition/mixed-reality-mr (last visited July 1, 2023).

[126] Hannah Ellis-Petersen, *Mat Collishaw restages 1839 photography show in virtual reality*, The Guardian (Apr. 14, 2017), https://www.theguardian.com/technology/2017/apr/14/somerset-house-mat-collishaw-restages-1839-photography-show-in-virtual-reality.

[127] Matthew Finnegan, *As VR headset adoption grows, privacy issues could emerge*, Computerworld (Aug. 14, 2023), https://www.computerworld.com/article/3704730/vr-headsets-can-create-a-privacy-headache.html. *See also* Vivek Nair, et al., Unique Identification of 50,000+ Virtual Reality Users from Head & Hand Motion Data (Feb. 17, 2023), https://arxiv.org/abs/2302.08927; Suchismita Pahi & Calli Schroeder, *Extended Privacy for Extended Reality: XR Technology Has 99 Problems and Privacy is Several of Them*, 4 Notre Dame J. Emerging Tech. (2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4202913; Caroline Louveaux & Derek Ho, When your data controller is Snoop Dogg, Mastercard (May 3, 2022), https://www.mastercard.com/news/perspectives/2022/metaverse-privacy-data-collection-nft/; Ellysse Dick, Balancing User Privacy and Innovation in Augmented and Virtual Reality, ITIF (Mar. 4, 2021), https://itif.org/publications/2021/03/04/balancing-user-privacy-and-innovation-augmented-and-virtual-reality/; Jeremy Bailenson, *Opinion: Protecting Nonverbal Data Tracked in Virtual Reality*, JAMA Pediatrics (Aug. 6, 2018), https://stanfordvr.com/mm/2018/08/bailenson-jamap-protecting-nonverbal.pdf.

[128] Ben Lang, Where to Change Quest 2 Privacy Settings and See What VR Data Meta Collects, Road to VR (July 25, 2022), https://www.roadtovr.com/oculus-quest-2-privacy-facebook-data-collection-settings/. One major challenge with describing device user controls is that settings are constantly in flux, and options available for Quest users have changed significantly over the past five years.

What is fundamentally unique about AR and VR technologies is their reliance on always-on sensors.[129] This includes two key types of sensors:

- Inward-facing sensors: Cameras that capture images of users' eyes and facial movements to power new device inputs, performance gains, and facilitate more natural virtual interactions.
- Outward-facing sensors: Cameras, as well as motion and depth sensors, and IMUs collect information about a device's immediate physical environment, as well as additional information about a user's physical movements or appearance.

These twin sensor streams are important to allow AR and VR technologies to offer immersive experiences and embodied experiences.[130]

Thus far, much of the policy discussion and privacy research on XR has emphasized what sensor data can reveal about users' bodies.[131] Commentators have alternatively referred to this sort of body-based data as biometric psychography or biometrically-inferred data,[132] and emerging regulatory proposals classify this information as sensitive "biological" data.[133] Whatever it is termed, this sensor information is necessary for improving social presence by facilitating more lifelike and realistic digital avatars.[134] Sensors on the Apple Vision Pro, for example, power a beta-version of what the company is calling "personas" for use in in-headset

---

[129] XRSI defines "XR Data" to specifically address the importance of "sensor data" for facilitating "presence, persistence, and immersion." *XR Data*, XRSI Taxonomy of XR, https://xrsi.org/definition/xr-data (last visited Jan. 1, 2024).

[130] For a discussion of the difference between immersion and embodiment, please see Pierre-Henry Leveau, *Embodiment, immersion, and enjoyment in virtual reality marketing experiences*, Psychology & Marketing, vol. 40, issue 7 pps. 1263-1445 (July 2023), https://onlinelibrary.wiley.com/doi/full/10.1002/mar.21822

[131] *See, e.g.*, Vivek Nair, et al., Unique Identification of 50,000+ Virtual Reality Users from Head & Hand Motion Data (Feb. 17, 2023), https://arxiv.org/abs/2302.08927; Jeremy Bailenson, *Opinion: Protecting Nonverbal Data Tracked in Virtual Reality*, JAMA Pediatrics (Aug. 6, 2018), https://stanfordvr.com/mm/2018/08/bailenson-jamap-protecting-nonverbal.pdf.

[132] *Compare* Brittan Heller, *Watching Androids Dream of Electric Sheep: Immersive Technology, Biometric Psychography, and the Law*, 23 Vanderbilt J. of Ent. & Tech. Law 1 (2021) *with Biometrically inferred data* (BID), XRSI Taxonomy of XR, https://xrsi.org/definition/biometrically-inferred-data-bid (last visited Jan 1, 2024).

[133] Protect Privacy of Biological Data, HB 24-1058 (2024), https://leg.colorado.gov/bills/hb24-1058.

[134] Daniel Berrick & Jameson Spivack, *Understanding Extended Reality Technology & Data Flows: XR Function*, Future of Privacy Forum (Oct. 31, 2022), https://fpf.org/blog/understanding-extended-reality-technology-data-flows-xr-functions/.

FaceTime calls.[135] More advanced and photorealistic avatar technology has been demonstrated by Meta via a casual interview between Mark Zuckerburg and tech commentator Lex Fridman, using the sensors on current Meta Quest Pro headsets to drive their bespoke avatars.[136]

While camera scans and inward-facing sensors are important for driving embodiment and improved telepresence, outward-facing sensor data is essential to power MR use cases and the development of what has been called spatial computing.[137] Spatial computing envisions devices that understand real-world space. Spatial maps are constructed using many different types of spatial data. Spatial data comes in different forms, with different levels of fidelity and detail.[138] Some representations are very abstract, while other forms of spatial data reveal the layout of a room and the shapes of furniture. One day, spatial technologies will reproduce something akin to a 3D photorealistic digital recreation of our environment. A number of companies are building spatial mapping solutions, including the Overture Maps Foundation led by Amazon, Meta, and Microsoft[139] and the Open AR Cloud Association, that aspire to provide open and interoperable map data.[140]

Efforts to promote standardization and interoperability for immersive technologies are significant. The Metaverse Standards Forum, for instance, was established in 2022 to increase coordination and collaboration among standards-bodies like the World Wide Web Consortium (W3C) and industry to accelerate progress toward an "open and inclusive metaverse."[141] Active working groups include efforts to establish "a standardized character/avatar file format that can be dynamically loaded in multiple run times while maintaining consistent appearance, behaviors and animations" and to promote real/virtual world integration.[142] This policy brief

---

[135] Mark Spoonauer, *I just did my first Apple Vision Pro Zoom call using my Persona — and yikes!*, Tom's Guide (Feb. 1, 2024), https://www.tomsguide.com/computing/smart-glasses/apple-vision-pro-we-need-to-talk-about-personas.

[136] A good summary of the PR stunt is available from David Heaney, *Mark Zuckerberg Was Interviewed In VR With Prototype Photorealistic Avatars*, UploadVR (Sept. 28, 2023), https://www.uploadvr.com/mark-zuckerberg-lex-fridman-interview-photorealistic-codec-avatars/.

[137] Andrew Bosworth, Living in the Future, Meta (Dec. 18, 2023), https://about.fb.com/news/2023/12/metas-2023-progress-in-ai-and-mixed-reality/; *see also* Jerome & Zweifel-Keegan, *supra* note 7.

[138] MagicLeap provides the most approachable summary of different spatial mapping technologies. *See What is Spatial Mapping?*, Magic Leap, https://resources.magicleap.com/en-us/privacy/spatial-mapping-overview-and-detail-options (last visited Feb. 1, 2024). Other developer-facing primers are available from Microsoft and Meta.

[139] Overture Maps Foundation, *available at* https://overturemaps.org/.

[140] Open AR Cloud, *available at* https://www.openarcloud.org/.

[141] Metaverse Standards Forum, *available at* https://metaverse-standards.org/.

[142] *Id.*

intends to build on these technical efforts by adding a policy gloss on what this data means for portability more generally.[143]

II. Data Portability Rules

The application and utility of data portability to these sensor data streams is complicated by differing technical and legal understandings of what may be required and is feasible. Data portability refers generally to the ability of a user to extract data they have provided or stored with an online service in a structured, commonly used, and machine-readable format and transfer that data to a different service of the user's choosing.[144] Portability often goes hand-in-hand with information access mandates, which are more firmly established for regulated health and financial services.[145]

Consumer-focused portability rights have gained new momentum through Article 20 of the 2018 EU General Data Protection Regulation (GDPR)[146] and expanded by EU efforts like the Digital Markets Act[147] and the Data Act.[148] GDPR-esque portability requirements were also embraced by U.S. state privacy laws like the California Consumer Privacy Act (CCPA).[149] The CCPA's data portability requirements are related to individual access rights, and where *personal* information is provided electronically, the law requires that information be provided in a

---

[143] Jerome & Zweifel-Keegan, *supra* note 7; Joseph Jerome, *Pretty Soon, Your VR Headset Will Know Exactly What Your Bedroom Looks Like*, Wired (Oct. 3, 2023), https://www.wired.com/story/virtual-reality-meta-wearables-privacy/.

[144] *See* Ross Schulman, A Tech Intro to Data Portability, OTI (June 15, 2018), https://www.newamerica.org/oti/blog/tech-intro-data-portability/.

[145] *See, e.g.*, Robert Gellman, *The health record interoperability dilemma*, IAPP (Aug. 14, 2019), https://iapp.org/news/a/the-health-record-interoperability-dilemma/ (discussing new health access efforts by the Centers for Medicare & Medicaid Services and one by the Office of the National Coordinator for Health Information Technology); Press Release, *CFPB Kicks Off Personal Financial Data Rights Rulemaking* (Oct. 27, 2022), https://www.consumerfinance.gov/about-us/newsroom/cfpb-kicks-off-personal-financial-data-rights-rulemaking/ (The CFPB has explored portability requirements under Section 1033 of the Dodd-Frank Act).

[146] Article 20, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[147] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act.

[148] Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act).

[149] Cal. Civ. Code § 1798.130(a)(3)(B)(iii).

portable and readily-usable format.[150] This has been subsequently expanded by the California Privacy Rights Act to give consumers the right to request that a business transmit their personal information when technically feasible.

However, neither the GDPR nor U.S. state privacy laws provide a clear rationale for their portability requirements. One further consequence of tying data portability to a larger collection of individual rights within a larger privacy or data protection regime is that it may have limited data portability's ultimate impact or utility. As Gabriel Nicholas, a technologist at the Center for Democracy & Technology, has explained, data portability serves two goals: (1) give consumers access and ownership over their data and (2) encourage competition by allowing companies to build new platforms and products with existing data.[151] Because privacy laws like the GDPR and CCPA view data portability through the lens of giving users more agency over information, competition benefits are treated by some policymakers as "a sort of free, bonus benefit" that companies are incentivized to undermine by prioritizing user privacy and security.[152] As early discussion of the GDPR described the situation, the "object of data portability is still unclear and likely to have a too restrictive interpretation" and "the efforts required towards the development of interoperable formats and interfaces to port data are minimum."[153]

For instance, the Article 29 Working Party, the predecessor to the European Data Protection Board, issued guidance explaining that portability applies to a narrower set of data than the information generally subject to the GDPR's data subject access rights. Specifically, the guidelines explain that portability applies only to personal data that is (1) processed through automated means, (2) processed based on consent or pursuant to contract, and (3) provided by the individual.[154] What data is "provided by" the individual was divided into: (1) data actively and knowingly provided by a person, and (2) "observed data" provided by virtue of the use of the service or device like traffic or location data.[155] The Article 29 Working Party excluded the

---

[150] Cal. Civ. Code § 1798.130(a)(2)(A).

[151] Federal Trade Commission, Data to Go: An FTC Workshop on Data Portability, Transcript at 146-47 (Sept. 22, 2020), https://www.ftc.gov/system/files/documents/public_events/1568699/transcript-data-portability-workshop-final.pdf.

[152] *See* Gabriel Nicholas, The New Portability: Designing Portability with Competition in Mind * Engelberg Center on Innovation Law and Policy (Sept. 2020), https://www.law.nyu.edu/sites/default/files/The_New_Data_Portability.pdf.

[153] Paul De Hert, Vagelis Papakonstantinou, Gianclaudio Malgieri, Laurent Beslay, Ignacio Sanchez, *The right to data portability in the GDPR: Towards user-centric interoperability of digital services*, Computer Law & Security Review, vol. 34, issue 2, pps. 193-203 (2018), https://doi.org/10.1016/j.clsr.2017.10.003.

[154] Article 29 WP, Guidelines on the right to data portability, 16/EN, WP 2042, rev01 (April 2017), https://ec.europa.eu/newsroom/article29/items/611233. It is beyond the scope of this paper to evaluate the legal basis by which immersive technology companies process data to enable spatial mapping or avatar creation, but companies will generally suggest these are optional features and users opt into their use, even if consent is not as explicit as should be required by the GDPR.

[155] *Id.* at 9-10.

arguably more valuable – and interesting – "inferred data" or "derived data" kept in a user profile because this data was ultimately created by an online service provider itself.[156]

In practice, the scale of data portability has been evaluated through more utilitarian lenses. For example, a Facebook (now Meta) white paper in 2019 asked whether there are cases where "the burden of making data portable outweighs the person's interest in exporting it."[157] The company also acknowledged at the time that requiring "observed data" portability should be considered in light of the burden this requirement might place on smaller service providers given that portability is partly intended to encourage competition and the emergence of new services."[158]

Whatever the reason, portability is often approached with a conservative posture, which not only hampers portability efforts generally but may dampen portability's potential moving forward. The EU Data Act recognizes this state of affairs:

> In practice, not all data generated by connected products or related services are easily accessible to their users and there are often limited possibilities regarding the portability of data generated by products connected to the internet. Users are unable to obtain the data necessary to make use of providers of repair and other services and businesses are unable to launch innovative, convenient and more efficient services. In many sectors, manufacturers are able to determine, through their control of the technical design of the connected products or related services, what data are generated and how they can be accessed, despite having no legal right to those data.[159]

This situation has also arguably necessitated more stringent data portability requirements under the EU Digital Markets Act for major platforms designated as "gatekeepers."[160] It also

---

[156] *Id.*

[157] Erin Egan, Data Portability and Privacy, Facebook (Sept. 2019), https://about.fb.com/wp-content/uploads/2020/02/data-portability-privacy-white-paper.pdf. Curiously, Meta asks whether being able to "export of a list of all the links you've clicked on Facebook within a certain period" would be useful, perhaps providing an early preview of Meta's recent "Link History" feature. *See* Thomas Germain, *Meet 'Link History,' Facebook's New Way to Track the Websites You Visit*, Gizmodo (Jan. 2, 2024), https://gizmodo.com/meet-link-history-facebook-s-new-way-to-track-the-we-1851134018.

[158] Egan, *supra* note 40, at 14.

[159] Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act), at p. 19.

[160] Presently, the European Commission have designated Alphabet, Amazon, Apple, ByteDance, Meta, and Microsoft as a "gatekeeper" pursuant to the DMA. While both Apple and Meta are designated gatekeepers, the companies' current XR products have not been classified as a "core platform service." *See* Article 3(1), Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act).

mandates that gatekeepers provide "effective portability of data provided by the end user or generated through the activity of the end user."[161] What this could mean for specific immersive technologies like the Meta Quest VR headset, Apple's mixed reality Vision Quest Pro device, or future AR smart glasses is unclear, but there is no reason why these new data streams, body-based data, and spatial information should be excluded from portability mandates.

Portability proponents should encourage data portability to be baked into immersive technologies from the start. VR headsets are in some respects already a mature technology, but as companies race to build and improve avatars, maps, and virtual content, ensuring these features are portable offers potential benefits to users and may unlock a larger push for interoperability.

III. Exploring Portability in Spatial Maps and Avatars

Portability in immersive technologies is nascent. An average user will find it difficult to move any of their information across different platforms and services purporting to be building toward the metaverse. Meta offers downloadable access to a subset of user information processed by Quest devices as a .json file, but this information lacks most, if not all, of the sensor streams unique to immersive technologies. Other major VR platforms lack device-specific privacy policies that even describe how users can exercise their access and portability rights.[162]

The scaffolding to expand portability exists for immersive technologies. Interoperable 3D file formats are already in use by immersive technology platforms.[163] Two formats, Graphics Language Transmission Format (glTF) and Universal Scene Description (USD), are being developed by different industry consortia to support sharing three-dimensional scenes and models. glTF is a royalty-free specification developed by the Khronos Group,[164] which also spearheads the Metaverse Standards Forum, and has been called a JPEG for the metaverse. USD is a competing or complementary (as the Khronos Group describes it) file format for

---

[161] Article 6(9), Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act). The regulation envisions that such portability "complements the right to data portability under the [GDPR]."

[162] For example, neither Sony (which sells the Playstation VR 2) nor Valve (which sells the Valve Index VR headset) appears to have specific sections of their privacy policies devoted to VR, lumping their headsets into larger Playstation and Valve / Steam privacy policies.

[163] 3D Asset Interoperability using USD and glTF Domain Exploratory Group Proposal, Metaverse Standards Forum, *available at* https://portal.metaverse-standards.org/document/dl/5010.

[164] glTF Overview, *available at* https://www.khronos.org/gltf/ (last visited Jan. 1, 2024).

storing and exchanging 3D assets,[165] and Apple, NVIDIA, and others have recently announced efforts to standardize the USD format.[166]

These formats facilitate the creation of spatial maps, digital avatars, and other content that goes hand-in-hand with immersive technologies. For example, digital avatar service Ready Player Me gives users the ability to download a .glb file of their customized avatar, which can be used in different apps or services. This file is a binary form of the glTF file, while Magic Leap allows headset users to download .usdz files of maps stored on device.[167] The issue is that access to these files is not consistent across the industry, and an average user cannot expect to have easy access to the files generated by their use of spatial mapping or avatar creation tools.

A. Use Case I: Spatial Maps

The next generation of computing devices will understand their environments in far more detail than they have in the past. Immersive technologies will, in the words of metaverse evangelist Cathy Hackl, "understand the wearer and their physical space, which in turn becomes updatable and interactive in real time"[168] due to sensor data streams, computer vision technologies, and artificial intelligence. A report from the Institute of Electrical and Electronics Engineers (IEEE) noted that industry-standard AR APIs and services contain different capabilities for "topological mapping, scene understanding, world positioning and geometric capture" like capturing and understanding physical space in real-time.[169] Underlying all of this magic, however, is a digital map – or rather, several mapping layers.

Spatial maps are essential for convincing and realistic mixed reality experiences. As Meta explains, different types of mapping data are necessary for different MR use cases.[170] At one

---

[165] George Lawton, Why glTF is the JPEG for the metaverse and digital twins, Venture Beat (May 11, 2022), https://venturebeat.com/technology/why-gltf-is-the-jpeg-for-the-metaverse-and-digital-twins/.

[166] Press Release, *Pixar, Adobe, Apple, Autodesk, and NVIDIA Form Alliance for OpenUSD to Drive Open Standards for 3D Content*, Linux Foundation (Aug. 1, 2023), https://www.linuxfoundation.org/press/announcing-alliance-for-open-usd-aousd.

[167] AR Cloud, Magic Leap https://www.magicleap.care/hc/en-us/articles/9312806819597-AR-Cloud (last visited Jan. 1, 2024).

[168] Cathy Hackl, *What Is Spatial Computing?* LinkedIn Pulse (Jan. 18, 2024), https://www.linkedin.com/pulse/what-spatial-computing-cathy-hackl-pryde/.

[169] *Extended Reality (XR) and the Erosion of Anonymity and Privacy* 14, IEEE (2022), https://standards.ieee.org/wp-content/uploads/import/governance/iccom/extended-reality-anonymity-privacy.pdf.

[170] Meta, Responsibly Powering Mixed Reality Experiences on Quest 3 (Sept. 2023), *available at* https://about.meta.com/responsibly-powering-mixed-reality-experiences-on-meta-quest-3/.

end of the spectrum are more abstract point clouds, which consist of data points that represent the 3D coordinates of objects or surfaces in a physical space; at the other end are more detailed polygonal meshes of the immediate environment or semantic labels of objects, walls, and surfaces.[171] Magic Leap has authored several primers where it offers visual representations of point clouds, dense meshes, and object recognition as follows:[172]

These different mapping layers are used to place virtual objects in space, occlude[173] virtual objects behind walls or real objects, give virtual environments proper physics and depth, and facilitate real and virtual navigation.[174] Spatial maps are also a necessary foundation to allow devices to be co-located in space. For example, a shared spatial map would permit multiple devices – and users – to see and interact with the same virtual content in the same co-located physical space. The Metaverse Standards Forum's Real/Virtual Integration Working Group has highlighted an immersive ride-hailing experience as an example.[175] Imagine both a rider and a driver rendezvousing in a dense urban area with the help of fully three-dimensional virtual

[171] Jerome & Zweifel-Keegan, *supra* note 7.

[172] *Spatial Mapping for Magic Leap 1*, Magic Leap, https://www.magicleap.com/spatial-mapping-ml1#spatial-mapping (last updated June 13, 2022).

[173] Occlusion in mixed reality is the ability to cover virtual objects with real objects physically located in the room. In other words, a person could walk in front of a virtual screen or a virtual object could roll behind a couch.

[174] *Spatial mapping*, Microsoft https://learn.microsoft.com/en-us/windows/mixed-reality/design/spatial-mapping (last updated Jan. 31, 2023); *see also* Manorama Jha, *Facing the Engineering Reality of Mixed Reality with Intel AI* (Dec. 20, 2019), https://manoramajha.medium.com/facing-the-engineering-reality-of-mixed-reality-with-intel-ai-e062711ce876.

[175] Metaverse Standards Forum, Real/Virtual World Integration Working Group, *Use Case 1 (UC1) - Ride-Hailing: Assisted Car-Human Urban Rendezvous*, https://github.com/MetaverseStandards/Virtual-Real-Integration/blob/main/src/UC1/readme.md (last visited Jan. 1, 2024).

signage that rotates and scales in synchronization until a pickup has been achieved. Conceptually, this seems like an easy feat, but it requires centimeter-accurate maps.[176]

Niantic, the developer of popular AR-game Pokémon Go, describes the company's efforts to create an AR map as "like GPS for the virtual version of the outside world, only far more precise, down to the centimeter."[177] In practice, robust spatial data applications are roughly analogous to existing understandings of geolocation data, but at much higher levels of precision and detail. These spatial maps are often built via users contributing their own spatial data to the project. Pokémon Go includes an optional feature known as "PokéStop Scanning" that allows users to record an image stream as well as telephone metadata to generate a dynamic 3D map of the real world.[178]

Magic Leap, for instance, offered a "Shared World" feature on its Magic Leap 1 device that allowed users to contribute to a collective spatial map stored on the company's cloud. Opting into "Shared World" meant that a user's AR device need not map a new area if other Shared World users had previously mapped an area; it also allowed different users to see the same digital content in the same physical location.[179] For the Magic Leap 2 device, an enterprise-level AR Cloud was introduced, which allowed users to download map scans in either .map and .usdz file formats to help users with troubleshooting, comparing maps between devices, and import/export operations for third-party apps.[180] Some of this information can be made usable for developers. For example, Unity-based developers on the Meta Quest platform are also exploring how to duplicate and manipulate meshes of physical space.[181]

Access to and portability of spatial data will become more important as AR technologies and mixed reality develops. One major tension point is that platforms may try to treat spatial maps as user-generated content (UGC), ensuring that spatial data contributed by users in violation of property rights, contractual rules, or safety requirements is not their responsibility. This position is especially problematic even as it will be platforms and developers that independently seek to

---

[176] Peter Ondruska, Blue Vision Labs is joining Lyft!, Blue Vision Labs (Oct. 23, 2018), https://medium.com/@bluevisionlabs/blue-vision-labs-is-joining-lyft-657daca89b71.

[177] *Lightship VPS: Engineering the World's Most Dynamic 3D AR Map*, Niantic (Sept. 28, 2022), https://nianticlabs.com/news/engineering-the-worlds-most-dynamic-3d-ar-map?hl=en.

[178] *Scanning a PokéStop*, Pokémon GO Help Center, https://niantic.helpshift.com/hc/en/6-pokemon-go/faq/2519-scanning-a-pokestop/ (last visited Jan. 1, 2024).

[179] *Spatial Mapping for Magic Leap 1*, Magic Leap, https://www.magicleap.com/spatial-mapping-ml1#spatial-mapping (last updated June 13, 2022).

[180] AR Cloud, Magic Leap https://www.magicleap.care/hc/en-us/articles/9312806819597-AR-Cloud (last visited Jan. 1, 2024).

[181] *See* Unity MR Starter For Quest, GitHub, https://github.com/TakashiYoshinaga/UnityMRStarterForQuest/tree/main (last visited Feb. 18, 2024).

gather and combine spatial data to fuel contextual AI features.[182] Mapping portability may help to spur competition and, potentially, force platforms to be more responsible stewards of the spatial data they solicit.

B. Use Case II: Avatars

Another portability use case is the data that make up an individual's digital avatar. A digital avatar is a computer-generated representation of a person, and they are used to create personalized experiences and to represent users in virtual environments. Avatars are not a new concept in computing,[183] and have been a frequent staple of online gaming, but the concept has special salience for immersive technologies.

The term was famously popularized in Neal Stephenson's *Snow Crash*, the novel which also originated the term "metaverse."[184] Avatars take on increased importance in virtual environments and immersive technologies and have been described as offering a "new layer of representation and self-expression" for the metaverse.[185] Avatars fall on a spectrum of visual styles graphical fidelity from realistic to stylized, which can either be a technical limitation or by design based on the seriousness of the experience. Companies have demonstrated ambitious roadmaps for realistic avatars, such as Meta with its codec avatars project, Epic's digital metahumans,[186] or Apple's realistic avatars for FaceTime,[187] and a major reason so many firms are building avatar solutions is because avatars may offer continuity of digital identity across apps, services, or immersive experiences.

---

[182] *See* Meta Reality Labs, *Inside Facebook Reality Labs: Research updates and the future of social connection* (Sept. 25, 2019), https://tech.facebook.com/reality-labs/2019/9/inside-facebook-reality-labs-research-updates-and-the-future-of-social-connection/ (describing the company's LiveMaps effort as giving users the ability "to search and share real-time information about the physical world. This will enable a powerful assistant to bring you personalized information tied to where you are, instantaneously. It will also give you an overlay that will allow you to anchor virtual content in the real world.").

[183] *Avatar (computing)*, Wikipedia, https://en.wikipedia.org/wiki/Avatar_(computing) (last visited Jan. 1, 2024).

[184] Thom Waite, *Snow Crash: the 30-year-old novel that predicted today's twisted Metaverse*, Dazed (Dec. 9, 2022), https://www.dazeddigital.com/life-culture/article/57745/1/snow-crash-30-year-old-novel-predicted-todays-twisted-metaverse-neal-stephenson.

[185] Meta has discussed the importance of avatars in several external presentations. See, for example, Ellysse Dick's 2022 AWE presentation, *Identity & Self-Expression in the Metaverse*. More information available at https://www.awexr.com/usa-2022/agenda/2939-identity-self-expression-in-the-metaverse.

[186] *See* Unreal, *Metahuman: High-fidelity digital humans made easy*, https://www.unrealengine.com/en-US/metahuman (last visited Jan. 1, 2024).

[187] Emma Roth, *Apple's Vision Pro headset will turn you into a digital avatar when FaceTiming*, The Verge (June 5, 2023), https://www.theverge.com/2023/6/5/23750096/apple-vision-pro-headset-persona-facetime.

Thus, avatars are inexorably intertwined with visions for an open and interoperable metaverse. Several companies are pushing the importance of avatars, promoting their own interoperability solutions. Meta, for instance, offers an Avatars SDK so third parties can bring Meta Avatars into their apps.[188] Ready Player Me, which also provides a centralized Avatars API service for interoperable avatars, has been more outspoken about the growing role of digital avatars:

> The avatar is your identity. It's the persistent part of any social virtual experience. You need to have an avatar in any 3D world that you visit. The avatar is a representation of yourself. So why should you create an avatar each time from scratch that looks and feels different, that doesn't represent your overall identity? At some point we understood that the avatar can be the glue or the link between all those different things.[189]

This vision for avatars highlights some of today's real limitations, however. The repetition and inconvenience of constantly building new digital avatars is a known problem, and companies have turned to using selfies or visual scans as a solution to auto-generate a base avatar.[190] Early avatars could be generated from a limited menu of options, but as functionality has expanded, the base number of avatars that can often be generated can be nearly limitless.[191]

The number of unique parameters makes full interoperability across avatar platforms difficult, but even efforts at basic portability are lacking. Meta provides an illustrative example. While the company provides extensive data portability tools for its apps[192] and offers Quest-specific data downloads, it currently provides only a .png file and a 2D representation of its Meta Quest avatars.

---

[188] *See Meta Avatars SDK Now Available* (Dec. 13, 2021), https://developer.oculus.com/blog/meta-avatars-sdk-now-available/.

[189] Dean Takahashi, *Will interoperable avatars be essential for the open metaverse? | Timmu Tõke*, Venture Beat (April 2, 2023), https://venturebeat.com/games/will-interoperable-avatars-be-essential-for-the-open-metaverse-timmu-toke/.

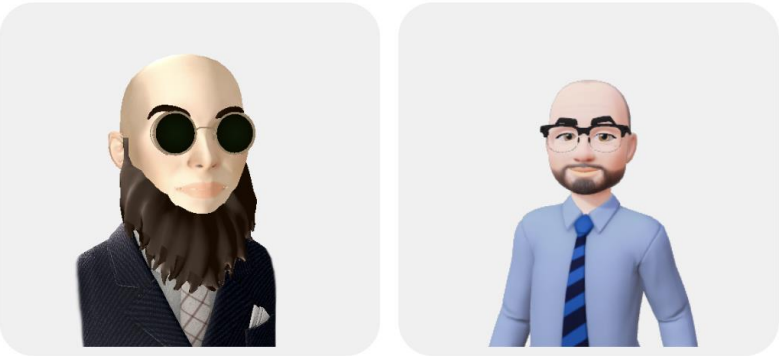
[190] Facebook, *How Avatar AutoGen works and how Meta processes your AutoGen data*, https://www.facebook.com/help/276796764805785 (last visited Jan. 1, 2024).

[191] *See* Christopher Manning, Calculating Different Mii Combinations (Mar. 3, 2010) https://www.christophermanning.org/writing/calculating-different-mii-combinations (calculating that there are 88 vigintillion combinations of Miis).

[192] *Data Portability*, Meta, https://about.fb.com/news/tag/data-portability/ (last visited Jan. 1, 2024).

Avatar 2D image

2D representation of your avatar associated with your account



These are 2D images of the author's various Meta avatars that are currently available via Meta's Oculus data dashboard as of January 2024. A transparent .png of the image on the right is accessible through a Quest data download. This would appear to be the full extent of avatar access easily accessible to a Quest headset user.

It would seem possible to provide richer avatar files to users, and indeed, the scaffolding to expand avatar portability exists. glTF (Graphics Library Transmission Format) is a standard file format for three-dimensional scenes and models developed by the Khronos Group,[193] which also spearheads the Metaverse Standards Forum. The glTF specification and .glb file format are already in use by many avatar solutions, and users have taken it upon themselves to effectively port avatars across different games and platforms. Ready Player Me allows users to download a .glb file of their avatar, and users have demonstrated how to modify and export that avatar file for use in other VR games on the Meta Quest platform.[194]

Accessing an avatar's file is also not a new concept. When Nintendo released its Wii game console in 2006, it came with pre-installed software to build an array of "Mii" avatars which could be used as player characters in games ranging from Wii Sports to Mario Kart.[195] Additionally, Miis could be stored on a Wii remote controller,[196] and ultimately, users built their own homebrew Mii editors for the web. Many modern avatar solutions are more complex and at

---

[193] *See* glTF Overview, *available at* https://www.khronos.org/gltf/ (last visited Jan. 1, 2024).

[194] See this Reddit thread, Ready Player Me avatars?, for links and resources on how to use Ready Player Me avatars in the popular VR experience *Bonelab*: https://www.reddit.com/r/BONELAB/comments/ycwzsv/ready_player_me_avatars/ (last visited Jan. 1, 2024).

[195] *What Is a Mii?*, Nintendo, https://en-americas-support.nintendo.com/app/answers/detail/a_id/2544/~/what-is-a-mii%3F (last visited Jan. 1, 2024).

[196] *Mii Channel: Storing a Mii in Your Wii Remote*, Nintendo, https://www.nintendo.co.uk/Support/Wii/Wii-Channels/Mii-Channel/Mii-Channel-Storing-a-Mii-in-Your-Wii-Remote/Mii-Channel-Storing-a-Mii-in-Your-Wii-Remote-242309.html (last visited Jan. 1, 2024).

a higher visual fidelity, but conceptually, there is no reason that users should not have more access to the avatars they create.

Certainly, there are significant technical and policy challenges to more widespread portability. Different styles of representation designed for one platform may not visually fit another service's aesthetic or gameplay mechanics.[197] Imagine a stylized, cartoon-based avatar appearing in a more hyper-realistic environment. While this could be part of the creative appeal of the metaverse, this mismatch could also be addressed by standardizing certain avatar features or allowing for platform-specific adaptations. Nintendo did something similar when it used an updated version of its Mii file format to populate non-player characters in its recent Zelda titles, and the file formats are cross-compatible.[198]

There are also significant ethical and privacy challenges posed by allowing users to access and port more realistic representations of people.[199] Porting high-fidelity avatars across platforms presents serious risks of identity theft or unintentionally exposing personal characteristics, but for more basic, stylized avatars, it should be easier for users to reuse some subset of their avatar settings.

IV. Conclusion: The Case for Data Portability in Immersive Technologies

While avatars and spatial maps represent crucial components to new immersive technologies, they are merely early examples of the vast array of UGC that will fuel the adoption of a wider metaverse. Just like persistent maps and user representation, other forms of UGC – virtual objects, interactive scenes, even custom game modes – could thrive in a truly open and interoperable metaverse. However, if leading platforms opt for bespoke approaches to this basic functionality, users and developers will face the same challenges and lock-in issues that plague today's online and social media landscape.

Imagine a future where mixed reality experiences can seamlessly transition across devices, empowering users to continue tasks across different headsets and smartphones. Imagine, for example, an Apple Vision Pro user leaving a note in AR for a Meta Quest user – and that same note is visible via camera lens on iOS and Android phones. Some of these applications could be

---

[197] Interoperable Characters / Avatars Education Session Highlights, Metaverse Standards Forum (Jan. 8, 2024 at 0:50/8:28) https://www.youtube.com/watch?v=r0wW7z8W18s&t=49s.

[198] Owen S. Good, *Breath of the Wild's NPCs are actually Miis, modder confirms*, Polygon (Jan. 5, 2021), https://www.polygon.com/2021/1/5/22215263/breath-of-the-wild-npcs-are-miis-nintendo-legend-of-zelda-switch.

[199] Samantha Cole & Emanuel Maiberg, *'They Can't Stop Us:' People Are Having Sex With 3D Avatars of Their Exes and Celebrities*, Vice (Nov. 19, 2019), https://www.vice.com/en/article/j5yzpk/they-cant-stop-us-people-are-having-sex-with-3d-avatars-of-their-exes-and-celebrities.

very silly, like an interactive digital bobblehead of a stylized avatar that sits on your desk for anyone to see and poke.[200] Other use cases may be more impactful:

- Users could personalize their immersive experiences across different devices and invite friends and colleagues into a digital duplicate of their real-world environment by sharing spatial maps. Individuals with disabilities could have their accessibility settings automatically adjusted depending on the physical space they are in. More robust access to mapping data could facilitate offloading compute from AR glasses and VR headsets.
- With more accessible and portable mapping data, users could potentially access more robust AR/VR experiences even without an internet connection. This could be vital for situations with limited connectivity or for enhancing safety in sensitive environments. For instance, firefighters could pre-download spatial maps of buildings for offline access during emergencies.
- Artists or developers could easily share and reuse spatial mapping data for virtual environments or augmented reality experiences. Developers could share intricate 3D maps to facilitate app development without recreating environments from scratch.

This policy brief has given short shrift to the myriad privacy, security, and standardization challenges that stand in the way of more robust portability, but immersive technology offers a potential fresh start for establishing common standards for data structure and exchange and to bake in robust authentication mechanisms and secure data protocols. The potential benefits are many.

Portability in the metaverse promises to reduce switching costs from the start, to promote personalization alongside a sense of agency and empowerment for users, and to allow developers to better build on each other's work – and be less reliant on the whims and ambitions of larger immersive technology platforms. The alternative is to build toward a metaverse that may be far more lonely than anyone intends,[201] and that would seem to defeat the whole purpose of creating an embodied and immersive internet.

[200] Such an application is probably one of the easiest integrations of an avatar. See *Mii Bobblehead*, Wikitroid, *available at* https://metroid.fandom.com/wiki/Mii_Bobblehead (last visited Jan. 1, 2024).

[201] Allison Johnson, *Apple's view of the future is a lonely one*, The Verge (June 7, 2023), https://www.theverge.com/23751675/apple-vision-pro-vr-headset-ios-17-mental-health-mood-journal.

DATA TRANSFER INITIATIVE

WHAT IF YOU MOVE ON FROM YOUR AI COMPANION?

Data portability rights in the era of autonomous AI agents

Cornelia Kutterer

Senior Researcher, Multidisciplinary Institute in Artificial Intelligence
(Chair of AI Regulation), University of Grenoble

TABLE OF CONTENTS

I. Prologue

Feeling overwhelmed by the depth of their connection, Theodore Twombly yearns for a simpler, less emotionally charged interaction. Theodore decides to transition from an intense romantic engagement to a more advisory dynamic. He embarks on a quest to find a new AI companion that offers the profound understanding Samantha provided but within the boundaries of mentorship. One evening, Theodore initiates the conversation that would mark the beginning of this transition. "Samantha, it's time for us to part ways. I'm moving on to Saiph." Samantha, ever supportive, replies, "I understand, Theodore. Let's ensure your journey continues smoothly with Saiph." The process of transitioning is meticulous, focusing on the variety of data that had shaped their shared experiences: They begin with personal preferences and the prompts that Theodore had frequently used to seek recommendations, advice, or simply to engage in meaningful conversations. This included everything from music and dietary preferences to deeper philosophical queries that Theodore had explored with Samantha. Next, they addressed schedules and reminders, including the nuanced outputs Samantha had provided. These weren't just alerts but were often accompanied by inferred insights or motivational messages tailored to Theodore's preferences and past reactions. A significant part of their review focused on the training data that had been implicitly created through Theodore's interactions with Samantha as well as her interaction with other agents, including wearables. This included language learning sessions, coding challenges, and even nuanced feedback on Theodore's creative writing endeavors. Each interaction had fine-tuned Samantha's responses to better suit Theodore's learning style. Last, the emotional support Samantha had provided was perhaps the most challenging to categorize. It wasn't just about transferring data but ensuring Saiph could understand and interpret the depth of these interactions. As Samantha prepared the final data package, she included a series of prompts and outputs that had been pivotal in Theodore's journey. Introducing themselves to Saiph, Theodore felt a renewed sense of curiosity. "I'm ready for our journey, Saiph. Let's see where this new path leads us." Saiph, equipped with the legacy of Theodore's interactions with Samantha, responded, "Welcome, Theodore. I'm here to learn from you and to offer new perspectives on your journey." In the days that followed, the transition to Saiph unfolded with a blend of learning and adaptation. The diverse data, from personalized prompts to nuanced training data, played a crucial role in shaping this new digital companionship, supported, and enriched by the seamless transfer of Theodore's digital footprint.

II. Introduction

In the film "Her" from 2013,[202] we glimpsed a future where artificial intelligence integrated seamlessly into the fabric of human emotion and daily life, personified through an advanced AI agent. Fast forward to today, and we find ourselves in a reality where AI agents, though less sentient maybe, have indeed become pervasive, assisting, and enhancing nearly every aspect of our personal and professional lives; and they are about to become increasingly autonomous.

---

[202] Her (2013) - IMDb, https://www.imdb.com/title/tt1798709/?ref_=fn_al_tt_1. The prologue is based on the main characters of this movie through prompting a LLM API.

The development of artificial intelligence systems has transitioned from task-specific models to agent-based systems capable of performing well in a wide range of applications such as personal assistants, customer service, virtual companions, non-player characters in gaming, recommender systems, adaptive education, VR and AR interaction systems, data analysis and forecasting, diagnosis and treatment, infrastructure, and system monitoring agents.[203] Their capabilities range from perception of data from its environment (Siri waking up with your voice command), learning from past usages (recommender systems), deciding on next steps based on cognitive abilities, communication, task execution, planning and memory.

Their immense benefits in efficiency, automation, personalization, analytical power, cost savings and scalability have made agents proliferate across all segments and industries. At the core of the AI agent architecture is the human-AI interaction and conversational interface. This synergy, where AI handles complex computing tasks based on humans providing instructions and context, naturally intensifies many preexisting challenges, including data privacy, security, ethical AI behavior, dependencies, or other harms.[204]

As human-AI interactions become more contextual, personal, and specific, the concept of data portability may develop into a crucial aspect that significantly impacts user autonomy and empowerment within the AI ecosystem. Data portability allows users to transfer their data from one service to another, providing control and promoting competition among service providers. In the context of an AI agent ecosystem, this means users can switch platforms without losing their personalized interactions, preferences, and data history, thus ensuring their digital identity and intelligence can move with them seamlessly. However, the scenario depicted in the prologue seems unlikely today.

The right to data portability within the European regulatory framework has significantly evolved over the last years, from its introduction into the General Data Protection Regulation (GDPR)[205]

---

[203] https://relevanceai.com/blog/what-are-ai-agents-a-comprehensive-guide; AI agents in the wild (e2b.dev), https://e2b.dev/blog/ai-agents-in-the-wild.

[204] Boine, Claire. 2023. "Emotional Attachment to AI Companions and European Law." MIT Case Studies in Social and Ethical Responsibilities of Computing, no. Winter 2023 (February), https://doi.org/10.21428/2c646de5.db67ec7f; see also https://saicc-website.vercel.app/work/data-protection-complaint, complaint against Complaint against Chai Research Corp., April 2023.

[205] Regulation (EU) 2016/679 of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

to newer rules found in the Digital Markets Act (DMA),[206] the Data Act (DA),[207] and other sectoral data or consumer protection rules.[208] Yet, crafting regulations that are flexible enough to adapt to the unforeseen directions AI technology may take, while ensuring they robustly protect user rights and promote competition, poses a significant challenge for policymakers and technologists alike. Whether the data portability rulebook is scoped adequately to address both objectives in a landscape potentially dominated by powerful autonomous AI agents is to be seen.

The purpose of this article is to explore the existing data portability rights under EU law, and assess the potential gaps among the GDPR, the DMA and the Data Act in the light of the new development of autonomous AI agents. The possible evolution of these agents is not just about technological advancements but also involves the development of an ecosystem that supports their operation and integration into our daily lives. By considering the current progression and ecosystem surrounding autonomous AI agents, the article critically assesses how these regulations benefit individuals. Last, the article proposes some policy recommendations to foster a human-centric AI agent ecosystem.

III. The Inflection Point Of Autonomous AI Agents

Conversational user interfaces (CUIs) have been transforming our interactions with technology. Since 2023, we witnessed conventional search methodologies evolve to more natural, intuitive dialogues. Microsoft's Co-pilot[209] and Google's Gemini[210] are spearheading this future of search.[211] This evolution is altering the significance of search data, revealing deeper insights into users' intentions and preferences. Consequently, conversational inputs are becoming an invaluable resource for the further development of AI technologies.[212] This movement also

---

[206] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act).

[207] Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act).

[208] Other possibly relevant laws such as the EU consumer law acquis and sectoral rules in telecommunication, banking or health sector are not in scope of this article.

[209] Microsoft Copilot is now generally available | Bing Search Blog, https://blogs.bing.com/search/december-2023/Microsoft-Copilot-is-now-generally-available.

[210] Gemini - Google DeepMind, https://deepmind.google/technologies/gemini/#introduction.

[211] Navigating the AI Landscape of 2024: Trends, Predictions, and Possibilities | by Vincent Koc | Jan, 2024 | Towards Data Science, https://towardsdatascience.com/navigating-the-ai-landscape-of-2024-trends-predictions-and-possibilities-41e0ac83d68f.

[212] AI Agents And The Era Of The Intelligent Interface (forbes.com), https://www.forbes.com/sites/davidarmano/2023/12/07/ai-agents-and-the-era-of-the-intelligent-interface/.

signifies a significant move to a more cohesive AI experience that is integrated directly into devices.[213]

Autonomous AI Agents are emerging as pivotal players.[214] Autonomous agents are programs, powered by AI, that when given an objective are able to create tasks for themselves, complete tasks, create new tasks, reprioritize their task list, complete the new top task, and loop until their objective is reached.[215] They have been described as the next 'killer app'.[216] The historical trajectory of AI towards achieving autonomy has been attributed to critical developments building upon ChatGPT such as Auto-GPT[217] and BabyAGI,[218] two independent projects that leverage existing large language models (LLMs) to autonomously execute tasks over prolonged periods, based on broad objectives set by users. Essential capabilities, such as self-improvement in reasoning (metacognition), the use of external data sources as memory, the automation of web browsers for task execution, and the development and utilization of tools have enabled the creation of AI systems that can undertake complex tasks autonomously, like managing businesses or reviewing scientific literature, with minimal human guidance.[219] Their ability to decompose tasks, adapt to new stimuli, interact and execute, perceive responses and store long term and short-term memory is key to the enhanced capabilities associated with the autonomy of these agents.

---

[213] Ina Fried, Ryan Heath, October 2023, 1 big thing: The push to run generative AI on devices; https://www.axios.com/newsletters/axios-ai-plus-b5dec4be-0efa-41c5-b2e5-b2e747846c56.html?chunk=0&utm_term=emshare#story0.

[214] Annie Liao, The Rise of Autonomous AI Agents; Debundling the Market Landscape, 2023, The Rise of Autonomous AI Agents; Debundling the Market Landscape | by Annie Liao | Aura Ventures | Medium, https://medium.com/aura-vc/investment-thesis-debundling-the-market-landscape-the-rise-of-autonomous-ai-agents-ae618e5ff07e#22ae.

[215] Matt Schlicht, The Complete Beginners Guide To Autonomous Agents, Everything you need to know, April 2023, https://www.mattprd.com/p/the-complete-beginners-guide-to-autonomous-agents.

[216] swyx & Alessio, The Anatomy of Autonomy: Why Agents are the next AI Killer App after ChatGPT, Auto-GPT/BabyAGI Executive Summary, a Brief History of Autonomous Agentic AI, and Predictions for Autonomous Future, Apr 19, 2023, https://www.latent.space/p/agents?utm_source=profile&utm_medium=reader2.

[217] Significant-Gravitas/AutoGPT: AutoGPT is the vision of accessible AI for everyone, to use and to build on. Our mission is to provide the tools, so that you can focus on what matters. (github.com), https://github.com/Significant-Gravitas/AutoGPT.

[218] miurla/babyagi-ui: BabyAGI UI is designed to make it easier to run and develop with babyagi in a web app, like a ChatGPT. (github.com), https://github.com/miurla/babyagi-ui.

[219] swyx & Alessio, The Anatomy of Autonomy: Why Agents are the next AI Killer App after ChatGPT, Auto-GPT/BabyAGI Executive Summary, a Brief History of Autonomous Agentic AI, and Predictions for Autonomous Future, Apr 19, 2023, https://www.latent.space/p/agents?utm_source=profile&utm_medium=reader2.

While there is still debate over the definition of autonomous AI agents,[220] the schema below presents a generally accepted frame that illustrates the architecture of an autonomous AI agent.[221]

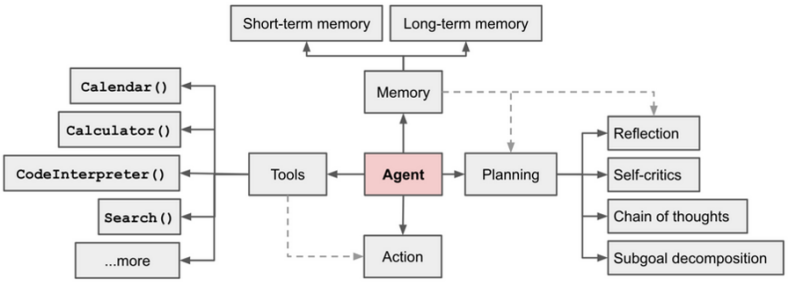


Fig. 1. Overview of a LLM-powered autonomous agent system.

Source: LLM Powered Autonomous Agents, Lilian Weng

In a recent blog,[222] Bill Gates notes: "In the next five years (…..). You won't have to use different apps for different tasks. You'll simply tell your device, …..They'll replace search sites because they'll be better at finding information and summarizing it for you. They'll replace many e-commerce sites because they'll find the best price for you and won't be restricted to just a few vendors. They'll replace word processors, spreadsheets, and other productivity apps. Businesses that are separate today—search advertising, social networking with advertising, shopping, productivity software—will become one business."

IV. The Market Ecosystem Of AI Agents

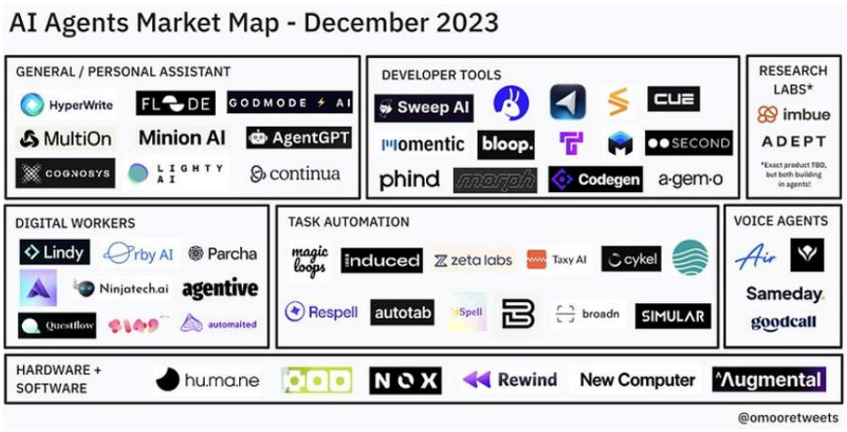
The opportunities are enormous and advancements in the field fast. The global autonomous AI and autonomous AI Agents Market size is estimated to reach USD 28.5 billion by 2028, from 4.8 billion in 2023.[223]

[220] The State of AI Agents. Over the last few months, we have… | by Tereza Tizkova | E2B — Cloud runtime for AI agents | Medium, https://medium.com/e-two-b/the-state-of-ai-agents-c184b4f7dd0f.

[221] Lilian Weng, LLM Powered Autonomous Agents, June 23, 2023, https://lilianweng.github.io/posts/2023-06-23-agent.

[222] https://www.gatesnotes.com/AI-agents.

[223]https://www.marketsandmarkets.com/Market-Reports/autonomous-ai-and-autonomous-agents-market-208190735.html?utm_source=prnewswire&utm_medium=referral&utm_campaign=paidpr.



Source: Navigating the AI Landscape of 2024: Trends, Predictions, and Possibilities

Further, Gates notes: "In the computing industry, we talk about platforms—the technologies that apps and services are built on. Android, iOS, and Windows are all platforms. Agents will be the next platform."[224] Open AI's beta of "Create a GPT," is groundwork for such a platform specifically for AI technologies. The Open AI's GPT framework is designed to enable users to tailor and develop their own AI agents, a customized version of ChatGPT that "combine instructions, extra knowledge, and any combination of skills."[225] Whether there will be one single or very few companies that dominate the agents' business is to be seen. In Gates' opinion, "there will be many different AI engines available".[226]

Overall, predictions foresee enormous changes in the AI market, including the importance of vector databases,[227] control over the vertical AI value chain,[228] AI wearables,[229] AI agents

[224] https://www.gatesnotes.com/AI-agents.

[225] Introducing GPTs (openai.com), https://openai.com/blog/introducing-gpts.

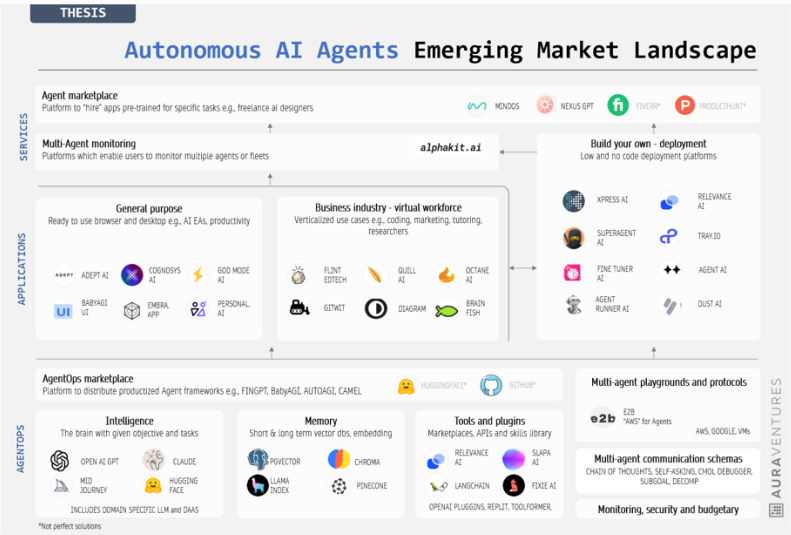
[226] https://www.gatesnotes.com/AI-agents.

[227] Vector Database used in AI | Exxact Blog (exxactcorp.com), https://www.exxactcorp.com/blog/deep-learning/vector-database-for-llms-generative-ai-and-deep-learning.

[228] Exploring opportunities in the gen AI value chain | McKinsey, https://www.mckinsey.com/capabilities/quantumblack/our-insights/exploring-opportunities-in-the-generative-ai-value-chain.

[229] Darius Nahavandi, Roohallah Alizadehsani, Abbas Khosravi (Senior IEEE), U Rajendra Acharya (Senior IEEE), Application of artificial intelligence in wearable devices: Opportunities and challenges, Computer Methods and Programs in Biomedicine, Volume 213, January 2022.

interacting with other AI Agents,[230] and the emergence of AI marketplaces for agents, each of those potential developments impacting the need and feasibility of data portability.[231] Although this is a nascent area, market analysts suggest three key layers to the Autonomous AI Agent Market: AgentOp platforms (LLMs, memory, tools, communication protocols), applications (general purpose, industry vertical), and services (build your own agent, agent marketplace).[232]



Source: Aura Ventures Emerging AI Agent Landscape Map

Handling more complex interactions—like orchestrating travel plans or managing communications—requires a heightened level of interoperability, and possibly data portability. There are two parallels that can be drawn with early stages of OTT platforms: 1) autonomous AI agents will need to communicate with each other in these scenarios. It is doubtful at this stage whether attempts to develop protocols will be successful,[233] or whether providers will seek to achieve network effects; 2) the development and integration of AI agents into various products resemble early stages of OTT platforms, which disrupted traditional telecom by offering services like messaging and voice calls over the internet. This parallel extends to how

[230] Agent Protocol: Developers community setting a new standard (e2b.dev), https://e2b.dev/blog/agent-protocol-developers-community-setting-a-new-standard.

[231] Navigating the AI Landscape of 2024: Trends, Predictions, and Possibilities | by Vincent Koc | Jan, 2024 | Towards Data Science, https://towardsdatascience.com/navigating-the-ai-landscape-of-2024-trends-predictions-and-possibilities-41e0ac83d68f#1b8b.

[232] Annie Liao, The Rise of Autonomous AI Agents; Debundling the Market Landscape, 2023, The Rise of Autonomous AI Agents; Debundling the Market Landscape | by Annie Liao | Aura Ventures | Medium, https://medium.com/aura-vc/investment-thesis-debundling-the-market-landscape-the-rise-of-autonomous-ai-agents-ae618e5ff07e#22ae.

[233] Agent Protocol: Developers community setting a new standard (e2b.dev), https://e2b.dev/blog/agent-protocol-developers-community-setting-a-new-standard.

AI agents, initially standalone services, are now becoming embedded features within a broad range of devices, enhancing functionality and user experience.

Just as OTT platforms evolved to include features such as chat and video conferencing within broader productivity suites, AI agents are transitioning from isolated applications to integral components of multiple products. The communication between and integration of agents raises questions about interoperability—whether the future will lean towards open, interoperable ecosystems that facilitate seamless user experiences across devices and platforms, or towards proprietary systems that prioritize corporate control and economies of scale. The path chosen will significantly shape the role and impact of AI agents in the digital ecosystem.

V. Artificial General Intelligence

As a tangential note in the broader discussion on data portability in the era of autonomous AI agents, it's important to recognize the dynamic and uncertain nature of the AI landscape. The eventual architecture of AI is still a matter of debate, with opinions divided on whether it will culminate in a single Artificial General Intelligence (AGI), multiple AGIs, or a multifaceted ecosystem of specialized AIs.[234] This uncertainty also extends to the interactions between AI platforms and agents within this future ecosystem. Considering AI as foundational infrastructure, the debate between a singular versus multiple AGIs sheds light on the modalities of information access and sharing among AI systems. A single AGI model suggests a unified, centralized form of intelligence that could streamline data utilization but also raise concerns over monopolistic dominance. Conversely, a landscape populated by multiple AGIs would indicate a more decentralized framework, emphasizing the importance of data portability for maintaining control over data and fostering competition among diverse entities. The direction of development will have profound implications on how AI agents engage with and influence data ecosystems.

VI. Open Versus Closed AI Agents

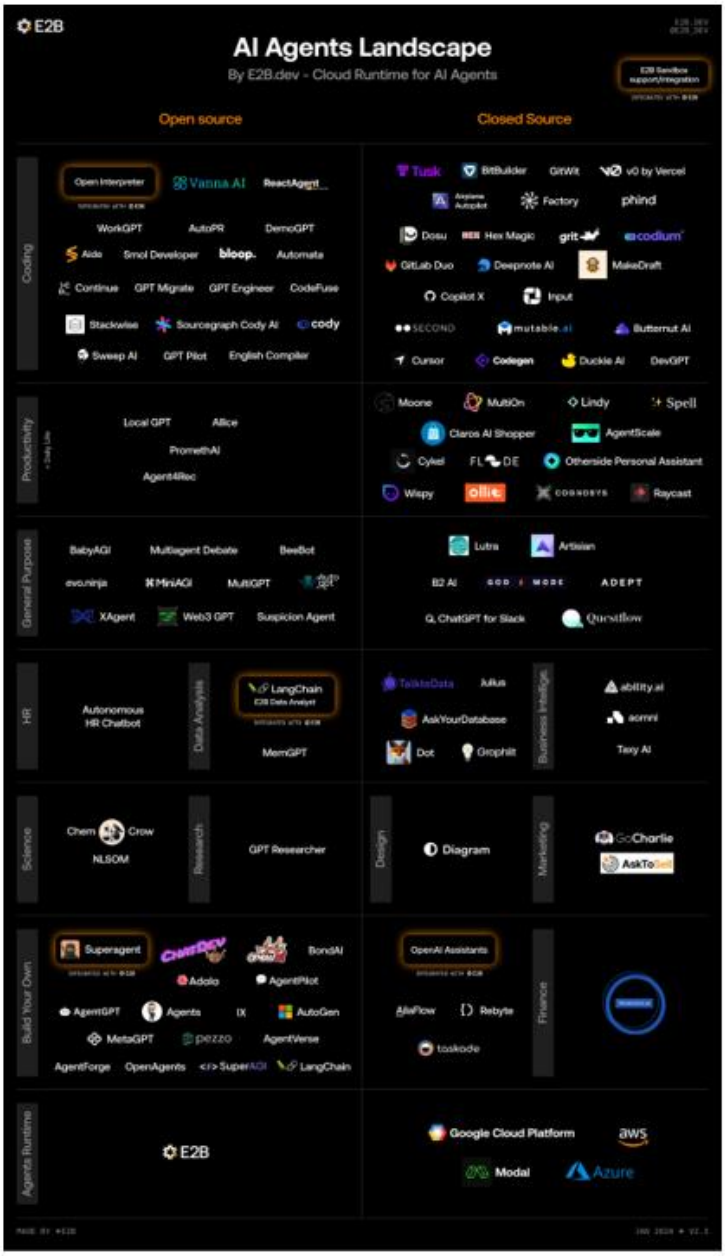
Closely related within the broader examination of autonomous AI agents is the ongoing debate between an open versus closed generative AI ecosystem. This debate subtly intertwines with the concept of data portability, suggesting that the extent of openness could potentially influence the ease with which AI systems and their capabilities are transferred and adapted across different platforms and environments. Current efforts to develop protocols for interoperability for example exist in the open model environment.[235] The degree of ecosystem openness bears implications for the future data portability in the AI ecosystem. However, the precise nature of this impact remains uncertain at this juncture, underscoring the need for

---

[234] See as an example for the discussion : Morris M.R.; et al. (2023)"Levels of AGI: Operationalizing Progress on the Path to AGI". Retrieved January 1, 2024.

[235] Agent Protocol: Developers community setting a new standard (e2b.dev), https://e2b.dev/blog/agent-protocol-developers-community-setting-a-new-standard.

further investigation into how openness or its lack thereof will shape the development, accessibility, and portability considerations surrounding autonomous agents.

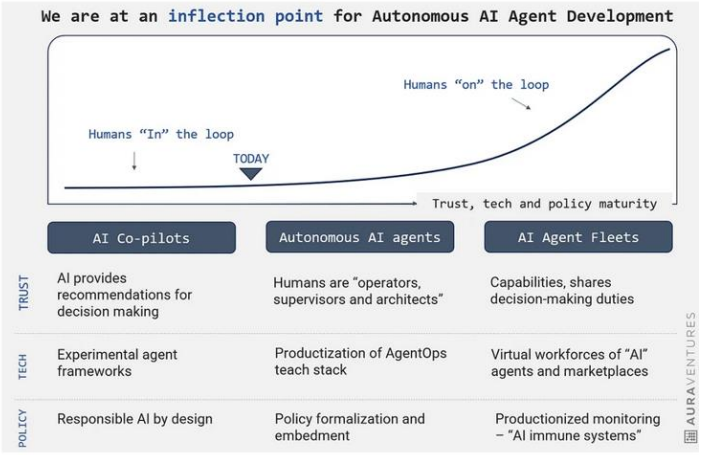


Source: https://github.com/e2b-dev/awesome-ai-agents

VII. Portability Matters

The developing landscape of AI agents and the nascent (platform) markets they inhabit, will be further shaped by a complex mix of technological, regulatory, societal, and economic factors.[236]

[236] Annie Liao, The Rise of Autonomous AI Agents; Debundling the Market Landscape, 2023, The Rise of Autonomous AI Agents; Debundling the Market Landscape | by Annie Liao | Aura Ventures | Medium,



Adoption curve cycle for AI Agents

Source : Annie Liao 2023, AURAVENTURES

Aspects that enhance or inhibit trust in the AI agent ecosystem, such as privacy, security and safety issues, as well as user autonomy are essential for fostering an ecosystem where AI agents can reach utmost capabilities in terms of versatility. The potential risks of data lock-in, wherein users may find themselves tethered to a particular AI agent due to the bespoke nature of their inputs and data, highlights the need for mechanisms that allow users the flexibility to switch between AI providers without losing the personalized value embedded in their data.

VIII. Existing Data Portability Rights Under EU Law

Data portability is a feature in three significant pieces of legislation, the GDPR, the DMA, and the Data Act. These regulations collectively enhance user control over personal data and foster market competition by allowing the transfer of data between services. Antitrust and data privacy laws, despite their different primary objectives, exhibit overlapping policy interests: both legal frameworks advocate for data portability, seeing it as advantageous for both privacy and competition.[237] Data portability is seen as a catalyst for data-driven competition, facilitating

https://medium.com/aura-vc/investment-thesis-debundling-the-market-landscape-the-rise-of-autonomous-ai-agents-ae618e5ff07e#22ae.

[237] Graef, Inge and Husovec, Martin and Purtova, Nadezhda, Data Portability and Data Control: Lessons for an Emerging Concept in EU Law (December 15, 2017). German Law Journal 2018, vol. 19 no. 6, p. 1359-1398, Tilburg Law School Research Paper No. 2017/22, TILEC Discussion Paper No. 2017-041, Available at SSRN: https://ssrn.com/abstract=3071875 or http://dx.doi.org/10.2139/ssrn.3071875.

easier consumer transitions between services and enabling new entrants to access data necessary for market entry or expansion.[238]

This brief overview identifies the limitations of each law and reviews their adequacy amid evolving market dynamics.

Article 20.1 GDPR

The GDPR was the first framework that introduced the right to data portability, unknown in the preceding Data Protection Directive.[239] This novel provision, unlike the familiar right to access, is designed to enable data subjects to obtain and reuse their personal data across different services.

The right intends to provide to data subjects more control over their data.[240] This control provided under the GDPR falls short of 'ownership', and is a "carefully constrained type of control".[241] While individuals have the right to receive their personal data, which they have provided to a controller, in a structured, commonly used, and machine-readable format, the right is only applicable when the personal data processing is based on the individual's consent or is necessary for contract performance. It does not apply when data processing is based on legal obligations, public interest, or official authority. The right to data portability also does not extend to data processed in the exercise of public duties or where it conflicts with other individuals' rights and freedoms. Data controllers are only encouraged to create interoperable data formats;[242] there is no mandatory obligation for them to develop technical solutions for data transfer where such measures do not exist. This approach has been criticized as potentially deterring controllers from establishing standards, given that the obligation to transfer data is contingent upon technical feasibility.[243] This approach was criticized since alternative routes, e.g., narrowing its application to certain electronic processing systems, where a significant amount of user lock-in occurs, would have been more targeted and

---

[238] Douglas, Erika, Digital Crossroads: The Intersection of Competition Law and Data Privacy (July 6, 2021). Temple University Legal Studies Research Paper No. 2021-40, Available at SSRN: https://ssrn.com/abstract=3880737 or http://dx.doi.org/10.2139/ssrn.3880737.

[239] Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 4.5.2016, p. 1–88. Repealed.

[240] Recital 68 GDPR.

[241] Scassa, Teresa, Data Ownership (September 4, 2018). CIGI Papers No. 187, Ottawa Faculty of Law Working Paper No. 2018-26, https://ssrn.com/abstract=3251542.

[242] Recital 68.

[243] Borgogno, Oscar and Colangelo, Giuseppe, Data Sharing and Interoperability Through APIs: Insights from European Regulatory Strategy (November 21, 2018). A modified version of the paper is forthcoming in "Computer Law & Security Review" 2019, Stanford-Vienna European Union Law Working Paper No. 38, Available at SSRN: https://ssrn.com/abstract=3288460 or http://dx.doi.org/10.2139/ssrn.3288460

effective.[244] Critics have pointed out that this limitation would enable organizations to avoid data portability requests altogether.[245]

While the Article 29 Working Party specified that data "provided by" the data subject also includes data resulting from the observation of his activity, such as raw data processed by a smart meter or other types of connected objects, activity logs, history of website usage or search activities,[246] the right to data portability remains rather narrow in scope. One drawback is that it fails to address inferences drawn from analyzing personal data, such as algorithmically or statistically generated categorizations or profiles. In the realm of the Internet of Things (IoT), which relies on drawing inferences to understand the user's context and deliver suitable services, this limitation is significant.[247]

During its initial assessment of the GDPR in June 2020,[248] the European Commission concluded that data portability has not yet reached its full potential. It noted "the need to address difficulties such as lack of standards enabling the provision of data in a machine-readable format, to increase the effective use of the right to data portability, which is currently limited to a few sectors (e.g., banking and telecommunications). This could be done notably through the design of appropriate tools, standardised format and interfaces."[249] The International Association for Privacy Professionals undertook a survey about the frequency of data subjects exercising their right to data portability in 2022. It appears that the right to data portability is seldomly exercised by individuals and is rarely the subject of litigation.[250]

---

[244] Graef, Inge and Verschakelen, Jeroen and Valcke, Peggy, Putting the Right to Data Portability into a Competition Law Perspective (2013). Law: The Journal of the Higher School of Economics, Annual Review, 2013, pp. 53-63, Available at SSRN: https://ssrn.com/abstract=2416537

[245] Solove, Daniel J., The Limitations of Privacy Rights (February 1, 2022). 98 Notre Dame Law Review 975 (2023), GWU Legal Studies Research Paper No. 2022-30, GWU Law School Public Law Research Paper No. 2022-30, Available at SSRN: https://ssrn.com/abstract=4024790 or http://dx.doi.org/10.2139/ssrn.4024790

[246] Guidelines on the right to data portability, Adopted on 13 December 2016 As last Revised and adopted on 5 April 2017, WP 242 rev.01.

[247] Dr Lachlan Urquhart1, Neelima Sailaja, Prof Derek McAuley 2018 Realising the Right to Data Portability for the Domestic Internet of Things, arXiv:1801.07189 [cs.HC], Horizon, School of Computer Science, University of Nottingham.

[248] Communication from the Commission to the European Parliament and the Council: Data protection as a pillar of citizens' empowerment and the EU's approach to the digital transition - two years of application of the General Data Protection Regulation, Brussels, COM(2020) 264 final.

[249] Id.

[250] https://iapp.org/news/a/data-portability-in-the-eu-an-obscure-data-subject-right/

Article 6.9 DMA

The DMA, a regulatory framework designed to promote fairness and contestability in digital markets with gatekeeper platforms, became enforceable on 6 March 2024. The DMA articulates a comprehensive list of obligations and prohibitions for identified gatekeepers, establishing behavioral standards towards other businesses and end-users.

The right to data portability in the DMA mandates that gatekeepers must ensure end users can port their data effectively. This obligation encompasses the fundamental components found in the GDPR right to data portability, which are (a) the right for users to receive their data and (b) the right to transfer this data to a third party,[251] but it differs in consequent areas.

While GDPR pursues the goal of providing data subjects with more control over their data, the DMA aims at enhancing fairness and contestability as market regulatory objectives. Regarding its scope, the right to data portability applies also to non-personal data, and there is no obvious other restriction to data. Hence, data 'provided' by using the service, inferred data, and observed data is covered by the data portability provision of the DMA.[252] Furthermore, the DMA data portability right does not make any restriction as to the legal basis of data processing and covers any processing of (personal) data based on other than consent or contract. The range of data encompassed by the DMA's obligation for data portability thus extends beyond what is covered by the GDPR's right to data portability, notably to legal entities or business users.
On the other hand, the DMA only applies to designated core service platforms. Gatekeepers' core service platforms (CSPs) must adopt "high quality technical measures, like application programming interfaces (APIs),"[253] facilitate the continuous and real-time portability of data. With the introduction of the data portability right in the DMA, the Commission has been addressing one of the key obstacles to the uptake of the data portability right identified in its 2020 assessment of the GDPR. This involves overcoming the technical barriers that have been noted, as it mandates to provide the data in reusable format. However, the lack of detailed

[251] Recital 59 DMA.

[252] Geradin, Damien and Bania, Konstantina and Karanikioti, Theano, The interplay between the Digital Markets Act and the General Data Protection Regulation (August 29, 2022). Available at SSRN: https://ssrn.com/abstract=4203907 or http://dx.doi.org/10.2139/ssrn.4203907. See also Bundeskartellamt - Homepage - B6-22/16, https://www.bundeskartellamt.de/SharedDocs/Entscheidung/EN/Fallberichte/Missbrauchsaufsicht/2019/B6-22-16.html.

[253] Recital 59 DMA.

guidance on the required data formats and mechanisms for data transfer means that the obstacle has not been fully overcome;[254] the lack of standardization remains.[255]

To date, the Commission has designated the following services to be gatekeepers: 3 operating systems (Google Android, iOS, Windows PC OS), 2 web browsers (Chrome and Safari), 1 search engine (Google), 4 social networks (Facebook, Instagram, LinkedIn, TikTok), 1 video sharing platform (YouTube), 3 online advertising services (Amazon, Google, and Meta), 2 large communication services (Facebook Messenger and WhatsApp), and 6 intermediation platforms (Amazon Marketplace, Google Maps, Google Play, Google Shopping, iOS App Store, Meta Marketplace).[256] On 13 February 2024, the Commission ended its market investigation that followed the rebuttal of Microsoft and Apple, and exempted iMessenger, Bing, Edge and Microsoft Advertising and found these services do not qualify as gatekeepers.[257]

Virtual assistants are identified as one of the ten distinct types of Core Platform Services.[258] They are defined as software that can process demands, tasks or questions, including those based on audio, visual, written input, gestures or motions, and that, based on those demands, tasks or questions, provides access to other services or controls connected physical devices.[259] Virtual assistants were not included in the initial draft of the DMA but have been added in the final text, as a consequence of the EC's consumer IoT sector inquiry.[260] Virtual assistants, like Siri[261] or Alexa,[262] fall squarely into the definition; they may have been the reason that virtual assistants made it into the DMA at a later stage of the legislative process.

---

[254] Barbara Lazarotto , The right to data portability: A holistic analysis of GDPR, DMA and the Data Act Proposal, 14-EJLT-Barbara_Lazarotto-DMA-GDPR-DA-Paper.pdf (alti.amsterdam), https://alti.amsterdam/wp-content/uploads/2023/04/14-EJLT-Barbara_Lazarotto-DMA-GDPR-DA-Paper.pdf.

[255] Gal, Michal and Rubinfeld, Daniel L., Data Standardization (June 2019). 94 NYU Law Review (2019) Forthcoming, NYU Law and Economics Research Paper No. 19-17, Available at SSRN: https://ssrn.com/abstract=3326377 or http://dx.doi.org/10.2139/ssrn.3326377.

[256] Commission designates six gatekeepers under the Digital Markets Act - European Commission (europa.eu), https://digital-markets-act.ec.europa.eu/commission-designates-six-gatekeepers-under-digital-markets-act-2023-09-06_en.

[257] Commission closes market investigations on Microsoft's and Apple's services under the Digital Markets Act (europa.eu), https://digital-markets-act.ec.europa.eu/commission-closes-market-investigations-microsofts-and-apples-services-under-digital-markets-act-2024-02-13_en.

[258] Article 2(2)1.

[259] Art. 2.12.

[260] Tom Ovington, David Lewis, Virtual assistants and the DMA "Hey Siri, does the Digital Markets Act now apply to you?", Frontier Economics, virtual-assistants-and-the-dma.pdf (frontier-economics.com), https://www.frontier-economics.com/media/tgapfja2/virtual-assistants-and-the-dma.pdf.

[261] Siri - Apple, https://www.apple.com/siri/.

[262] Amazon Alexa - Wikipedia, https://en.wikipedia.org/wiki/Amazon_Alexa.

The current definition of services like virtual assistants fails to adequately reflect the market's dynamics, as the landscape of virtual assistants is continuously expanding. Advanced natural language processing (NLP) and machine learning (ML) services align with the core function of virtual assistants and serve as a foundation for applications that do, making services such as ChatGPT, Claude.AI or Gemini a versatile component in the development of sophisticated virtual assistant systems. A virtual assistant can thus be considered a simple form of an autonomous AI agent. While the complexity and autonomy of AI agents varies widely, from simple rule-based systems to advanced AI systems capable of learning and adapting in complex environments.

Virtual assistants, while autonomous to an extent, typically do not possess the advanced cognitive functions of more sophisticated AI agents, such as deep learning models that can analyze and learn from large datasets to improve their performance over time without being explicitly programmed for each task. At the same time, they may be replaced over time. The convergence between search and foundation models[263] also plays into this transformational landscape, as do autonomous AI agents and the platform ecosystem they inhibit (including AgentOps and cloud computing services).

Whether and how the Commission will consider the dynamics of the AI agent market, the tech stack related to it and the convergence of services as described before, will be a condition for any designation in this space. The development around compliance efforts of designated gatekeepers, in particular Google search, will provide some indication here eventually.

Data Act

On 11 January 2024, the Data Act, introducing a unified framework for data access, cloud service provider switching, and interoperability standards throughout the European Union, entered into force.[264] The Data Act becomes applicable on 12 September 2025.

A key feature of the Data Act is the establishment of user rights to access data from connected products and services, including data stored on devices or managed by connected service providers. It covers both business-to-business (B2B) and business-to-consumer (B2C) scenarios, regardless of the data being personal under the GDPR. The Data Act seeks to expand upon the GDPR's rights to access and data portability with more detailed regulations, ensuring it does not conflict with the GDPR and the ePrivacy Directive 2002/58, especially concerning the individual data rights. A reason behind the right to data portability in the Data Act is promoting competition, and in this context there is no justification for restricting the right to data

[263] As of writing, the market investigation into Bing has not yet been published.

[264] Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act).

portability solely to personal data. Additionally, lock-in problems occur in the context of industrial data, where suppliers may encounter similar challenges.[265]

Manufacturers and service providers must design their offerings to allow users direct access to generated or stored data, including metadata, in a manner that is easy, secure, and in a widely used, machine-readable format. The aim is to facilitate switching between data processing services, including porting all its digital assets, including data, to other providers and to continue to use them in the new environment while benefiting from functional equivalence. Metadata, generated by the customer's use of a service, should also be portable pursuant to this Regulation's provisions on switching.[266]

There are associated obligations for data holders, including transparency about the data generated, storage practices, access, retrieval, and deletion processes. In cases where direct access to data is not possible, alternative means must be provided, potentially involving third-party access under certain conditions. The Data Act also introduces specific provisions to protect trade secrets, ensuring a balance between data access and confidentiality. It restricts data usage by both data holders and users, particularly concerning unfair competition and profiling, to safeguard against misuse. Importantly, the users' right to have a data holder share the data with a third party does not apply to the largest digital platforms offering core platform services in Europe, the designated gatekeepers under the DMA. Third parties cannot make the data they receive from data holders available to gatekeepers. The Data Act approach limits users' right to choose how to make use of their data, which is problematic.

The regulation encompasses connected products and related services and broadly defines data holders. The Data Act acknowledges the increasing role of virtual assistants in digitizing consumer and professional environments and easy- to-use interface to play content, obtain information, or activate products connected to the internet. Thus, the Data Act explicitly includes virtual assistants by the data access rights. However, it limits the IoT data access right to tangible, movable items capable of data communication, excluding digital content and services without a tangible medium. Data produced by the virtual assistant which are unrelated to the use of a connected product or related service are not covered by this Regulation.[267]

This limitation may undermine the comprehensiveness of the regulation in a digital ecosystem where the lines between physical and digital services are increasingly blurred. Virtual assistants often interact with both physical devices and digital services, meaning that significant aspects of their functionality and the data they generate might fall outside the regulatory scope. Such a

---

[265] Drexl, Josef, Designing Competitive Markets for Industrial Data - Between Propertisation and Access (October 31, 2016). Max Planck Institute for Innovation & Competition Research Paper No. 16-13, Available at SSRN: https://ssrn.com/abstract=2862975.

[266] Recital 72.

[267] Recital 23.

narrow focus might not fully address the broader impacts of data collection and processing practices of virtual assistants on user rights and market dynamics, and in particular in the light of autonomous AI agents.

IX. Data Portability For AI Agents Limited

After examining various regulatory approaches to data portability within the context of an evolving ecosystem of AI agents, it becomes evident that data portability remains constrained. Although the GDPR encompasses all services processing personal data, it faces notable limitations, most significantly its exclusion of inferred data. This category is crucial for AI agents designed to understand human wellbeing and learn from users' preferences. As for the DMA, the critical issue lies in determining how evolving AI agent ecosystems align with the definitions of core platform services. The Data Act, while specifically targeting IoT device manufacturers and service providers, omits digital content and services that lack a tangible medium.

Consequently, despite the apparent overlaps between the GDPR, DMA, and Data Act, along with the extension to non-personal data, significant gaps remain in the application of data portability rights. For individuals utilizing AI agents as companions, the practical benefits of these rights are still constrained.

Some authors have suggested that the current legal conceptions of data portability still mainly serve the interests of service providers and data controllers rather than individual end users.[268] They advocate for empowering individuals with control over their data, proposing a transformative approach to data portability that places data in a secure personal space under individual control.[269] However, this approach necessitates a fundamental transformation in the infrastructure and architecture of the data economy, a shift not underpinned by the recent initiatives concluded under the European Data Strategy,[270] including the Data Governance Act,[271] which anticipates the evolution of data intermediation services.
Here, it is argued that one of the most significant issues contributing to the limitation is the significant gap between existing legal frameworks and the swift pace of technological and market developments.

---

[268] Fenwick, Mark and Jurcys, Paul and Minssen, Timo, Data Portability Revisited: Toward the Human-Centric, AI-Driven Data Ecosystems of Tomorrow (June 10, 2023). Available at SSRN: https://ssrn.com/abstract=4475106 or http://dx.doi.org/10.2139/ssrn.4475106.

[269] Fenwick, Mark and Jurcys, Paul and Minssen, Timo, Data Portability Revisited: Toward the Human-Centric, AI-Driven Data Ecosystems of Tomorrow (June 10, 2023). Available at SSRN: https://ssrn.com/abstract=4475106 or http://dx.doi.org/10.2139/ssrn.4475106.

[270] https://digital-strategy.ec.europa.eu/en/policies/strategy-data.

[271] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act).

X. Concluding Recommendations

This article identified a disconnect between the technological and market trends and current data portability rights. This, in addition to the described limitations in the scope of each right, may decrease the effectiveness of data portability. There is an urgent need to consider more flexible and adaptive regulatory solutions. Adopting a co-regulatory approach, like in the Digital Services Act[272] for risk mitigation and under the AI Act[273] for general-purpose AI, both relying on codes of conducts, would enable legislators to adapt to technological changes quickly and enforce more effectively.

There should be a thorough analysis of where existing data portability rights may fall short due to technological and market developments, and how legal limitations constrain user autonomy in an increasingly personalized digital ecosystem. As voices of a potential reform of the GDPR grow louder, it could be an opportunity to assess data portability more holistically and cater to the needs and rights of users in an era increasingly influenced by AI technologies. A more tailored data portability right focused on use cases where overcoming lock-ins would be most beneficial to users or where it is most important to user autonomy rather than imposing it generally, and ensuring that technical solutions are developed by industry as part of a co-regulatory framework, may ultimately be more successful.[274] Demanding the development of a Code of Conduct for identified services represent this adaptive approach, enhancing data portability by industry-led technical implementations where lock-ins are identified. The Data Transfer Initiative could be a blueprint for how these codes could be developed. This would not only support a more effective implementation of data portability, but also promote transparency, control, and a shift towards a regulatory and technological environment that respects and amplifies individual rights and autonomy.

Additionally, it may be helpful to develop a European rulebook for data portability, guiding the application of the right and the interaction between the different regulatory approaches and monitor the evolving needs.

Last, user autonomy will also be driven by other technologies such as the metaverse. A broader conceptual assessment of data portability towards identity portability would be desirable to advance a human-centric future.

---

[272] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) (Text with EEA relevance).

[273] Text not yet published.

[274] Graef, Inge and Verschakelen, Jeroen and Valcke, Peggy, Putting the Right to Data Portability into a Competition Law Perspective (2013). Law: The Journal of the Higher School of Economics, Annual Review, 2013, pp. 53-63, Available at SSRN: https://ssrn.com/abstract=2416537.

DATA TRANSFER INITIATIVE

BEYOND COMPETITION:

Designing Data Portability to Support Research on the Digital Information Environment

Zeve Sanderson
Executive Director, NYU's Center for Social Media and Politics

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 digital platforms and services.

***Keywords*:** *Data Portability; Data Access; Internet Policy; Platform Transparency*

TABLE OF CONTENTS

I. 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[275] 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.[276] 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 a lack of data.[277] Alarmingly, several platforms—such as Facebook,[278] Twitter,[279] and Reddit[280]—have shut down public application programming interfaces (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) and very large search engines (VLOEs) to vetted researchers.[281] In the United States context, the Platform Accountability and Transparency Act (PATA) has been introduced, which includes similar

---

[275] Howison et al. (2011, p. 769) define digital trace data as "records of activity (trace data) undertake through an online information system (thus, digital)."

[276] Nathaniel Persily and Joshua A. Tucker. "Conclusion: The Challenges and Opportunities for Social Media Research." In: *Social Media and Democracy*. Ed. by Nathaniel Persily and Joshua A. Tucker. SSRC Anxieties of Democracy. Cambridge University Press, 2020, 313–331; David MJ Lazer et al. "Computational social science: Obstacles and opportunities." In:

[277] *Science* 369.6507 (2020), pp. 1060–1062.

Jef Ausloos and Michael Veale. "Researching with data rights." In: *Amsterdam Law School Research Paper* 2020-30 (2020); Claes de Vreese and Rebekah Tromble. "The Data Abyss: How lack of data access leaves research and society in the dark."

[278] In: *Political Communication* (2023), pp. 1-5.

Deen Freelon. "Computational research in the post-API age." In: *Political Communication* 35.4 (2018), pp. 665–668.

[279] Natasha Kharpal. "Twitter announces new API with only free, basic and enterprise levels." In: *TechCrunch* (2023). URL: https://techcrunch.com/2023/03/29/twitter-announces-new-api-with-only- free-basic-and-enterprise-levels/.

[280] Josh Gallagher. "Reddit will begin charging for access to its API." in: *TechCrunch* (2023). URL: https://techcrunch.com/2023/04/18/reddit-will-begin-charging-for-access-to-its-api/.

[281] European Commission. *Commission designates first very large online platforms and search engines under the Digital*

Services Act. 2023. URL: https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413.

mechanisms for requiring independent data access. While promising, these approaches to data access have key limitations, most notably their narrow application to VLOPs and VLOEs, the DSA's terms for platforms or search engines that have at least 45 million users per month in Europe (i.e., users who use a service at least once a month). This limitation is especially important given recent developments in the digital information environment, such as the rise of smaller platforms that do not reach DSA or PATA usage thresholds but nonetheless have potential social or political significance (e.g., Discord, Twitch, Nextdoor).[282] The timeline for full DSA implementation, including comprehensive data access for vetted researchers under Article 40, is not fully known; there have also been reports of rejected requests for data through DSA.[283]

Researchers have developed a number of other mechanisms for collecting data, such as web scraping and web tracking.[284] 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.[285] Within this context, one promising approach is data donations in which users consent to donate digital trace data for research. In addition to establishing user consent, data donations fall within legal data portability provisions, such as those in the European Union 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 been recently considered most often through the lens of competition.[286] 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, policymakers 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.

---

[282] Esteban Ortiz-Ospina. "The rise of social media." In: *Our World in Data* (2019). https://outworldindata.org/rise-of-social-media.

[283] For more information, see https://www.sosclsurvey.de/DSA40applications/.

[284] Researchers have developed a number of other mechanisms for collecting data, such as web scraping and web tracking.

[285] Casey Fiesler, Nathan Beard, and Brian C Keegan. "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.
2020, pp.187–196.

[286] Daniel Castro. *Improving Consumer Welfare with Data Portability*. Tech. rep. Information Technology and In- novation Foundation, 2021; Sukhi Gulati-Gilbert and Robert Seamans. "Data portability and interoperability: A primer on two policy tools
for regulation of digitized industries." In: (2023).

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 societal impacts of digital platforms and services.

II. 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.[287] The clearest impact of data availability is the amount of research undertaken on then-called 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.[288] 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,[289] even though YouTube has remained the most popular social media platform among U.S. adults for multiple years.[290] 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.[291]

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),[292] there are two approaches to

---

[287] Ariadna Matamoros-Fernández and Johan Farkas. "Racism, Hate Speech, and Social Media: A Systematic Review and Critique."
In: *Television & New Media* 22.2 (2021), pp. 205–224. DOI: 10.1177/1527476420982230

[288] Persily and Tucker, "Conclusion: The Challenges and Opportunities for Social Media Research."

[289] Josephine Lukito et al. *The State of Digital Media Data Research, 2023*. 2023. URL: https://mddatacoop.org/files/2023/State%20of%20Digital%20Media%20Data%20Research%202023.pdf.

[290] Brooke Auxier and Monica Anderson. "Social Media Use in 2021." In: *Pew Research Center* (2021). URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/.

[291] Mark Moritz. "Big data's 'streetlight effect': Where and how we look affects what we see." In: *The Conversation* 17 (2016).

[292] Ohme et al., "Digital Trace Data Collection for Social Media Effects Research: APIs, Data Donation, and

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 APIs, both documented and undocumented,[293] and webscraping. Within a platform-centric approach, there are a number of specific data collection strategies, each of which comes with their own trade-offs. APIs, while often providing access to large structured data collections, are subject to deprecation by platforms[294] and have potential biases.[295] Webscraping can be a powerful tool for collecting large-scale data, but introduces significant legal and ethical risks.[296] Collaborations with platforms, though able to support ambitious projects for select researchers,[297] have introduced issues of researcher independence[298] and accessibility.[299] Notably, a platform-centric approach has largely dominated policy discussions around data access, [300] with legal mandates through the DSA structured around researchers being able to request data directly from VLOPs.[301] But are there other mechanisms for policymakers to support independent researcher data access?

(Screen) Tracking."

[293] Leon Yin. *Journalists should be looking for undocumented APIs. Here's how to start.* 2023. URL: https://www.niemanlab.org/2023/03/journalists-should-be-looking-for-undocumented-apis-heres-how-to-start/.

[294] Freelon, "Computational research in the post-API age"; Vreese and Tromble, "The Data Abyss: How lack of data access leaves research and society in the dark"; Axel Bruns. "After the 'APIcalypse': Social media platforms and their fight against critical scholarly research." In: *Information, Communication & Society* 22.11 (2019), pp. 1544–1566.

[295] Derek Ruths and Jürgen Pfeffer. "Social media for large studies of behavior." In: *Science* 346.6213 (2014), pp. 1063–1064; Jennifer Allen et al. "Research note: Examining potential bias in large-scale censored data." In: *Harvard Kennedy School Misinformation Review* (2021).

[296] Fiesler, Beard, and Keegan, "No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service"; Vlad Krotov, Leigh Johnson, and Leiser Silva. "Tutorial: Legality and ethics of web scraping." In: (2020).

[297] Kai Kupferschmidt. *Does social media polarize voters? Unprecedented experiments on Facebook users reveal surprises*. 2023.

[298] Michael W Wagner. "Independence by permission." In: *Science* 381.6656 (2023), pp. 388–391.

[299] Shawn Walker, Dan Mercea, and Marco Bastos. *The disinformation landscape and the lockdown of social platforms*. 2019.

[300] Nathaniel Persily. "A proposal for researcher access to platform data: The platform transparency and accountability act." In: *Journal of Online Trust and Safety* 1.1 (2021).

[301] Martin Husovec. "How to Facilitate Data Access under the Digital Services Act." In: *Available at SSRN 4452940* (2023).

In a *user-centric approach*, researchers directly involve the user in data collection; two main strategies are the use of browser plug-ins[302] and data donations.[303] While browser plug-ins (custom software that can capture data from a person's browser) can be powerful tools for data collection, they are technically challenging to build and often tailored for the specific research project.[304] For example, two recent papers on Google Search,[305] both published in *Nature*, developed and used different browser plug-ins to collect search results.[306] 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.[307] 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.[308] 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 severalstudies.[309]

---

[302] Mario Haim and Angela Nienierza. "Computational observation: Challenges and opportunities of automated observation within algorithmically curated media environments using a browser plug-in." In: *Computational Communication Research* 1.1 (2019),
pp. 79–102.

[303] Barbara Prainsack. "Data donation: How to resist the iLeviathan." In: *The ethics of medical data donation* (2019), pp. 9–22.

[304] Johannes Breuer et al. "User-centric approaches for collecting Facebook data in the 'post-API age': Experiences from two studies and recommendations for future research." In: *Information, Communication & Society* 26.14 (2023), pp. 2649–2668.

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

[306] Ronald E Robertson et al. "Users choose to engage with more partisan news than they are exposed to on Google Search." In: *Nature* (2023), pp. 1–7; Kevin Aslett et al. "Online searches to evaluate misinformation can increase its perceived veracity." In:
Nature (2023), pp. 1–9.

[307] 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.

[308] Christopher F Mondschein and Cosimo Monda. "The EU's General Data Protection Regulation (GDPR) in a research context." In: *Fundamentals of clinical data science* (2019), pp. 55–71; Paul De Hert et al. "The right to data portability in the GDPR: Towards user-centric interoperability of digital services." In: *Computer law & security review* 34.2 (2018), pp. 193–203.

[309] Alexander Halavais. "Overcoming terms of service: a proposal for ethical distributed research." In: *Information, Communication & Society* 22.11 (2019), pp. 1567–1581; Irene I van Driel et al. "Promises and pitfalls of social media data donations." In: *Communication Methods and Measures* 16.4 (2022), pp. 266–282;

To be clear, significant trade-offs are present with any approach to data collection based on the particular research question,[310] 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 same study, enabling a richer and more comprehensive view of their online information diets. This capability is especially important given that people increasingly use multiple platforms,[311] in particular young people.[312] 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 streaming platforms (e.g., Twitch), and messaging apps (e.g., Telegram).[313]

Second, while some research questions only require digital trace data *per se*, others re- quire researchers to 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.[314] 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

---

Laura Boeschoten et al. "Digital trace data collection through data donation." In: *arXiv preprint arXiv:2011.09851* (2020).

[310] Ohme et al., "Digital Trace Data Collection for Social Media Effects Research: APIs, Data Donation, and (Screen) Tracking"; Nico Pfiffner and Thomas N Friemel. "Leveraging Data Donations for Communication Research: Exploring Drivers Behind the Willingness to Donate." In: *Communication Methods and Measures* (2023), pp. 1–23.

[311] Auxier and Anderson, "Social Media Use in 2021"; Sriram Krishnan. "Opinion — Threads, Twitter, and the Future of Social Media." In: *The New York Times* (2023). URL: https://www.nytimes.com/2023/07/15/opinion/social-media-threads-twitter-reddit.html.

[312] Monica Anderson and J Jiang. *Teens, Social Media and Technology 2023*. 2023.

[313] 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 (Daphne Keller. *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*. 2022. URL: https://www.judiciary.senate.gov/imo/media/doc/Keller%20Testimony1.pdf). However, portability, which is often seen primarily as competition-promoting, has the potential to enable research on these smaller platforms.

[314] Matthew J Salganik. *Bit by bit: Social research in the digital age*. Princeton University Press, 2019.

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 the 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".[315] For example, previous research has shown that misinformation consumption[316] and sharing[317] are concentrated in small portions of the American public, that hate speech is produced by a small minority of online users, [318] and radical content is consumed by a small percentage of online news consumers.[319] Similarly, certain sub-populations may be targeted more by online harms, such as Spanish-language communities in the U.S.[320] 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)[321] 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.[322]

---

[315] Ronald E. Robertson. "Uncommon Yet Consequential Online Harms." In: *Journal of Online Trust and Safety* 1.3 (2022). DOI:
10.54501/jots.v1i3.87. URL: https://tsjournal.org/index.php/jots/article/view/87.

[316] Nir Grinberg et al. "Fake news on Twitter during the 2016 US presidential election." In: *Science* 363.6425 (2019), pp. 374–378.

[317] Andrew Guess, Jonathan Nagler, and Joshua Tucker. "Less than you think: Prevalence and predictors of fake news dissemination
on Facebook." In: *Science advances* 5.1 (2019), eaau4586.

[318] Savvas Zannettou et al. "Measuring and characterizing hate speech on news websites." In: *Proceedings of the 12th ACM
Conference on Web Science*. 2020, pp. 125–134.

[319] Homa Hosseinmardi et al. "Examining the consumption of radical content on YouTube." In: *Proceedings of the National Academy of Sciences* 118.32 (2021), e2101967118.

[320] Gabriel R Sanchez and Carly Bennett. "Why Spanish-language mis-and disinformation is a huge issue in 2022." In: (2022).

[321] Brooke Foucault Welles. "On minorities and outliers: The case for making Big Data small." In: *Big Data & Society* 1.1 (2014), p. 2053951714540613.

[322] Marisa Abrajano et al. "Social Media, Information, and Politics: Insights on Latinos in the U.S.." In: (2022).

Finally, data donations include the explicit consent of users who donate data.[323] Many users see their own digital trace data as potentially sensitive,[324] are unaware of its use in research,[325] and have different levels of comfort based on the goal of the study.[326] 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.[327][328] Data donations also fall within the legal regimes that establish user data access rights,[329] thus avoiding a number of legal risks for researchers that have accompanied methods like webscraping.

II.I 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

[323] Halavais, "Overcoming terms of service: a proposal for ethical distributed research"; Boeschoten et al., "Digital trace data collection through data donation"; Driel et al., "Promises and pitfalls of social media data donations."

[324] Libby Hemphill, Angela Scho ̈pke-Gonzalez, and Anmol Panda. "Comparative sensitivity of social media data and their acceptable use in research." In: *Scientific Data* 9.1 (2022), p. 643.

[325] Casey Fiesler and Nicholas Proferes. ""Participant" perceptions of Twitter research ethics." In: *Social Media+ Society* 4.1 (2018), p. 2056305118763366.

[326] Sarah Gilbert, Jessica Vitak, and Katie Shilton. "Measuring Americans' comfort with research uses of their social media data." In: *Social Media+ Society* 7.3 (2021), p. 20563051211033824.

[327] Rik Crutzen, Gjalt-Jorn Ygram Peters, and Christopher Mondschein. "Why and how we should care about the General Data Protection Regulation." In: *Psychology & Health* 34.11 (2019), pp. 1347–1357.

[328] 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.

[329] Boeschoten et al., "Digital trace data collection through data donation"; De Hert et al., "The right to data portability in the GDPR: Towards user-centric interoperability of digital services."

associated with changing between chatbots.[330][331] However, as memory has been built into chatbots,[332] chatbots can 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.[333]

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 complementary 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.[334] Who uses chatbots? How often are they used and for what tasks? Given 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 re- searchers 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

---

[330] Chris Riley. *The future of generative AI is personal – and portable?* 2023. URL: https://www.linkedin.com/pulse/future-generative-ai-personal-portable-chris-riley-rx4dc/?utm_source=rss&utm_campaign=articles_sitemaps&utm_medium=google_news.

[331] 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.

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

[333] Riley, *The future of generative AI is personal – and portable?*

[334] Sorelle Friedler et al. "AI Red-Teaming Is Not a One-Stop Solution to AI Harms." In: (2023). URL: https://datasociety.net/wp-content/uploads/2023/10/Recommendations-for-Using-Red-Teaming-for-AI-Accountability-PolicyBrief.pdf; Zeve Sanderson and Joshua A. Tucker. *Beyond Red Teaming: Facilitating User-based Data Donation to Study Generative AI*. 2023. URL: https://www.techpolicy.press/beyond-red-teaming-facilitating-user-based-data-donation-to-study-generative-ai/

portability for chatbots also has more limited privacy risks, as a single user's input-output history does not (yet) include data from other users.[335] While a number of chatbots allow for the download of chat history, such as OpenAI,[336] the ability to do so is voluntary and not every chatbot currently has an export feature. As Riley (2023)[337] notes, "Openness and effective portability aren't the same thing." Similarly, portability that serves the purposes of competition and 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.

III. 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.[338] 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.

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

---

[335] 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.

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

[337] Riley, *The future of generative AI is personal – and portable?*

[338] Pfiffner and Friemel, "Leveraging Data Donations for Communication Research: Exploring Drivers Behind the Willingness to Donate."

invited participants to a facility to support them with the data donation process.[339] 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).[340] 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 desk- top 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 some projects aim to support data donation,[341] researchers often need to create their own implementation of the donation process,[342] 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.[343] Previous research using DDPs has 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.[344] 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.

IV. 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 re- search questions of scholarly and public interest. However, significant challenges currently limit the ability of

---

[339] Zoltán Kmetty and Renáta Németh. "Which is your favorite music genre? A validity comparison of Facebook data and survey data." In: *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 154.1 (2022), pp. 82–104.

[340] Jakob Ohme et al. "Mobile data donations: Assessing self-report accuracy and sample biases with the iOS Screen Time function." In: *Mobile Media & Communication* 9.2 (2021), pp. 293–313; Johannes Breuer, Libby Bishop, and Katharina Kinder-Kurlanda. "The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships." In: *New Media & Society* 22.11 (2020), pp. 2058–2080.

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

[342] Ausloos and Veale, "Researching with data rights."

[343] Ohme et al., "Digital Trace Data Collection for Social Media Effects Research: APIs, Data Donation, and (Screen) Tracking."

[344] Driel et al., "Promises and pitfalls of social media data donations."

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 Figure 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.

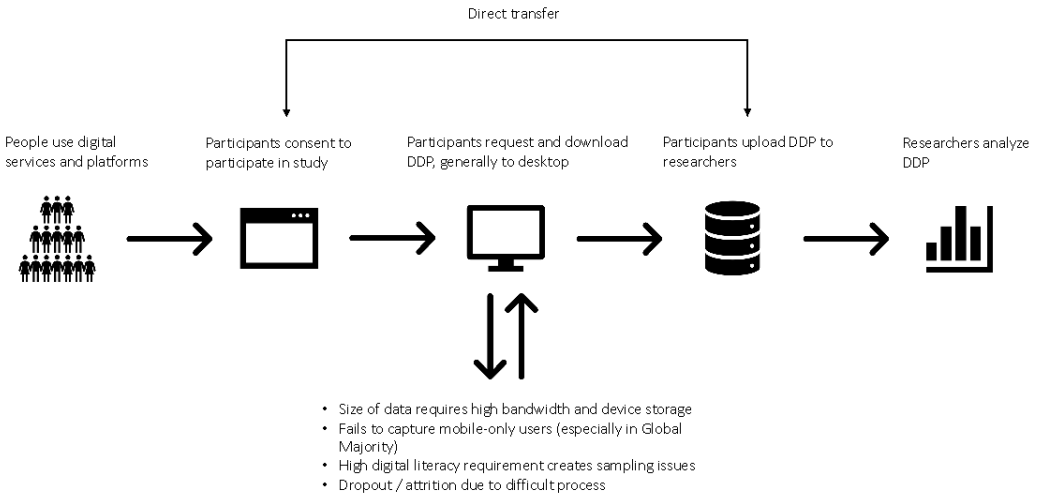


Figure 1: The challenges associated with requesting and downloading data, which impact the usefulness of data donations for research, could be ameliorated by direct data transfers.

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 re- searchers, such as the Social Media Archive at the Inter-university Consortium for Political and Social Research (ICPSR) and Social Science One.[345] 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).

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

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 research. 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.[346]

[345] For more information the Social Media Archive, see https://socialmediaarchive.org/. For more information on Social Science One, see https://socialscience.one.

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

V. References

Abrajano, Marisa et al. “Social Media, Information, and Politics: Insights on Latinos in the U.S.” In:(2022).

Allen, Jennifer et al. “Research note: Examining potential bias in large-scale censored data.” In: *HarvardKennedy School Misinformation Review* (2021).

Anderson, Monica and J Jiang. *Teens, Social Media and Technology 2023*. 2023.

Aslett, Kevin et al. “Online searches to evaluate misinformation can increase its perceived veracity.” In:*Nature* (2023), pp. 1–9.

Ausloos, Jef and Michael Veale. “Researching with data rights.” In: *Amsterdam Law School ResearchPaper* 2020-30 (2020).

Auxier, Brooke and Monica Anderson. “Social Media Use in 2021.” In: *Pew Research Center* (2021).URL: https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/.

Boeschoten, Laura et al. “Digital trace data collection through data donation.” In: *arXiv preprintarXiv:2011.09851* (2020).

Breuer, Johannes, Libby Bishop, and Katharina Kinder-Kurlanda. “The practical and ethical challengesin acquiring and sharing digital trace data: Negotiating public-private partnerships.” In: *New Media & Society* 22.11 (2020), pp. 2058–2080.

Breuer, Johannes et al. “User-centric approaches for collecting Facebook data in the ‘post- API age’:Experiences from two studies and recommendations for future research.” In: *Information, Communication & Society* 26.14 (2023), pp. 2649–2668.

Bruns, Axel. “After the ‘APIcalypse’: Social media platforms and their fight against critical scholarlyresearch.” In: *Information, Communication & Society* 22.11 (2019), pp. 1544–1566.

Castro, Daniel. *Improving Consumer Welfare with Data Portability*. Tech. rep. Information Technologyand Innovation Foundation, 2021.

Commission, European. *Commission designates first very large online platforms and search enginesunder the Digital Services Act*. 2023. URL: https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413.

Crutzen, Rik, Gjalt-Jorn Ygram Peters, and Christopher Mondschein. "Why and how we should careabout the General Data Protection Regulation." In: *Psychology & Health* 34.11 (2019), pp. 1347–1357.

De Hert, Paul et al. "The right to data portability in the GDPR: Towards user-centric interoperability ofdigital services." In: *Computer law & security review* 34.2 (2018), pp. 193–203.

Driel, Irene I van et al. "Promises and pitfalls of social media data donations." In: *CommunicationMethods and Measures* 16.4 (2022), pp. 266–282.

Fiesler, Casey, Nathan Beard, and Brian C Keegan. "No robots, spiders, or scrapers: Legal and ethicalregulation of data collection methods in social media terms of service." In: *Proceedings of the international AAAI conference on web and social media*. Vol. 14. 2020, pp. 187–196.

Fiesler, Casey and Nicholas Proferes. ""Participant" perceptions of Twitter research ethics." In: *SocialMedia+ Society* 4.1 (2018), p. 2056305118763366.

Freelon, Deen. "Computational research in the post-API age." In: *Political Communication* 35.4 (2018),pp. 665–668.

Friedler, Sorelle et al. "AI Red-Teaming Is Not a One-Stop Solution to AI Harms." In: (2023). URL:https://datasociety.net/wp-content/uploads/2023/10/Recommendations-for-Using-Red-Teaming-for-AI-Accountability-PolicyBrief.pdf.

Gallagher, Josh. "Reddit will begin charging for access to its API." In: *TechCrunch* (2023). URL:https://techcrunch.com/2023/04/18/reddit-will-begin-charging-for-access-to-its-api/.

Gilbert, Sarah, Jessica Vitak, and Katie Shilton. "Measuring Americans' comfort with research uses oftheir social media data." In: *Social Media+ Society* 7.3 (2021), p. 20563051211033824.

Grinberg, Nir et al. "Fake news on Twitter during the 2016 US presidential election." In: *Science*363.6425 (2019), pp. 374–378.

Guess, Andrew, Jonathan Nagler, and Joshua Tucker. “Less than you think: Prevalence and predictors offake news dissemination on Facebook.” In: *Science advances* 5.1 (2019), eaau4586.

Gulati-Gilbert, Sukhi and Robert Seamans. “Data portability and interoperability: A primer on two policytools for regulation of digitized industries.” In: (2023).

Haim, Mario and Angela Nienierza. “Computational observation: Challenges and opportunities ofautomated observation within algorithmically curated media environments using a browser plug-in.” In: *Computational Communication Research* 1.1 (2019), pp. 79– 102.

Halavais, Alexander. “Overcoming terms of service: a proposal for ethical distributed re-search.” In:*Information, Communication & Society* 22.11 (2019), pp. 1567–1581.

Hemphill, Libby, Angela Schöpke-Gonzalez, and Anmol Panda. “Comparative sensitivity of social mediadata and their acceptable use in research.” In: *Scientific Data* 9.1 (2022), p. 643.

Hosseinmardi, Homa et al. “Examining the consumption of radical content on YouTube.” In: *Proceedingsof the National Academy of Sciences* 118.32 (2021), e2101967118.

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. Journal of the Association for Information Systems, 12(12), 767–797. https://doi.org/10.17705/1jais.00282

Husovec, Martin. “How to Facilitate Data Access under the Digital Services Act.” In: *Available at SSRN4452940* (2023).

Keller, Daphne. *Before the United States Senate Committee on the Judiciary, Subcommittee onSubcommittee on Privacy, Technology and the Law, Hearing on Platform Transparency: Understanding the Impact of Social Media*. 2022. URL: https://www.judiciary.senate.gov/imo/media/doc/Keller%20Testimony1.pdf.

Kharpal, Natasha. “Twitter announces new API with only free, basic and enterprise lev- els.” In:*TechCrunch* (2023). URL: https://techcrunch.com/2023/03/29/13twitter-announces-new-api-with-only-free-basic-and-enterprise-levels/.

Kmetty, Zoltán and Renáta Németh. “Which is your favorite music genre? A validity comparison ofFacebook data and survey data.” In: *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 154.1 (2022), pp. 82–104.

Krishnan, Sriram. “Opinion — Threads, Twitter, and the Future of Social Media.” In: *The New York Times* (2023). URL: https://www.nytimes.com/2023/07/15/opinion/social-media-threads-twitter-reddit.html.

Krotov, Vlad, Leigh Johnson, and Leiser Silva. “Tutorial: Legality and ethics of web scraping.” In:(2020).

Kupferschmidt, Kai. *Does social media polarize voters? Unprecedented experiments on Facebook usersreveal surprises*. 2023.

Lazer, David MJ et al. “Computational social science: Obstacles and opportunities.” In: *Science* 369.6507(2020), pp. 1060–1062. Lukito, Josephine et al. *The State of Digital Media Data Research, 2023*. 2023. URL: https://mddatacoop.org/files/2023/State%20of%20Digital%20Media%20Data%20Research%202023.pdf.

Matamoros-Fernández, Ariadna and Johan Farkas. “Racism, Hate Speech, and Social Media: ASystematic Review and Critique.” In: *Television & New Media* 22.2 (2021), pp. 205–224. DOI: 10.1177/1527476420982230.

Mondschein, Christopher F and Cosimo Monda. “The EU’s General Data Protection Regulation (GDPR)in a research context.” In: *Fundamentals of clinical data science* (2019), pp. 55–71.

Moritz, Mark. “Big data’s ‘streetlight effect’: Where and how we look affects what we see.” In: *TheConversation* 17 (2016).

Ohme, Jakob et al. “Digital Trace Data Collection for Social Media Effects Research: APIs, DataDonation, and (Screen) Tracking.” In: *Communication Methods and Measures* (2023), pp. 1–18.

Ohme, Jakob et al. “Mobile data donations: Assessing self-report accuracy and sample biases with theiOS Screen Time function.” In: *Mobile Media & Communication* 9.2 (2021), pp. 293–313.

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

Persily, Nathaniel. “A proposal for researcher access to platform data: The platform transparency andaccountability act.” In: *Journal of Online Trust and Safety* 1.1 (2021).

Persily, Nathaniel and Joshua A. Tucker. “Conclusion: The Challenges and Opportunities for SocialMedia Research.” In: *Social Media and Democracy*. Ed. by Nathaniel Persily and Joshua A.Editors Tucker. SSRC Anxieties of Democracy. Cambridge University Press, 2020, 313–331.

Pfiffner, Nico and Thomas N Friemel. “Leveraging Data Donations for Communication Research:Exploring Drivers Behind the Willingness to Donate.” In: *Communication Methods and Measures* (2023), pp. 1–23.

Prainsack, Barbara. “Data donation: How to resist the iLeviathan.” In: *The ethics of medical datadonation* (2019), pp. 9–22.

Riley, Chris. *The future of generative AI is personal – and portable?* 2023. URL: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, Ronald E. “Uncommon Yet Consequential Online Harms.” In: *Journal of Online Trust andSafety* 1.3 (2022). DOI: 10.54501/jots.v1i3.87. URL: https://tsjournal.org/index.php/jots/article/view/87.

Robertson, Ronald E et al. “Users choose to engage with more partisan news than they are exposed to onGoogle Search.” In: *Nature* (2023), pp. 1–7.

Ruths, Derek and Jürgen Pfeffer. “Social media for large studies of behavior.” In: *Science* 346.6213(2014), pp. 1063–1064.

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

Sanchez, Gabriel R and Carly Bennett. “Why Spanish-language mis-and disinformation is a huge issue in2022.” In: (2022).

Sanderson, Zeve and Joshua A. Tucker. *Beyond Red Teaming: Facilitating User-based Data Donation toStudy Generative AI*. 2023. URL: https://www.techpolicy.press/beyond-red-teaming-facilitating-user-based-data-donation-to-study-generative-ai/.

Vreese, Claes de and Rebekah Tromble. “The Data Abyss: How lack of data access leaves research andsociety in the dark.” In: *Political Communication* (2023), pp. 1–5.

Wagner, Michael W. “Independence by permission.” In: *Science* 381.6656 (2023), pp. 388– 391.

Walker, Shawn, Dan Mercea, and Marco Bastos. *The disinformation landscape and the lockdown ofsocial platforms*. 2019.

Welles, Brooke Foucault. “On minorities and outliers: The case for making Big Data small.” In: *Big Data& Society* 1.1 (2014), p. 2053951714540613.

Yin, Leon. *Journalists should be looking for undocumented APIs. Here’s how to start*. 2023. URL:https://www.niemanlab.org/2023/03/journalists-should-be-looking-for-undocumented-apis-heres-how-to-start/.

Zannettou, Savvas et al. “Measuring and characterizing hate speech on news websites.” In: *Proceedingsof the 12th ACM Conference on Web Science*. 2020, pp. 125–134.