ADDRESSING THE FAILURE OF ANONYMIZATION: GUIDANCE FROM THE EUROPEAN UNION'S GENERAL DATA PROTECTION REGULATION

Elizabeth A. Brasher*

It is common practice for companies to "anonymize" the consumer data that they collect. In fact, U.S. data protection laws and Federal Trade Commission guidelines encourage the practice of anonymization by exempting anonymized data from the privacy and data security requirements they impose. Anonymization involves removing personally identifiable information ("PII") from a dataset so that, in theory, the data cannot be traced back to its data subjects. In practice, however, anonymization fails to irrevocably protect consumer privacy due to the potential for deanonymization—the linking of anonymized data to auxiliary information to re-identify data subjects. Because U.S. data protection laws provide safe harbors for anonymized data, re-identified data subjects receive no statutory privacy protections at all—a fact that is particularly troublesome given consumers' dependence on technology and today's climate of ubiquitous data collection.

By adopting an all-or-nothing approach to anonymization, the United States has created no means of incentivizing the practice of anonymization while still providing data subjects statutory protections. This Note argues that the United States should look to the risk-based approach taken by the European Union under the General Data Protection Regulation and introduce multiple tiers of anonymization, which vary in their potential for deanonymization, into its data protection laws. Under this approach, pseudonymized data—i.e., certain data

* J.D. Candidate 2018, Columbia Law School; B.A. 2012, Bucknell University. Many thanks to Professor Ronald Mann for his insight throughout the Note-writing process. Additional thanks to the staff and editorial board of the *Columbia Business Law Review* for their assistance in preparing this Note for publication.

that has had PII removed but can still be linked to auxiliary information to re-identify data subjects—falls within the scope of the governing law, but receives relaxed requirements designed to incentivize pseudonymization and thereby reduce the risk of data subject identification. This approach both strikes a balance between data privacy and data utility, and affords data subjects the benefit of anonymity in addition to statutory protections ranging from choice to transparency.

I. INTRODUCTION

Data protection laws in the United States currently provide safe harbors to companies that "anonymize" their data—a practice that involves removing or obstructing the personally identifiable information ("PII") in a data set so that, in theory, data subjects can no longer be identified.[1] These safe harbors operate under the assumption that anonymization renders data subjects "anonymous," and thereby adequately protects consumer privacy, making it unnecessary to impose additional statutory privacy and data security protections. In practice, however, anonymization fails to permanently obstruct the identities of data subjects due to the potential for deanonymization—the linking of anonymized data to "auxiliary" information to re-identify data subjects.[2]

As consumer data is collected from various sources, consumers leave "data fingerprints" that render PII-based protections, including anonymization exemptions, ultimately futile; for, as one leading privacy expert explained it, "everything is PII to one who has the right outside

[1] *See infra* Part II (explaining various anonymization techniques and U.S. data protection laws that provide safe harbors for anonymized data).

[2] *See generally* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701 (2010) (providing a summary of anonymization techniques, explaining that anonymization has failed, and discussing the implications of this failure for privacy law). Paul Ohm, a leading privacy expert and Professor of Law, is widely recognized for energizing the anonymization debate and the subsequent reassessment of privacy policies.

information."[3] Today's climate of big data and ubiquitous data collection renders deanonymization increasingly feasible as it enables adversaries,[4] such as data brokers, to amass auxiliary information. And problematically, there exist strong financial incentives, such as targeted marketing, to re-identify consumer data.[5] As deanonymization becomes a viable and lucrative practice, the risk of deanonymization—and thus the risk of privacy harm—increases. However, because U.S. data protection laws exempt anonymized data, data subjects who are re-identified receive no statutory privacy protections.

This fact is problematic in light of consumers' growing reliance on technological advancements that make possible—and continually normalize—pervasive data collection.[6] For instance, consumers today heavily use their smart phones, which collect geolocation and other sensitive data, to satisfy society's demand for constant connectivity. And as it becomes more practical to perform everyday tasks online, consumers leave a digital trail of activity, from e-mail and social media communications to search engine queries and payment transactions, with private companies that have an interest in collecting, storing, and selling their data. Given consumers' growing reliance on such technologies, it is unrealistic to think that consumers have a real choice in whether to "opt-out" of using these products, and thereby "opt-out" of data sharing.

---

[3] *Id.* at 1723. To illustrate the enormity of this concept, consider the study conducted by Dr. Latanya Sweeney that used 1990 census data to show that 87.1% of people in the United States could likely be uniquely identified by their five-digit ZIP code, birth date, and sex alone. Latanya Sweeney, *Uniqueness of Simple Demographics in the U.S. Population* 16 (Lab. for Int'l Data Privacy, Working Paper No. LIDAP-WP4, 2000), http://dataprivacylab.org/projects/identifiability/paper1.pdf [perma.cc/L2X8-VJW8].

[4] The term "adversary" is used to refer to the individual or entity attempting to deanonymize data records. For a discussion of big data, data brokers, and ubiquitous data collection, see *infra* Section III.B.

[5] *See infra* Section III.B.3 (discussing the financial incentives that adversaries have to deanonymize consumer data).

[6] This Note merely seeks to provide a sample of such technological advancements in an effort to illustrate the current climate of ubiquitous data collection, and in no way endeavors to address the full spectrum of devices enabling data collection that exist today.

Moreover, technological advancements such as the Internet of Things ("IoT")—including home automation systems, autonomous cars, wearable computers, and smart medical devices—enable an unprecedented stream of data collection and facilitate the collection of new types of sensitive consumer data.[7] As these products offer enticing benefits, consumers choose to become users despite privacy concerns.[8] And in light of these realities, the lack of statutory privacy protections for re-identified data subjects is inappropriate.

The United States has created no means of incentivizing the practice of anonymization, which reduces the linkability of data to its subjects, while still providing those same data subjects statutory privacy and data security protections. This Note contributes to the growing literature on the anonymization debate by arguing that the United States should follow the European Union's approach to anonymization under the General Data Protection Regulation (the "GDPR") by incorporating multiple tiers of anonymization requirements, which vary their obligations with the risk of deanonymization, into its data protection laws. Under the European Union's approach, pseudonymized data, which has had PII removed but can still be linked to auxiliary information to re-identify data subjects, falls *within* the scope of the relevant data protection law—affording data subjects protections such as transparency, choice, and security—while receiving certain relaxed requirements designed to incentivize the practice of pseudonymization and thereby reduce the linkability of data to its data subjects.

---

[7] Consider, for instance, when users of Fitbit, a popular wearable fitness tracker, discovered that the device incidentally collected their sexual activity records. *See* Kashmir Hill, *Fitbit Moves Quickly After Users' Sex Stats Exposed,* FORBES (July 5, 2011, 7:58 AM), http://www.forbes.com/sites/kashmirhill/2011/07/05/fitbit-moves-quickly-after-users-sex-stats-exposed [perma.cc/KKC9-BA3P]. *See also infra* Section III.B.3 (discussing the Internet of Things (the "IoT")).

[8] *See infra* Section III.B.1 (discussing the "convenience/privacy" tradeoff). *See also, e.g.*, Robert Stroud, *The Convenience/Privacy Trade-Off on the Internet of Things*, WIRED (Dec. 17, 2013, 11:05 AM), http://insights.wired.com/profiles/blogs/the-convenience-privacy!-trade-off-on-the-internet-of-things [perma.cc/5K73-DCE9].

Part II of this Note will provide an overview of anonymization and the safe harbors that exist for anonymized data under U.S. data protection laws and Federal Trade Commission ("FTC") guidelines. Part III will explain the failure of anonymization and problematize this failure in light of society's reliance on technology, the prevalence of big data and data brokers, and technological advancements that enable unprecedented sensitive data collection. Part IV will engage in a comparative analysis and discuss the European Union's approach to anonymization under the GDPR. Finally, Part IV will conclude that the United States should embrace a framework similar to the European Union's by introducing the concept of pseudonymization into its data protection laws—an approach that artfully balances data privacy with data utility, and affords data subjects the benefit of quasi-anonymity as well as a range of statutory privacy protections.

II. ANONYMIZATION AND U.S. DATA PROTECTION LAWS

A. Anonymization

1. How Anonymization Works

Anonymization forms "the core of standard procedures for storing or disclosing personal information."[9] Anonymization involves modifying a dataset to remove or encrypt PII. By obstructing this PII, anonymization protects data subjects' privacy by reducing the linkability of the data to its subjects.

PII is a legal concept, and the data elements that constitute PII are defined in the data protection law governing the data at issue.[10] Traditionally, definitions of PII contemplated "direct identifiers"—or facially identifiable data such as name, social security number, or date of birth. However, in recognition of the potential for deanonymization, definitions of PII now also contemplate "quasi-identifiers"—or non-

9 Ohm, *supra* note 2, at 1707.

10 *See infra* Section II.B.2 (discussing the data that constitute PII under the Health Insurance Portability and Accountability Act ("HIPAA")).

facially identifiable data that can be linked to auxiliary information to re-identify data subjects.[11] For example, the FTC has broadly defined PII to include data "reasonably link[able] to a specific customer, computer, or other device."[12]

There are a variety of anonymization techniques that satisfy anonymization exemptions under existing data protection laws.[13] These techniques balance privacy with utility to varying extents. This Note will provide a brief overview of five of the most common anonymization techniques: (1) suppression, (2) generalization, (3) aggregation, (4) noise addition, and (5) substitution.[14]

---

[11] *See, e.g.*, Fed. Trade Comm'n, Comment Letter on Proposed Rule to Protect the Privacy of Customers of Broadband and Other Telecommunications Services 9 (May 27, 2016), https://www.ftc.gov/system/files/documents/advocacy_documents/comment-staff-bureau-consumer-protection-federal-trade-commission-federal-communications-commission/160527fcccomment.pdf [perma.cc/D4VX-5PWW] ("[T]he definition of PII should not be confined to information that is already linked to an individual. . . . Not only is it possible to link information historically considered non-PII to specific individuals or devices, but businesses have strong incentives to do so. . . . [The use of the term] 'linkable' extends stronger privacy protections to consumers.").

[12] FED. TRADE COMM'N, PROTECTING CONSUMER PRIVACY IN AN ERA OF RAPID CHANGE: RECOMMENDATIONS FOR BUSINESSES AND POLICY MAKERS vii (2012), http://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf [perma.cc/BA29-SYXK].

[13] *See id.* at 21 ("A variety of technical approaches to de-identification may be reasonable, such as deletion or modification of data fields, the addition of sufficient 'noise' to data, statistical sampling, or the use of aggregate or synthetic data.").

[14] For an overview of anonymization techniques and their strengths and weaknesses, see ARTICLE 29 DATA PROTECTION WORKING PARTY, OPINION 05/2014 ON ANONYMISATION TECHNIQUES (Apr. 10, 2014), http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf [perma.cc/GQ6C-6MNW]; GREGORY S. NELSON, PRACTICAL IMPLICATIONS OF SHARING DATA: A PRIMER ON DATA PRIVACY, ANONYMIZATION, AND DE-IDENTIFICATION, 13–15 (ThotWave Technologies 2015), http://support.sas.com/resources/papers/proceedings15/1884-2015.pdf [perma.cc/2WFK-PT6K]; Cédric Burton & Sára Hoffman, *Personal Data, Anonymization, and Pseudonymization in the EU*, WSGR DATA ADVISOR (Sept. 15, 2015),

Suppression involves removing PII from a dataset entirely. While this technique provides the most protection to data subjects, it significantly reduces the utility of the anonymized data records. Generalization, in contrast, simply involves modifying identifier values to display, for instance, the year of a person's birth instead of the full date. While this technique better preserves the data's utility, it provides a much weaker privacy protection to data subjects. Professor Paul Ohm, the privacy expert who energized the anonymization debate, has described these two techniques as "release-and-forget" techniques: Once a data administrator modifies the data and releases the records, she "forgets, meaning she makes no attempt to track what happens to the records after release."[15]

Other anonymization techniques "work by relaxing either the release or the forget requirement."[16] For instance, aggregation, noise addition, and substitution substantially reduce the linkability of anonymized data to its data subjects by obstructing the raw data from view.[17] Aggregation provides summary statistics by grouping data subjects that share some personal data element,[18] while noise addition inserts imprecision into the original dataset.[19] Finally, substitution involves directly replacing data values in the original dataset with other parameters—for instance, by replacing a given height (e.g., five feet) with a given color (e.g., blue).[20]

http://www.wsgrdataadvisor.com/2015/09/personal-data-anonymization-and-pseudonymization-in-the-eu/ [perma.cc/88M7-7JEL].

15 Ohm, *supra* note 2, at 1712.

16 *Id.* at 1755.

17 Even Paul Ohm has noted, "In most cases, reidentifiers will find it much more difficult to link answers like these to identity than if they had access to the underlying raw data." *Id.*

18 *K*-anonymity, a popular form of aggregation, ensures that a set of potential target records cannot be reduced to fewer than *k* records in the dataset. *See* Ira Rubinstein & Woodrow Hartzog, *Anonymization and Risk*, 91 WASH. L. REV. 703, 758–59 (2016); Burton & Hoffman, *supra* note 14.

19 For instance, shuffling the values of attributes in a table so that some attributes are artificially linked to different data subjects is a form of noise addition known as permutation. *See, e.g.*, ARTICLE 29 DATA PROTECTION WORKING PARTY, *supra* note 14, at 12.

20 Burton & Hoffman, *supra* note 14.

2. The Purpose of Anonymization

Privacy laws impede the free flow of information, which is instrumental to many essential political and economic functions, such as transparency and research. As Professors Ira S. Rubinstein and Woodrow Hartzog have remarked, "[B]lanket and robust prohibitions on information collection and disclosure would be incredibly costly to organizations and society as a whole. Shutting down research and the information economy would be devastating. Even if such restrictions were wise and politically palatable, they would likely be ineffective given the existing data ecosystem."[21]

Anonymization is appealing because it balances the free flow of information with the risk of privacy harm; it enables the release of valuable but sensitive information while reducing the linkability of that data to its subjects. For this reason, anonymization has been praised as a "best-of-both-worlds compromise."[22] However, its ability to achieve this balance has been called into question since the weakness of anonymization as a privacy protection came to light.

The failure of anonymization to permanently obstruct data subjects' identifies, discussed *infra*, has energized a vibrant debate around whether anonymization is a sufficient means of protecting data subjects' privacy.[23] And because there is a negative correlation between data privacy and data utility, defining the proper balance lies at the heart of this debate. On one side of the debate, scholars defend anonymization by emphasizing the opportunity costs of reduced data sharing that stem from alternative methods of privacy protections.[24] On the other side of the debate, scholars suggest abandoning

[21] Rubinstein & Hartzog, *supra* note 18, at 731.

[22] Ohm, *supra* note 2, at 1703, 1736 ("Legislatures have deployed a perfect, silver bullet solution—anonymization—that has absolved them of the need to engage in overt balancing. Anonymization liberated lawmakers by letting them gloss over the measuring and weighing of countervailing values like security, innovation, and the free flow of information.").

[23] *See infra* Section III.A (discussing the failure of anonymization, the anonymization debate, and policy solutions proposed by academics).

[24] *See* Jane Yakowitz, *Tragedy of the Data Commons*, 25 HARV. J.L. & TECH. 1, 4 (2011) (discussing the importance of broad data accessibility).

reliance on anonymization, arguing that "the benefits of being free from data controls do not outweigh the cost of relinquishing control and protection."[25] These considerations are important in weighing alternative privacy solutions and the extent to which they should employ anonymization as a method of striking this balance and enabling information flow.

B. Safe Harbors for Anonymized Data Under U.S. Data Protection Laws

1. Overview of U.S. Data Protection Laws

The United States does not have a general, comprehensive data protection law. Rather, the United States has a patchwork system of laws that regulates data on a channel- or industry-specific basis at both the federal and state levels, with certain industries deemed sensitive enough to warrant heightened regulation by one or both levels of government.[26]

At the federal level, general industry regulators enforce industry-specific privacy regulations. For instance, the Health Insurance Portability and Accountability Act ("HIPAA") is enforced by the Department of Health and Human Services ("HHS").[27] In the absence of a specific regulation governing the data at issue, the FTC is the primary federal regulatory authority pursuant to section 5 of the Federal Trade Commission Act (the "FTCA"), a consumer protection law that

[25] Rubinstein & Hartzog, *supra* note 18, at 739 ("[W]e argue that sound process-based policy minimizes or eliminates 'release-and-forget' deidentification as an acceptable strategy. . . . We argue that the data controls are just as important as deidentification in safely releasing data sets."). *See also* Ohm, *supra* note 2, at 1732, 1768 (categorizing anonymization techniques as a "shared hallucination" and explaining that "the idea that we can single out fields of information that are more linkable to identify than others has lost its scientific basis and must be abandoned").

[26] Lisa J. Sotto & Aaron P. Simpson, *United States*, *in* LEGAL BUSINESS RESEARCH LTD., GETTING THE DEAL THROUGH—DATA PROTECTION & PRIVACY 2015 (Rosemary P Jay ed., 2014), https://www.huntonprivacyblog.com/wp-content/uploads/sites/18/2011/04/DDP2015_United_States.pdf [perma.cc/4EUM-DK5H]. *See also* Paul Ohm, *Sensitive Information,* 88 S. CAL. L. REV. 1125 (2015) (providing an overview of U.S. data privacy laws).

[27] Pub. L. No. 104-191, 110 Stat. 1936 (1996).

prohibits "unfair or deceptive acts or practices in or affecting commerce."[28] At the state level, state attorneys general have the authority to bring actions for unfair or deceptive trade practices or enforce violations of state data protection laws.[29]

There are two kinds of federal data protection laws in the United States: "sensitive information" laws and "protected channel" laws.[30] Protected channel laws, such as the U.S. Wiretap Act and the Stored Communications Act, regulate specific channels of communication.[31] Sensitive information laws regulate data on an industry-specific basis, with some industries receiving more stringent protections than others. Data that are currently regulated on an industry-specific basis include medical data under HIPAA;[32] consumer report and background screening data under the Fair Credit Reporting Act (the "FCRA") and the Fair and Accurate Credit Transactions Act ("FACTA");[33] children's data under the Children's Online Privacy Protection Act ("COPPA");[34] financial data under Gramm-Leach-Bliley Act (the "GBLA");[35] and educational records under the Family Educational Rights

[28] Federal Trade Commission Act, Pub. L. No. 63-203 (codified as amended at 15 U.S.C. §§ 41–58). Note that the Federal Trade Commission (the "FTC") also has the authority to enforce a number of industry-specific laws. *See Privacy & Security Update (2016)*, FTC (Jan. 2017), https://www.ftc.gov/reports/privacy-data-security-update-2016 [perma.cc/365V-U6VK] (describing FTC enforcement authority).

[29] Divonne Smoyer & Aaron Lancaster, *State AGs: The Most Important Regulators in the U.S.?*, IAPP (Nov. 26, 2013), https://iapp.org/news/a/state-ags-the-most-important-regulators-in-the-us/ [perma.cc/RY9U-JY57].

[30] Ohm, *supra* note 26, at 1132–36.

[31] The U.S. Wiretap Act regulates the data collected by providers of communication services. *See* 18 U.S.C., ch. 119, §§ 2510-2522 (1968). The Stored Communications Act regulates communications stored with certain types of online intermediaries. *See* 18 U.S.C., ch. 121, §§ 2701–2712 (1986).

[32] Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-109, 110 Stat. 1936 (1996).

[33] Fair Credit Reporting Act, Pub. L. No. 91-508, 84 Stat. 1127 (1970) (codified at 15 U.S.C. § 1681); Fair and Accurate Credit Transactions Act, Pub. L. No. 108-159, 117 Stat. 1952.

[34] Children's Online Privacy Protection Act of 1998, Pub. L. No. 105-277, 112 Stat. 2681-728 (1998)

[35] Gramm-Leach-Bliley Act, Pub. L. No. 106-102, 113 Stat. 1338 (1999).

and Privacy Act ("FERPA").[36] There is no accepted definition for when data meets the requisite sensitivity to warrant its own law, and existing laws were enacted on an ad hoc basis.[37]

Many categories of data that are deemed sensitive enough to warrant heightened protections in other legal systems, such as in the European Union, do not receive heightened protections in the United States. For instance, the GDPR designates "special categories of personal data" that are subject to heightened protections, including "data concerning a natural person's sex life or sexual orientation" in this category.[38] In the United States, such data, and all other data not governed by a channel-specific or industry-specific federal law, are left under the protection of the FTC pursuant to section 5 of the FTCA, as well as state attorneys general.[39]

2. Exemptions Under HIPAA

HIPAA was enacted with the broad purpose of improving health insurance coverage and health care delivery, and is among the most significant privacy laws in the United States. The portion of HIPAA that regulates the privacy of personal

[36] Family Educational Rights and Privacy Act, Pub. L. No. 93-380, 88 Stat. 484 (codified as amended at 20 U.S.C. § 1232g).

[37] *See* Ohm, *Sensitive Information*, *supra* note 26, at 1130, 1140 ("[N]ew categories of sensitive information are rarely added to the positive law of privacy, and categories already enshrined in law are never removed.").

[38] Council Regulation 2016/679, art. 9(1) 2016 O.J. (L119) The full list of "special categories" includes "personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation."

[39] This Note is primarily focused on statutory law in the United States. However, the dissemination of certain data, including sexually explicit data, may also violate common law privacy torts. The common law includes a requirement that the privacy tort of disclosure involve disclosure of facts that would be "highly offensive to a reasonable person," which have been found to include information about sexual activity. Ohm, *supra* note 26, at 1134. *See, e.g.*, Michaels v. Internet Entm't Grp., 5 F. Supp. 2d 823, 842 (C.D. Cal. 1998) (enjoining defendants from disseminating a video depicting plaintiffs' sexual activity, finding that plaintiffs made the requisite showing of success on the merits for their claim of violation of the right to privacy).

health information ("PHI") is known as the HIPAA Privacy Rule.[40] Among other things, the HIPAA Privacy Rule requires appropriate safeguards to protect the privacy of PHI, sets limits and conditions on the use and disclosure of PHI with and without patient authorization, and gives patients rights over their PHI, including the right to obtain a copy of their health records and to request corrections.[41] To illustrate the regulatory approach to anonymized data in the United States, this Section will discuss the exemption for de-identified health information ("DHI") under the HIPAA Privacy Rule.[42]

A major goal of the HIPAA Privacy Rule is to "strike[] a balance that permits important uses of information, while protecting the privacy of people who seek care and healing."[43] HIPAA expressly exempts DHI from regulation under the HIPAA Privacy Rule if the information is anonymized per HIPAA's de-identification standard and implementation specifications.[44] HIPAA defines DHI as "[h]ealth information that does not identify an individual and with respect to which there is no reasonable basis to believe that the information

---

40 The HIPAA Privacy Rule is located at 45 C.F.R. Part 160 and Subparts A and E of Part 164. *See also The HIPAA Privacy Rule*, HHS (last updated Apr. 16, 2015), https://www.hhs.gov/hipaa/for-professionals/privacy/index.html [perma.cc/S9CT-HQBG].

41 *See The HIPAA Privacy Rule*, *supra* note 40.

42 *See* NELSON, *supra* note 14, at 8 ("[HIPAA is] one of the primary standards used to provide guidance for de-identifying [PII] and [PHI]."); Ohm, *supra* note 2, at 1737 ("[HIPAA is] the high-water mark for the use of PII to balance privacy risks against valuable uses of information.").

43 *Summary of the HIPAA Privacy Rule*, HHS (last updated July 26, 2013), https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html [perma.cc/RDV6-F5KU] (stating that a key objective was to "assure that individuals' health information is properly protected while allowing the flow of health information needed to provide and promote high quality health care and to protect the public's health and well being").

44 45 C.F.R. § 164.502(d)(2) (2017) ("Health information that meets the standard and implementation specifications for de-identification under § 164.514(a) and (b) is considered not to be individually identifiable health information, *i.e.*, de-identified. The requirements of [Subpart E – Privacy of Individually Identifiable Health Information] do not apply to information that has been de-identified.").

can be used to identify an individual."[45] To constitute DHI under HIPAA, information that is anonymized must satisfy the requirements of one of two standards: (1) the Expert Determination standard[46] or (2) the Safe Harbor standard.[47]

To satisfy the Expert Determination standard, a person with "appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable" must conclude that the risk is "very small" that the information "could be used, alone or in combination with other reasonably available information, by an anticipated recipient" to identify a data subject.[48] It is noteworthy that this standard contains several qualifiers that fail to preclude, in absolute terms, the possibility of reidentification: (1) the risk of identification need only be "very small"; (2) the auxiliary information must be "reasonably" available; and (3) the risk of reidentification is limited to that by the "anticipated recipient" as opposed to that by any adversary who gains access to the information.

DHI is alternatively exempt from HIPAA if it satisfies the Safe Harbor standard. The Safe Harbor standard requires (1) the removal of eighteen enumerated identifiers of the individual or of relatives, employers, or household members of the individual,[49] and (2) that "the Covered Entity does not have actual knowledge that the information could be used alone, or in combination with other information, to identify"

45 45 C.F.R. § 164.514(a).

46 *See* 45 C.F.R. § 164.514(b)(1).

47 *See* 45 C.F.R. § 164.514(b)(2).

48 45 C.F.R. § 164.514(b)(1).

49 45 C.F.R. § 164.514(b)(2)(i). These data elements are: (A) names, (B) geographic subdivisions smaller than a State, (C) all elements of dates (except year) directly related to the individual including birth date and date of death, (D) telephone numbers, (E) fax numbers, (F) e-mail addresses, (G) social security numbers, (H) medical record numbers, (I) health plan beneficiary numbers, (J) account numbers, (K) certificate/license numbers, (L) vehicle identifiers such as license plate numbers, (M) device identifiers and serial numbers, (N) URLs, (O) IP addresses, (P) biometric identifiers including finger and video prints, (Q) full face photographic images, and (R) "any other unique identifying number, characteristic or code" beyond those permitted in 45 C.F.R. § 164.514(c).

the data subject.[50] This standard too contains qualifiers that fail to preclude the possibility of reidentification: (1) only the eighteen identifiers enumerated must be removed;[51] and (2) the Covered Entity must have actual knowledge that the information could be used to identify the data subject.

In addition to creating an exemption for DHI and prescribing mandatory mechanisms for anonymization, the HIPAA Privacy Rule creates implementation specifications that allow covered entities to re-identify DHI. Specifically, HIPAA allows a covered entity to "assign a code or other means of record identification to allow information de-identified . . . to be re-identified by the covered entity" as long as (1) the means of record identification is not "derived from or related to" information about the data subject or otherwise capable of being "translated" so as to identify the data subject, and (2) the covered entity does not use or disclose the means of record identification "for any other purpose" or disclose the mechanism for re-identification.[52] HIPAA clarifies that if de-identified information is re-identified, the covered entity's use and disclosure of the information must comply with the requirements of the HIPAA Privacy Rule.[53]

3. Exemptions Under FTC Guidelines

The FTC is the primary federal regulatory authority for all data not governed by a channel- or industry-specific federal law.[54] In 2012, the FTC issued a final report setting forth best practices for commercial entities that collect or use consumer

[50] 45 C.F.R. § 164.514(b)(2)(ii). The term "Covered Entity" is defined under the HIPAA Privacy Rule to include (1) a health plan; (2) a health care clearinghouse; and (3) a health care provider who transmits any health information in electronic form in connection with a transaction covered by Subchapter C of HIPAA. 45 CFR § 160.103. *Id.*

[51] *See also infra* Section III.A.1 (problematizing data protection laws that define PII by reference to an enumerated list of data elements).

[52] 45 C.F.R. § 164.514(c).

[53] 45 C.F.R. § 164.502(d)(2)(i)–(ii).

[54] *See supra* Section II.B.1.

data.[55] The report contained a framework (the "FTC Privacy Framework") for protecting consumer privacy that urged companies to adopt three practices: (1) privacy by design; (2) simplified choice; and (3) greater transparency. The report called on companies to make "privacy the 'default setting' for commercial data practices and giv[e] consumers greater control over the collection and use of their personal data through simplified choice and increased transparency."[56]

Importantly for purposes of this Note, the FTC Privacy Framework contains an exemption for anonymized data: The scope of the FTC Privacy Framework is limited to commercial entities that collect or use consumer data that can be "*reasonably linked* to a specific consumer, computer, or other device."[57] The FTC's inclusion of "reasonably linked" ignited concerns that the FTC Privacy Framework provided "less incentive for a business to try to de-identify the data it maintains" since, "with improvements in technology and the ubiquity of public information, more and more data could be 'reasonably linked' to a consumer, computer or device."[58] In response to these concerns, the FTC revised its privacy framework to qualify the definition of "reasonably linked."

---

[55] FED. TRADE COMM'N, *supra* note 12. The FTC explained the purpose of the Report, stating:

> The final framework is intended to articulate best practices for companies that collect and use consumer data . . . [and] intended to assist Congress as it considers privacy legislation. To the extent the framework goes beyond existing legal requirements, the framework is not intended to serve as a template for law enforcement actions or regulations under laws currently enforced by the FTC.

Id. at iii.

[56] *Id.* at i.

[57] *Id.* at iv (emphasis added). In order to address concerns about "undue burdens on small businesses," the report additionally exempts commercial entities that collect only non-sensitive data from fewer than 5000 consumers a year, provided that they do not share the data with third parties.

[58] *Id.* For the FTC's full discussion of anonymization and response to these comments, see *id.* at 18–22 (explaining that the exemption generated the most comments to the report from a wide range of interested parties).

These revisions clarified that data is not reasonably linkable—and thus outside the scope of the FTC's Privacy Framework—to the extent that a company (1) takes "reasonable measures" to ensure that the data is de-identified; (2) publicly commits not to try to re-identify the data; and (3) contractually prohibits downstream recipients from trying to re-identify the data.[59] The FTC explained that if a company takes steps to re-identify de-identified data in violation of this "reasonable linkability" standard, "its conduct could be actionable under Section 5 of the FTC Act."[60] The FTC further elucidated the term "reasonable measures," stating that a company "must achieve a reasonable level of justified confidence that the data cannot reasonably be used to infer information about, or otherwise be linked to, a particular consumer, computer, or other device."[61] The FTC explained that "reasonable measures" entail a subjective and factual inquiry that depends on the circumstances at issue, including the available methods and technologies, the nature of the data, and the purposes for which the data will be used.[62]

By broadening its exemption to incentivize commercial entities to de-identify consumer data, the FTC operates under the assumption that anonymization per its standards protects consumer privacy better than the combined privacy and data security rules imposed by FTC's Privacy Framework—namely, privacy by design, simplified choice, and greater transparency. This is the case for any data protection law, such as HIPAA, that exempts anonymized data from its scope.

III. DEANONYMIZATION AND PRIVACY HARM

Data anonymization fails to irrevocably protect the privacy of data subjects due to the potential for deanonymization—the

[59] *Id.* at iv, 22 ("The clarification of the framework's reasonable linkability standard is designed to help address the concern that the standard is overly broad [and] gives companies an incentive to collect and use data in a form that makes it less likely the data will be linked to a particular consumer or device, thereby promoting privacy.").

[60] *Id.* at 21.

[61] *Id.*

[62] *Id.*

linking of anonymized data records to auxiliary information to re-identify data subjects. As a consequence of deanonymization risk, the exemptions that exist for anonymized data under U.S. data protection laws do not provide absolute privacy protections to anonymized data subjects, and in fact result in re-identified data subjects receiving no statutory privacy protections at all.

The prevalence of big data and ubiquitous data collection today renders deanonymization increasingly feasible, as adversaries such as data brokers amass consumer data, and information about consumers becomes publicly available online. Moreover, there are strong financial incentives to re-identify consumer data for purposes such as targeted marketing. As the re-identification of consumer data becomes a viable and lucrative business, the risk of deanonymization increases and consumers become more exposed to privacy harm. Accordingly, this Note argues that the current lack of statutory privacy protections for re-identified data subjects is inadequate—particularly in light of pervasive data collection, technological advancements enabling new forms of sensitive data collection, and society's growing reliance on technology.

Section III.A will discuss deanonymization, the U.S. regulatory response to the risk of deanonymization, and the scholarly debate around the future of anonymization as a tool to protect consumer privacy. Section III.B will problematize the failure of anonymization against the modern realities of technological dependency, omnipresent data collection, and technological advancements that enable unprecedented forms of sensitive data collection.

A. The Failure of Anonymization

1. Deanonymizaiton

Deanonymization[63] occurs when adversaries re-identify anonymized data subjects by linking anonymized data records

[63] Some scholars prefer to use the terms "re-identification" and "de-identification" over "anonymization" and "deanonymization," since "the concept of 'anonymity' or 'anonymization' . . . implicitly guarantees

to outside, or auxiliary, information—a practice known as a linkage attack.[64] To understand this concept, consider Dr. Latanya Sweeney's re-identification of the hospitalization records of Governor Weld. An insurance agency had released anonymized hospitalization records to the public for research, removing direct identifiers but leaving demographic and sensitive health data. Dr. Sweeney matched these records to publicly available voter registration records containing similar demographic data to re-identify Governor Weld.[65]

The potential for deanonymization exists across the range of anonymization techniques, and certain techniques render deanonymization easier than others.[66] Of course, the more anonymous the data is, the lower the risk of deanonymization; however, even techniques that never release the raw data—such as aggregation, noise addition, and substitution—have the potential to be deanonymized using the right outside information.[67] To support this claim, privacy experts have pointed to proof that the Census Bureau provided aggregated, city-block-level data that—despite not identifying particular houses or families—helped locate and send Japanese Americans to internment camps during World War II.[68]

The ability to execute a linkage attack is made easier by the growing availability of consumer information to private entities and the public: "The more information about a person

protection of identity." Rubinstein & Hartzog, *supra* note 18, at 707; *see also* Ohm, *supra* note 2, at 1744 ("[W]e need a new word for privacy-motivated data manipulation that connotes only effort, not success.").

64 Rubinstein & Hartzog, *supra* note 18, at 711.

65 Latanya Sweeney, *K-Anonymity: A Model for Protecting Privacy*, 10 INT'L J. ON UNCERTAINTY, FUZZINESS & KNOWLEDGE-BASED SYSTEMS 557, 558–59 (2002). *See also* Rubinstein & Hartzog, *supra* note 18, at 711.

66 *See supra* Section II.A.1 (discussing anonymization techniques).

67 *See supra* note 15 and accompanying text. *See also* Ohm, *supra* note 2, at 1756 (explaining that, not only do these techniques offer limited utility and require constant maintenance by the data administrator, but that they, too, can be reverse engineered with the requisite auxiliary data).

68 *See, e.g.*, Ohm, *supra* note 2, at 1756 (citing William Seltzer & Margo Anderson, *Population Association of America, After Pearl Harbor: The Proper Role of Population Data Systems in Time of War* (Mar. 28, 2000) (unpublished paper)). Ohm provides a number of other examples as well. *Id.*

that is known, the more likely it becomes that this information can be used to identify that person or determine further data about her."[69] And since the "phenomenon of data availability heightens the ability to turn non-PII into PII,"[70] data protection laws that enumerate a static list of identifiers to define PII, and that create anonymization exemptions for data obstructing those identifiers, have been criticized for their "arbitrar[y] . . . categorization[s]."[71] Ohm, for instance, has problematized HIPAA's Safe Harbor standard:

> By enumerating eighteen identifiers, the [HIPAA] Privacy Rule assumes that any other information that might be contained in a health record cannot be used to reidentify. We now understand the flaw in this reasoning, and we should consider revising the Privacy Rule as a result . . . . Easy reidentification makes PII-focused laws like HIPAA underproductive by exposing the arbitrariness of their intricate categorization and line drawing.[72]

Several high-profile anonymization failures have drawn attention to the risk of deanonymization.[73] For instance, in 2006, AOL published a sample of de-identified search queries, including searches such as "depression and medical leave," "fear that spouse contemplating cheating," and "how to kill your wife."[74] AOL suppressed the PII, including IP address, and substituted each AOL username with a unique numeric

---

[69] Paul M. Schwartz & Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 N.Y.U. L. REV. 1814, 1843 (2011); Scott Berinato, *There's No Such Thing as Anonymous Data,* HARV. BUS. REV. (Feb. 9, 2015), https://hbr.org/2015/02/theres-no-such-thing-as-anonymous-data [perma.cc/YF4C-6S6C] ("Anonymization . . . is 'inadequate' and ultimately doomed to fail with large metadata—the kind of publicly available big data that so many companies are tapping into.").

[70] Schwartz & Solove, *supra* note 69, at 1812.

[71] Ohm, *supra* note 2, at 1740.

[72] Ohm, *supra* note 2, at 1737–38, 1740. *See also* notes 49–51 and accompanying text (discussing HIPAA's Safe Harbor standard).

[73] *But see* Yakowitz, *supra* note 24, at 16, 36 (arguing that undue emphasis is placed on these publicized failures, and that data presents no more of risk of re-identification than other tolerated risks, such as garbage).

[74] *See* Ohm, *supra* note 2, at 1717.

code (e.g., No. 4417749).[75] However, the queries themselves contained personal information, and because every query was associated with a numeric code representing an AOL username, journalists were able to link the queries associated with the same AOL username to discern the identity of that user. Within a few days, the *New York Times* revealed the identity of sixty-two year old Thelma Arnold, who conducted multiple searches including "60 single men," "landscapers in Lilburn, Ga," "people with the last name Arnold," and "homes sold in shadow lake subdivision Gwinnet county georgia."[76]

In another publicized incident, Netflix published a dataset containing one hundred million anonymized movie ratings as part of its "Netflix Prize," an effort to improve its movie recommendations by awarding the first team to significantly improve its algorithms one million dollars. Each record included a unique subscriber ID, movie title, year of release, and rental date.[77] Shortly after the release, two researchers re-identified many of the Netflix subscribers contained in the dataset by matching their Netflix reviews with data from IMDb. The researchers found that if an adversary knew the movies that a Netflix subscriber had rented in a given time period, the adversary could reverse-engineer the data to discover the subscriber's entire viewing history.[78]

---

[75] *Id.*

[76] *See* Michael Barbaro & Tom Zeller Jr., *A Face Is Exposed for AOL Searcher No. 4417749*, N.Y. TIMES (Aug. 9, 2006), http://www.nytimes.com/2006/08/09/technology/09aol.html ("[I]t did not take much investigating to follow that data trail to Thelma Arnold . . . . [T]he detailed records of searches conducted . . . underscore how much people unintentionally reveal about themselves when they use search engines—and how risky it can be for companies like AOL, Google, and Yahoo to compile such data.").

[77] *See, e.g.*, Julianne Pepitone, *5 Data Breaches: From Embarrassing to Deadly: Netflix Accidentally Reveals Rental Histories*, CNN MONEY (2010), http://money.cnn.com/galleries/2010/technology/1012/gallery.5_data_breaches/ [perma.cc/S8EJ-6NC6].

[78] See Arvind Narayanan & Vitalty Schamtikov, *Robust De-anonymization of Large Sparse Datasets*, *in* 2008 IEEE SYMPOSIUM ON SEC. & PRIVACY 111 (2008). The researchers explained:

> [A]n adversary who knows a little bit about some subscriber can easily identify her record if it is present in the dataset, or, at the very least, identify a small set of records which

2. The U.S. Regulatory Response

The failure of anonymization has called into question whether anonymization can or should be exclusively relied upon to prevent privacy harm. However, the safe harbors for anonymized data that exist under U.S. data protection laws were created by lawmakers who knew of the potential for re-identification. For instance, in reference to the debate over its "reasonable linkability" standard,[79] the FTC's report containing the FTC Privacy Framework directly addressed the failure of anonymization, stating that "[t]here is significant evidence demonstrating that technological advances and the ability to combine disparate pieces of data can lead to identification of a consumer, computer, or device even if the individual pieces of data do not constitute PII."[80]

Nevertheless, the FTC declined to reject anonymization as an insufficient privacy protection. Explaining this decision, the FTC assessed the comments it received from privacy advocates and industry representatives. The FTC explained that privacy advocates support the "reasonable-linkability" standard due to (1) consumers' objections to being tracked even without the use of PII, (2) the ability of adversaries to re-identify "anonymous" data, (3) the existence of industries that have turned re-identification for marketing purposes into a commercial enterprise, and (4) consumers' privacy interest in data going beyond mere PII to aggregated or de-identified data.[81] The FTC also explained the opposition of industry representatives, citing their claims that (1) the risk associated with PII is not the same as that associated with data not containing PII, (2) the "reasonable linkability" standard is "potentially too open-ended to be practical" since, "given

> include the subscriber's record. The adversary's background knowledge need not be precise, e.g., the dates may only be known to the adversary with a 14-day error, the ratings may be known only approximately, and some of the ratings and dates may even be completely wrong.

Id. at 112.

79 *See supra* notes 57–62 and accompanying text.

80 FED. TRADE COMM'N, *supra* note 12, at 20.

81 *See id.* at 18–19.

enough time and resources, any data may be linkable to an individual," and (3) that requiring the same level of protection for all data would "undermine companies' incentive to avoid collecting data that is more easily identified or take steps to de-identify the data they use or collect."[82] In consideration of these opinions, the FTC settled on its qualified "reasonable-linkability" standard with exemptions for anonymized data.[83]

3. The Academic Debate over Anonymization

There has been lively academic debate surrounding the future of anonymization as a privacy protection among both legal scholars and privacy experts, including Paul Ohm, Jane Yakowitz, Ira S. Rubinstein, and Woodrow Hartzog.[84] Their views lend to very different regulatory approaches.

Ohm has taken a critical view of anonymization, categorizing it as a failure and pushing for its abandonment as an exclusive means of privacy protection.[85] Because the definition of PII expands as technological advancements enable quasi-identifiers to be used to re-identify data subjects, Ohm argues that privacy regulations should not enumerate specific data elements that constitute PII, and that any law that draws distinctions based solely on whether particular data types can be linked to identify should be reevaluated.[86] Ohm has also dismissed three possible solutions as insufficient: (1) harm compensation; (2) waiting for technology

---

82 *See id*. at 19–20.

83 *See supra* notes 57–62 and accompanying text.

84 For an illustration of their views on anonymization, see Ohm, *supra* note 2; Rubinstein & Hartzog, *supra* note 18; Yakowitz, *supra* note 24.

85 Ohm, *supra* note 2, at 1742–43 ("At the very least, we must abandon the pervasively held idea that we can protect privacy by simply removing personally identifiable information . . . . [Anonymization] should no longer be considered to provide meaningful guarantees of privacy.").

86 "No matter how effectively regulators follow the latest reidentification research, folding newly identified data fields into new laws and regulations, researchers will always find more data field types they have not yet covered. The list of potential PII will never stop growing until it includes everything." *Id.*

to "save us"; and (3) banning re-identification.[87] Rather, Ohm suggests weighing the benefits of information flow with the costs of privacy harm and incorporating risk assessment strategies.[88] He further argues that neither industry-specific nor comprehensive, cross-industry privacy reform alone will be effective, and suggests instead implementing general regulation that sets a realistic privacy floor while tailoring specific laws to address industry-specific privacy risks.[89]

Yakowitz represents the opposite side of the debate. While she recognizes the limitations of anonymization, she argues that Ohm overstates the risk of harm and that the benefits of information flow exceed the deanonymization risk.[90] She contends that guarding anonymized information leads to a "tragedy of the data commons": When a "data subject depletes the commons by removing his data," "the marginal detriment of his decision is externalized and shared across the entire population" while "he enjoys the full value of the avoided risk of re-identification."[91] She further criticizes current U.S. legislation for its regulation of the *release* of data, as opposed to its use.[92] She concludes that public research data should in fact be easier to disseminate, and argues that while there should be harsh punishment for adversaries who intentionally re-identify data subjects, data administrators should receive immunity from statutory and common law privacy claims if they undergo basic anonymization techniques.[93]

Finally, Rubinstein and Hartzog offer a middle-ground, process-based approach that focuses on risk assessment and the implementation of technical, physical, and procedural safeguards—including data flow controls—in addition to

87 *See id.*

88 *Id.* at 1759.

89 *Id.* at 1762.

90 *See* Yakowitz, *supra* note 24.

91 *Id.* at 4.

92 *Id.* at 42.

93 *Id.* at 5.

anonymization.[94] They state that, in lieu of today's output-based approaches that strive for perfect anonymization,

> A more sustainable approach would focus on the preconditions and processes necessary for protection. It is hard to ensure protection. It is easier, however, to ensure that data custodians follow appropriate processes for minimizing risk, which may include both deidentification in combination with legal and administrative tools. . . . Approaches that focus on transparency, disclosure, harm and permission all seem inadequate, at least by themselves, to respond to the failure of anonymization.[95]

They suggest that the solution be contextually sensitive to account for factors such as the motivation for re-identification, harms that can result from re-identification, and the utility of the de-identified data.[96] They also suggest that the solution be "risk tolerant": "By focusing on process instead of output, data release policy can aim to raise the cost of reidentification and sensitive attribute disclosure to acceptable levels without having to ensure perfect anonymization."[97]

While the anonymization debate in the United States is far from resolved, the regulatory status quo retains exemptions for anonymized data. This Note will now problematize the lack of statutory privacy protections for re-identified data subjects in light of several modern realities—specifically, consumers' growing reliance on technology; technological advancements that enable omnipresent and increasingly sensitive data collection; and the financial incentives that exist to re-identify consumer data.

---

[94] Rubinstein & Hartzog, *supra* note 18, at 702–03, 706, 737 ("[W]e recommend including both deidentification techniques and controls on data flow as part of data release policy.").

[95] *Id.* at 729–31.

[96] *Id.* at 735–36 ("All of these factors mean that a 'one size fits all' standard for data release policy will not be effective. Such attempts are doomed to be either over-protective or under-protective."). They accordingly applaud the FTC's "reasonably linkability" standard. *Id.* at 736.

[97] *Id.* at 736–37.

B. Modern Realities that Increase the Privacy Harm Resulting from Deanonymization

1. The Convenience/Privacy Tradeoff and Consumers' Reliance on Technology

Technology increasingly pervades consumers' lives, and while it adds enormous value, its pervasiveness comes at a cost: the "convenience/privacy tradeoff." As technologies ranging from smartphones and fitness trackers to search engines and geolocation-tracking applications offer compelling benefits, consumers "seem increasingly resigned to giving up fundamental aspects of their privacy for convenience . . . and have grudgingly accepted that being monitored by corporations . . . is just a fact of modern life."[98]

Similarly, a 2016 study by the Pew Research Center assessed the extent to which Americans would be willing to provide personal information in exchange for certain deals (e.g., discounts) and found that there are a "variety of circumstances under which many Americans would share personal information or permit surveillance in return for getting something of perceived value."[99] Yet, it also found that they are "frequently unhappy about what happens to that information once companies have collected it."[100] Such tradeoffs are increasingly common today. Consider, for

98 Liz Mineo, *On Internet Privacy, Be Very Afraid*, HARV. L. TODAY (Aug. 25, 2017), https://today.law.harvard.edu/internet-privacy-afraid/ [perma.cc/8XLY-VHK9) (interviewing cybersecurity expert Bruce Schneier). *See* Robert Stroud, *The Convenience/Privacy Trade-Off on the Internet of Things*, WIRED (Dec. 17, 2013), http://insights.wired.com/profiles/blogs/the-convenience-privacy-trade-off-on-the-internet-of-things [perma.cc/5K73-DCE9] ("Whether it is using geolocation tracking on photo-sharing apps or understanding search habits for targeted ads, our complex, interconnected world is increasingly asking customers and enterprises alike to weigh the pros and cons of convenience factors vs. threats to data privacy and security. As hype gives way to concerns, individuals and organizations must . . . account for both the opportunities and hazards . . . .").

99 Lee Rainie & Maeve Duggan, *Privacy and Information Sharing*, PEW RES. CTR. (Jan. 14, 2016), http://www.pewinternet.org/2016/01/14/privacy-and-information-sharing/ [perma.cc/U89Z-49JH].

100 *Id.*

instance, smart thermostats—devices that allow consumers to save on their energy bills by monitoring their movements around the home[101]—or telematics devices—devices that allow consumers to save on their car insurance by monitoring their driving habits, such as speed and distance driven.[102]

And consumers today are not merely encouraged to trade off privacy for perceived efficiency, financial, and other benefits. Frequently, they lack any real choice in the matter. As cybersecurity expert Bruce Schneier aptly explains,

> Consumers are concerned about their privacy and don't like companies knowing their intimate secrets. But they feel powerless and are resigned to the privacy invasions because they don't have any real choice. People need to own credit cards, carry cellphones, and have email addresses and social media accounts. That's what it takes to be a fully functioning human being in the 21st century. This is why we need the government to step in.[103]

As society adjusts to technological innovation, it develops a reliance on the benefits that technology brings, and consumers lose their ability to opt-out.[104] It is simply unrealistic to say that consumers have free choice over whether to avail themselves of technology, and whether to

---

[101] *See e.g.*, *id.*; *see generally* SMART THERMOSTAT GUIDE, https://smartthermostatguide.com/ [perma.cc/9RFV-CBUK].

[102] *See, e.g.*, Rainie & Duggan, *supra* note 99; *What Is a Telematics Device?*, ALLSTATE (Jan. 2014), https://www.allstate.com/tools-and-resources/car-insurance/telematics-device.aspx [perma.cc/CUL2-27TT].

[103] Mineo, *supra* note 99.

[104] *Id.* For just one example, a growing number of businesses today refuse to accept cash from their customers. *See, e.g.*, Andy Newman, *Cash Might Be King, but They Don't Care*, N.Y. TIMES (Dec. 25, 2017), https://www.nytimes.com/2017/12/25/nyregion/no-cash-money-cashless-credit-debit-card.html ("[In Midtown and some other neighborhoods across New York City, cashless is fast on its way to becoming normal."). For there is "no Federal statute mandating that a private business, a person, or an organization must accept currency or coins as payment for goods or services." *FAQs*, BD. OF GOVERNORS OF THE FED. RES. SYS. (June 17, 2011), https://www.federalreserve.gov/faqs/currency_12772.htm [perma.cc/LM8R-96F6]. As cashless becomes the new normal, the need for credit cards grows.

hand their personal data over to the businesses behind those technologies that collect, store, and share consumer data.[105]

2. Technological Advancements Enabling Unprecedented Data Collection

Consumers' growing resignation toward, and lack of control over, their data-sharing is problematic in light of technological advancements that enable unprecedented forms of sensitive data collection in every corner of their lives. For instance, the advancement of IoT increasingly connects everyday objects to the internet while both (1) creating new channels for collecting consumer data and (2) enabling the collection of new kinds of consumer data. These devices include home automation systems, autonomous cars, activity tracking devices, and smart medical devices[106]—technologies the FTC described in 2015 as "unimaginable a decade ago."[107]

In 2015, the FTC issued a report on IoT recognizing that these devices create significant privacy risks: some risks involving the "direct collection of sensitive personal information, such as precise geolocation, financial account numbers, or health information," and others arising from the "collection of personal information, habits, locations, and

[105] "And 'buyer beware' is putting too much onus on the individual." Mineo, *supra* note 99 (quoting cybersecurity expert Bruce Schneier).

[106] For a discussion of IoT and the medical industry, see, e.g., Karen Taylor, *By 2020 the Smart Hospital Will Be a Reality,* FUTURE HEALTH INDEX (June 13, 2017), https://www.futurehealthindex.com/2017/06/13/by-2020-the-smart-hospital-will-be-a-reality/ [perma.cc/LY3G-HPXQ]; Sarah Neville, *US Regulators Approve First Digital Pill with Tracking System*, FIN. TIMES (Nov. 14, 2017), https://www.ft.com/content/267b890a-a9b5-11e7-ab55-27219df83c97 [perma.cc/3KRY-CCBN].

[107] "Experts estimate that, as of this year, there will be 25 billion connected devices, and by 2020, 50 billion." FED. TRADE COMM'N, INTERNET OF THINGS: PRIVACY & SECURITY IN A CONNECTED WORLD 1 (2015), https://www.ftc.gov/system/files/documents/reports/federal-trade-commission-staff-report-november-2013-workshop-entitled-internet-things-privacy/150127iotrpt.pdf [perma.cc/UG4R-TEJD].

physical conditions over time, which may allow an entity that has not directly collected sensitive information to *infer it*."[108]

Regarding data volume, the FTC acknowledged:

> The sheer volume of data that even a small number of devices can generate is stunning: one participant indicated that fewer than 10,000 households using the company's IoT home-automation product can 'generate 150 million discrete data points a day' or approximately one data point every six seconds for each household.[109]

As the FTC observed, "Such a massive volume of granular data allows those with access to the data to perform analyses that would not be possible with less rich data sets."[110] And importantly, as this data becomes available to adversaries, linkage attacks become easier and deanonymization risk increases. Moreover, not only do technological advancements like IoT make data collection more pervasive and generate greater data volume, but they enable the collection of new *kinds* of sensitive data. The exposure of consumers' sexually explicit data, a category classified as sensitive under the GDPR, illustrates this point.[111]

In 2016, a lawsuit against an adult product company initiated a conversation around the collection of sexually explicit data.[112] The complaint alleged that the smartphone-controlled device used its internet connectivity to transmit data on consumers' usage, including user identifiers and the

---

[108] *Id.* at 14 (emphasis added) (noting commentators' concerns that "the trend towards abundant collection of data creates a 'non-targeted dragnet collection from devices in the environment'" and that "companies might use this data to make credit, insurance, and employment decisions").

[109] *Id.*

[110] *Id.* at 15 ("According to a [workshop] participant, 'researchers are beginning to show that existing smartphone sensors can be used to infer a user's mood; stress levels; personality type; bipolar disorder; demographics (*e.g.*, gender, marital status, job status, age); smoking habits; overall well-being; progression of Parkinson's disease; sleep patterns; happiness; levels of exercise; and types of physical activity or movement.'").

[111] *See supra* note 38 and accompanying text.

[112] *See* Complaint at 2, N.P. v. Standard Innovation (US) Corp., No. 1:16-cv-08655 (N.D. Ill. filed Sept. 2, 2016).

date, duration, and settings of each use, in real-time back to its manufacturer without consumers' consent.[113] In response, the company stated, "We do collect certain limited data to help us improve our products and for diagnostic purposes. As a matter of practice, we use this data in the aggregate, non-identifiable form."[114] Though the parties settled, the company updated its privacy policies and added an option for customers to opt-out of sharing anonymous app usage data.[115] The lawsuit highlighted the unprecedented capacity of IoT to collect sensitive consumer data—a fact that the lawyer for the firm representing plaintiffs acknowledged: "[O]f all the privacy violations our firm has prosecuted, this is among the most personal and invasive we have ever encountered."[116]

The potential for unprecedented sexually explicit data collection—and the collection of numerous other forms of sensitive data—is not limited to advancements in IoT. Widely popular dating sites and applications such as Tinder also collect information about their users' sex lives, such as sexual

---

113 *Id.*

114 *Our Commitment to Customer Privacy and Security,* WE-VIBE (Aug. 12, 2016), http://we-vibe.com/blog/our-commitment-to-customer-privacy-and-security [perma.cc/4LH6-DYCJ].

115 N.P. v. Standard Innovation Corp—*We-Vibe Settlement*, HEFFLER CLAIMS GRP., http://www.sicclassactionsettlement.com/ [perma.cc/K7DK-ZR97] (providing settlement details); *We-Connect App and Privacy Update*, WE-VIBE (Oct. 3, 2016), http://we-vibe.com/blog/we-connect-app-and-privacy-update/ [perma.cc/4QY9-GP4R] (explaining the privacy update).

116 The lawyer observed that "[w]hile this particular example shocks the conscience more than most, it should also put consumers on notice that the 'internet of things,' while providing convenience and assistance in their daily lives, is also ripe for unauthorized data collection." Molly Redden, *Tech Company Accused of Collecting Details of How Consumers Use Sex Toys,* GUARDIAN (Sept. 14, 2016), https://www.theguardian.com/us-news/2016/sep/14/wevibe-sex-toy-data-collection-chicago-lawsuit [perma.cc/M62G-RB2M]. Though this type of data collection is arguably limited to a narrow consumer base, the global adult toys market is expected to exceed $29 billion by 2020. *Global Adult Toys Market to Exceed USD 29 Billion by 2020, According to Technavio*, BUS. WIRE (Apr. 12, 2016), https://www.businesswire.com/news/home/20160412005747/en/Global-Adult-Toys-Market-Exceed-USD-29 [perma.cc/5MUE-K78S].

orientation and preferences.[117] In some contexts, this data can be sensitive, and its unauthorized release can be extremely harmful. In 2015, hackers stole user data collected by Ashley Madison, a dating website designed for extramarital affairs, and released the data onto the dark web. The release exposed thirty-six million users' data, including their names, street addresses, e-mail addresses, and interests.[118] People whose data had been released suffered harms ranging from public humiliation for themselves and their families to suicide and, in countries where infidelity and homosexuality are punishable offenses—such as Saudi Arabia, Pakistan, and the Philippines—the risk of prison, flogging, and execution.[119]

The Ashley Madison breach also highlighted the capacity for extortion in the data privacy context. Following the breach, extortionists blackmailed data subjects,[120] underscoring the fact that sensitive data can create leverage for extortion. This potential for extortion leads to an undesirable incentive for adversaries to re-identify sensitive consumer data.

---

117 *See generally* Judith Duportail, *I Asked Tinder for My Data. It Sent Me 800 Pages of My Deepest, Darkest Secrets*, GUARDIAN (Sept. 26, 2017), https://www.theguardian.com/technology/2017/sep/26/tinder-personal-data-dating-app-messages-hacked-sold [perma.cc/Q2ZY-G7K8].

118 Kim Zetter, *Hackers Finally Post Stolen Ashley Madison Data*, WIRED (Aug. 18, 2015), https://www.wired.com/2015/08/happened-hackers-posted-stolen-ashley-madison-data/ [perma.cc/D9JC-3KQY].

119 *See, e.g.*, *Ashley Madison Hack: 2 Unconfirmed Suicides Linked to Breach, Toronto Police Say*, CBC NEWS (Aug. 25, 2015), http://www.cbc.ca/news/canada/toronto/ashley-madison-hack-2-unconfirmed-suicides-linked-to-breach-toronto-police-say-1.3201432; Patrick Cain, *Where 1,296 Gay Ashley Madison Users Face Prison, Flogging, Execution,* GLOBAL NEWS (Sept. 2, 2015), http://globalnews.ca/news/2186587/where-1296-gay-ashley-madison-users-face-prison-flogging-execution/; Natasha Noman, *The Dark Side of the Ashley Madison Hack that Nobody's Talking About*, MIC (Aug. 20, 2015), https://mic.com/articles/124169/the-ashley-madison-hack-could-effect-those-who-live-in-country [perma.cc/5468-V8RN].

120 *See, e.g.*, Cory Bennett, *Extortion Begins for Ashley Madison Hack Victims*, HILL (Aug. 21, 2015), http://thehill.com/policy/cybersecurity/251682-extortion-begins-for-ashley-madison-hack-victims [perma.cc/MH6N-TJGH].

3. Targeted Marketing, Data Brokers, and Incentives to De-Identify Consumer Data

Consumers today leave a wealth of personal data in the hands of companies that have a financial interest in collecting and selling their data for a broad range of purposes, including targeted marketing, product development, and market research. As Schneier explained, "[s]urveillance is the business model of the internet. Everyone is under constant surveillance by many companies, ranging from social networks like Facebook to cellphone providers. The data is collected, compiled, analyzed, and used to try to sell us stuff. . . . We're the product, not the customer."[121] Regarding constraints on surveillance, Schneier remarked, "[o]ur system is optimized for companies that do everything that is legal to maximize profits, with little nod to morality"—a legal structure that has been termed "surveillance capitalism."[122]

For instance, targeted marketing has emerged as a strong financial incentive[123] to track and aggregate consumer data. As the FTC has explained, "The practice, which is typically invisible to consumers, allows businesses to align their ads more closely to the inferred interests of the audience. . . . [B]usinesses generally use 'cookies' to track consumers' [online] activities and associate those activities with a particular computer or device."[124] The FTC also recognized the significant potential for re-identification in this context:

---

121 Mineo, *supra* note 99 (quoting cybersecurity expert Bruce Schneier, "What we have is many 'Little Brothers:' Google, Facebook, Verizon, etc. They have enormous amounts of data on everybody, and they want to monetize it.").

122 *Id.*

123 Stakeholders include advertisers, who use targeted marketing to increase their brand awareness and thereby sell products, and publishers, who raise revenue by offering valuable ad placement. *See, e.g.*, *The True Cost of Ad Blockers for Advertisers and Publishers*, ALTITUDE, http://altitudedigital.com/byline/the-true-cost-of-ad-blockers-for-advertisers-and-publishers/ [perma.cc/Q5ME-ZGSD].

124 FED. TRADE COMM'N, SELF-REGULATORY PRINCIPLES FOR ONLINE BEHAVIORAL ADVERTISING 2–3 (2009), https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-staff-report-self-regulatory-

> For example, a consumer's Internet activity might reveal the restaurants in the neighborhood where she eats, the stores at which she shops, the property values of houses recently sold on her block, and the medical conditions and prescription drugs she is researching; when combined, such information would constitute a highly detailed and sensitive profile that is potentially traceable to the consumer.[125]

Meanwhile, an entire industry has emerged for the sole purpose of collecting, consolidating, monetizing, and selling consumer data. As the FTC recognized, "In today's economy, Big Data is big business. And data brokers—companies that collect consumers' personal information and resell or share that information with others—play a key role."[126] Data brokers collect data about consumers from a wide range of sources—social media accounts, public records, consumer purchase data, and web browsing activities, to name a few—without consumers' knowledge, and consolidate this data to resell it for purposes such as targeted marketing campaigns.

According to an FTC report examining nine data brokers, such firms "collect and store billions of data elements, including some on nearly every U.S. consumer. . . . one of the nine data brokers has 3,000 data segments for nearly every

principles-online-behavioral-advertising/p085400behavadreport.pdf [perma.cc/9HHC-24WD] (containing guidance "designed to serve as the basis for industry self-regulatory efforts to address privacy concerns").

[125] *Id.* at 22–23.

[126] Bridget Small, *FTC Report Examines Data Brokers*, FED. TRADE COMM'N (May 27, 2014), https://www.consumer.ftc.gov/blog/2014/05/ftc-report-examines-data-brokers [perma.cc/LE6P-558U]. *See also* FED. TRADE COMM'N, DATA BROKERS: A CALL FOR TRANSPARENCY AND ACCOUNTABILITY (2014), https://www.ftc.gov/system/files/documents/reports/data-brokers-call-transparency-accountability-report-federal-trade-commission-may-2014/140527databrokerreport.pdf [perma.cc/8SBM-BSAN] [hereinafter FTC DATA BROKER REPORT]; Press Release, Fed. Trade Comm'n, FTC Recommends Congress Require the Data Broker Industry to be More Transparent and Give Consumers Greater Control over Their Personal Information (May 27, 2014), https://www.ftc.gov/news-events/press-releases/2014/05/ftc-recommends-congress-require-data-broker-industry-be-more [perma.cc/2FTN-FZKV] [hereinafter FTC Report Press Release].

U.S. consumer."[127] In releasing the report, FTC Chairwoman Edith Ramirez stated, "The extent of consumer profiling today means that data brokers often know as much—or even more—about us than our family and friends, including our online and in-store purchases, our political and religious affiliations, our income and socioeconomic status, and more."[128]

The emergence of data brokers is especially alarming in the anonymization context because data brokers have an incentive to re-identify consumer data. That is, a largely unregulated[129] and highly lucrative[130] industry exists that enables and incentivizes adversaries to deanonymize data. The FTC recognized this fact in its report containing the FTC Privacy Framework, referencing a commentator who "pointed out that certain industries extensively mine data for marketing purposes and that re-identification is a commercial enterprise."[131] The FTC later acknowledged, "not only is it possible to re-identify non-PII data through various means, businesses have strong incentives to actually do so."[132]

The FTC also noted the growing risk of re-identification in its 2014 report on data brokers, which explained, "[D]ata brokers generally do not delete the consumer's information from their systems. Instead, they maintain the information in order to be able to match records that they may receive in the future. . . ."[133] To put it simply, these data brokers house a massive quantity of consumer data, which provides them with the ability to match anonymized data records to their data stores and thereby re-identify data subjects.

---

127 Small, *supra* note 126.

128 FTC Report Press Release, *supra* note 126. For instance, in one publicized incident, a teenager's father complained to Target about sending his daughter promotions for baby products, only to find out that his daughter was pregnant and had not told her parents yet. *See* Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES (Feb. 16, 2012), http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

129 *See, e.g.*, FTC Report Press Release, *supra* note 126.

130 One of the data brokers the FTC examined, Acxiom, reported over $800 million in revenue. FTC DATA BROKER REPORT, *supra* note 126, at 23.

131 FED. TRADE COMM'N, *supra* note 12, at 18–19.

132 *Id.* at 20 (citing Sorrell v. IMS Health Inc., 564 U.S. 552 (2011)).

133 FTC DATA BROKER REPORT, *supra* note 126, at 43.

IV. GUIDANCE FROM THE EUROPEAN UNION: ANONYMIZATION UNDER THE GENERAL DATA PROTECTION REGULATION

As Part II of this Note discussed, U.S. data protection laws currently provide safe harbors for data that has been anonymized in compliance with their standards. However, as Part III of this Note explained, data anonymization does not permanently protect the privacy of data subjects due to the potential for deanonymization—the re-identification of data subjects. Deidentification is a real risk, as data brokers have strong financial incentives to re-identify consumer data, and the ubiquity of data collection makes deanonymization easier.

Because U.S. data protection laws exempt anonymized data, data subjects who are re-identified ultimately receive no statutory privacy or data security protections in the United States. Technological advancements, such as those discussed in Section III.B, exacerbate privacy challenges by enabling pervasive and sensitive data collection, incentivizing users to trade privacy for convenience, and increasing consumers' reliance on technology. Given that consumers often lack any real choice in deciding whether to "opt-out" of the modern realities of technology and data collection, the current lack of protection in the face of deanonymization is inadequate.

This Part will examine the European Union's approach to data anonymization under the existing Data Protection Directive (the "Directive") and the forthcoming GDPR to suggest that the United States follow the European Union's lead. In particular, this Part will argue that the United States should amend its data protection laws to introduce a principle akin to the GDPR's "pseudonymization"—the concept that certain deidentified data that has not been rendered truly anonymous due to the potential for re-identification falls within the scope of the governing privacy rule, but receives relaxed requirements designed to incentive anonymization and thereby reduce the linkability of the data to its subjects.

A. Anonymization in the European Union

Generally, the European Union has more stringent privacy regulations than the United States.[134] Unlike the United States, the European Union—specifically, the European Commission (the "EC")—enforces comprehensive, cross-industry data privacy regulation. Currently, the European Union operates under the 1995 Directive. However, the EC adopted the GDPR on April 27, 2016, which is scheduled to replace the Directive on May 25, 2018. While both the Directive and the GDPR are comprehensive data privacy regulations, the GDPR imposes many heightened privacy protections, including key anonymization requirements.[135]

1. The 1995 Data Protection Directive

The existing Directive regulates the processing of "personal data."[136] The Directive defines "personal data" as "any information relating to an identified or identifiable natural person," and defines an "identifiable" person as one "who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural, or social identity."[137]

Recital 26 of the Directive exempts anonymized data from any regulation. In particular, it states: "[T]he principles of protection shall not apply to data rendered anonymous in such

134 As Bruce Schneier noted, there are more controls over government surveillance in the United States than in Europe, and more constraints on corporations in Europe than in the United States. Mineo, *supra* note 99 (explaining that this variance has occurred because "Americans tend to mistrust government and trust corporations," while "Europeans tend to trust government and mistrust corporations").

135 *See generally Top 10 Operational Impacts of the GDPR,* IAPP, https://iapp.org/resources/article/top-10-operational-impacts-of-the-gdpr/ [perma.cc/TA9A-D78U].

136 The concept of "personal data" under E.U. law is comparable to that of PII under U.S. law. Note that Article 8 of the Directive includes a separate category of "special" personal data (i.e., sensitive data), which receives extra protections under the Directive. *See supra* note 38.

137 *Id.* at art. 2(a).

a way that the data subject is no longer identifiable."[138] To determine whether a data subject is "identifiable," Recital 26 provides that "account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person."[139]

In 2014, the Article 29 Data Protection Working Party[140] released a non-binding opinion (the "Opinion") assessing the strengths and weaknesses of various anonymization techniques "by taking account of the residual risk of reidentification inherent in each of them."[141] In reference to Article 26, the Opinion emphasized that "[a]n important factor is that the processing must be irreversible," and "as permanent as erasure, i.e. making it impossible to process personal data."[142] The Opinion concluded that "anonymization techniques can provide privacy guarantees and may be used to generate efficient anonymization processes, but only if their application is engineered appropriately," which requires clearly defined prerequisites and objectives determined on a case-by-case basis as well as a combination of different techniques.[143] It also stated that data controllers should "not rely on the 'release and forget'

138 *Id*. at rec. 26.

139 *Id.*

140 The Article 29 Data Protection Working Party (the "WP") is an independent European advisory body on data protection and privacy set up under Art. 29 of the Directive. *Id.* at art. 29.

141 Article 29 Data Protection Working Party, *Opinion 05/2014 on Anonymization Techniques*, EUROPA 3 (2014), http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf [perma.cc/HHF4-K4QZ]. The Opinion analyzed the following anonymization techniques: noise addition, permutation, differential privacy, aggregation, k-anonymity, l-diversity, and t-closeness. The opinion clarified that "'identification' not only means the possibility of retrieving a person's name and/or address, but also includes potential identifiability by singing out, linkability and inference." *Id.* at 10.

142 *Id.* at 5–6, 9 ("An effective anonymization solution prevents all parties from singling out an individual in a dataset, from linking two records within a dataset (or between two separate datasets) and from inferring any information in such dataset.").

143 *Id.* at 3–4.

approach," but rather identify new risks and re-evaluate residual risks regularly, assess whether the controls for identified risks suffice, and monitor and control the risks.[144]

Importantly, the Opinion introduced the concept of "pseudonymization," which "consists of replacing one attribute (typically a unique attribute) in a record by another."[145] The Opinion clarified that pseudonymization is "not a method of anonymisation" but "merely reduces the linkability of a dataset with the original identity of a data subject, and is accordingly a useful security measure."[146] To this effect, the Opinion emphasized that it is still possible to single out an individual's records, link records related to an individual, and infer information concerning an individual from pseudonymized data. The Opinion described the five most common pseudonymization techniques: (1) encryption with a secret key;[147] (2) hash function;[148] (3) keyed-hash function with stored key;[149] (4) deterministic encryption or keyed-hash function with deletion of the key;[150] and (5)

144 *Id.* at 24.

145 *Id.* at 20.

146 *Id.* at 3, 10.

147 *Id.* at 20 ("[T]he holder of the key can trivially re-identify each data subject through decryption of the dataset because the personal data are still contained in the dataset, albeit in an encrypted form. Assuming that a state-of-the-art encryption scheme was applied, decryption can only be possible with the knowledge of the key.").

148 *Id.* ("[T]his corresponds to a function which returns a fixed size output from an input of any size (the input may be a single attribute or a set of attributes) and cannot be reversed; this means that the reversal risk seen with encryption no longer exists. However, if the range of input values the hash function are known they can be replayed through the hash function in order to derive the correct value for a particular record.").

149 *Id.* ("[T]his corresponds to a particular hash function which uses a secret key as an additional input (this differs from a salted hash function as the salt is commonly not secret). A data controller can replay the function on the attribute using the secret key, but it is much more difficult for an attacker to replay the function without knowing the key as the number of possibilities to be tested is sufficiently large as to be impractical.").

150 *Id.* at 21 ("[T]his technique may be equated to selecting a random number as a pseudonym for each attribute in the database and then deleting the correspondence table.").

tokenization,[151] foreshadowing the changes surrounding the anonymization exemption implemented under the GDPR.

2. The General Data Protection Regulation

The GDPR is similar to the Directive in that it regulates only "personal data," but affords a number of new protections to data subjects. For instance, the concept of identifiability under the GDPR accounts for new types of potentially identifying information. Specifically, the GDPR expands the definition of an "identifiable" person to "one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person."[152]

With respect to anonymization, Recital 26 of the GDPR includes several new provisions and is more focused on re-identification risk than the Directive. Though the GDPR continues to exempt "anonymous" data,[153] it heightens the threshold for determining whether a natural person is "identifiable" by specifically including *indirect* means of re-identification, stating that "account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly."[154] Further, unlike the

151 *Id.* ("[T]his technique is typically applied in (even if it is not limited to) the financial sector to replace card ID numbers by values that have reduced usefulness for an attacker. It is derived from the previous ones being typically based on the application of one-way encryption mechanisms or the assignment, through an index function, of a sequence number or a randomly generated number that is not mathematically derived from the original data.").

152 GDPR art. 4(1) (emphasis added).

153 *Id.* at rec. 26 (stating "[t]his Regulation does not therefore concern the processing of such anonymous information," meaning "information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable").

154 *Id.* (emphasis added).

Directive, the GDPR expands on the meaning of "reasonably likely," stating that "account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of processing and technological developments."[155]

Critically for purposes of this Note, Recital 26 relies on the concept of pseudonymization. In keeping with the Opinion,[156] the term "pseudonymization" is defined under the GDPR as:

> [T]he processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.[157]

In other words, pseudonymization is a security measure that is used to reduce the linkability of data to its data subjects by separating the data from direct identifiers, but does not prevent re-identification.[158] Pseudonymization under the GDPR includes encryption, and the GDPR requires that the encryption "key" be kept separate and secure, instructing data administrators to implement appropriate safeguards to prevent the "unauthorized reversal of pseudonymization.[159]

Pseudonymous data is not exempt from the GDPR. Recital 26 states, "[p]ersonal data which have undergone pseudonymisation . . . should be considered to be information

---

155 *Id.*

156 *See supra* notes 140–151 and accompanying text.

157 *Id.* at art. 4(5).

158 Pseudonymization has been aptly described as "the conversion of data about an identified person into data about a merely 'identifiable' person." Waltraut Kotschy, *The New General Data Protection Regulation—Is There Sufficient Pay-Off for Taking the Trouble To Anonymize or Pseudonimize Data?*, FUTURE PRIVACY FORUM (Nov. 2016), https://fpf.org/wp-content/uploads/2016/11/Kotschy-paper-on-pseudonymization.pdf [perma.cc/58FY-A5S8].

159 GDPR rec. 75.

on an identifiable natural person,"[160] and "[t]he explicit introduction of 'pseudonymisation' in this regulation is not intended to preclude any other measures of data protection."[161] However, in recognition of the fact that pseudonymization reduces the linkability of data to its data subjects, and thereby "can reduce the risks to the data subjects,"[162] the GDPR explicitly encourages the practice of pseudonymization. Specifically, the GDPR states that pseudonymization is a security measure that companies should implement "by default" or "as soon as possible,"[163] and encourages data controllers to adopt the "pseudonymisation of personal data" as a "code of conduct" to "contribute to the proper application of this Regulation."[164]

Though pseudonymized data is not exempt from the GDPR, the GDPR includes a number of important "incentives to apply pseudonymisation when processing personal data."[165] Emphasizing that pseudonymized data retains information value, Recital 29 states:

> [M]easures of pseudonymisation should, whilst allowing general analysis, be possible within the same controller when that controller has taken technical and organizational measures necessary to ensure, for the processing concerned, that this Regulation is implemented, and that this additional information for attributing the personal data to a specific data subject is kept separately.[166]

Moreover, the GDPR creates incentives for data controllers to apply pseudonymization by allowing pseudonymization to satisfy several of the requirements that it imposes.[167] First,

160 *Id.*

161 *Id.* at rec. 28.

162 *Id.*

163 *Id.* at rec. 78.

164 *Id.* at art. 40(1)–(2)(d).

165 *Id.* at rec. 29.

166 *Id.* at rec. 29.

167 The incentives created by the GDPR to pseudonymize data have been thoroughly summarized by IAPP. *See* Gabe Maldoff, *Top 10 Operational Impacts of the GDPR: Part 8—Pseudonymization*, IAPP (Feb.

data controllers who pseudonymize personal data are allowed more flexibility to process data beyond original collection purposes. The GDPR requires that data be collected for "specified, explicit, and legitimate purposes and not further processed in a manner that is incompatible with those purposes," but creates an exception to the further processing limitation if data is done so in a way "compatible" with the initial purposes of collection, and states that one factor in determining such compatibility is whether the data was pseudonymized.[168] Second, pseudonymization can satisfy the "appropriate safeguard" required for processing personal data for "archiving purposes in the public interest, scientific or historical research purposes or statistical purposes."[169] Third, pseudonymization can satisfy the "data protection by design and by default" requirement introduced in the GDPR, which requires data controllers to implement appropriate technical and organizational measures "both at the time of the determination of the means for processing and at the time of the processing itself."[170] Fourth, pseudonymization can satisfy the GDPR's data security requirements—namely the requirement that controllers implement risk-based measures for protecting data security—and may allow data controllers to avoid notification requirements.[171] Finally, if the data controller applies pseudonymization prohibiting it from identifying a data subject without the use of additional information (e.g., by permanently deleting the key), the controller receives an exemption from the rights to access, rectification, erasure, and portability required under the GDPR; however, this exemption is inapplicable if a data subject provides the controller with additional information enabling the controller to re-identify the data subject.[172]

12, 2016), https://iapp.org/news/a/top-10-operational-impacts-of-the-gdpr-part-8-pseudonymization/ [perma.cc/V3UQ-KGRZ].

168 GDPR art. 5, art 6(4)(e).

169 *Id.* at art. 89(1), rec. 156.

170 *Id.* at art. 25(1).

171 *Id.* at art. 32

172 *Id.* at art. 11, art. 15–20.

B. The United States Should Follow the European Union's Approach to Anonymization Under the General Data Protection Regulation

1. The Benefit of the European Union's Approach

The United States currently embraces an all-or-nothing approach to anonymization. Either data is anonymized in compliance with the standards set by the relevant regulation and therefore exempt from the regulation entirely, or data is subject to the full extent of the regulation's requirements.[173]

Insofar as the FTC and other entities are correct in thinking that some form of an exemption or other regulatory incentive is necessary to incentivize data controllers to anonymize their data,[174] U.S. data protection laws currently provide no means of incentivizing the de-identification of data subjects while providing statutory privacy protections to those same data subjects. The obstruction or removal of PII from a dataset protects consumer privacy to the extent that it reduces the linkability of data to its data subjects and thereby reduces the risk of identification. Thus, the practice of anonymization should indeed be encouraged. However, due to the failure of anonymization to permanently protect the privacy of data subjects—and given the sensitive nature of the data that may be at issue as well as consumers' growing inability to opt out of data-sharing—U.S. data protection laws should protect anonymized data subjects if they *are* re-identified—a realistic concern given the ubiquity of data collection and the incentives that data brokers have to re-identify data.[175] Currently, U.S. data protection laws provide no privacy protections to re-identified data subjects, regardless of the potential sensitivity of the data at issue.

An ideal solution from a privacy perspective would marry the practice of anonymization with additional statutory protections for consumers, such as transparency, choice, and data security. And while U.S. data protection laws fail to

173 *See supra* Section II.B.

174 *See supra* notes 58, 82 and accompanying text.

175 *See supra* notes 129–133 and accompanying text.

conceive of a less demanding standard of anonymization or any form of de-identification beyond that which is rewarded with a complete exemption, the European Union has accomplished just that. By introducing pseudonymization into the GDPR, the European Union has implemented a means to both (1) incentivize de-identification,[176] thereby reducing the linkability of data to its data subjects, and (2) keep de-identified data within the scope of the privacy regulation, thereby affording data subjects additional privacy protections.[177] The European Union has conceived of multiple tiers of de-identification—namely (1) anonymization and (2) pseudonymization—that vary in their degree of re-identification risk and information value, and has subjected those categories of data to different privacy requirements based on their relative risk of re-identification.

2. The European Union's Approach Accords with Solutions Proposed in the United States

The European Union's solution to the failure of anonymization falls within the range of proposals that have surfaced in the anonymization debate in the United States. The European Union does not entirely embrace the approach of Paul Ohm—the abandonment of de-identification—nor does the European Union fully adopt the alternative approach of Jane Yakowitz—prioritizing the free flow of information.[178]

Rather, the European Union meets both scholars half way, balancing privacy with utility by coupling de-identification with additional privacy requirements to address the failure of de-identification and encouraging the free flow of information by creating a less demanding standard of de-identification. In this respect, the European Union has become risk-tolerant, willing to accept a reality in which de-identification fails but additional safeguards exist to protect consumer privacy.

---

176 An analysis of the extent to which the European Union's incentives under the GDPR to pseudonymize data, discussed *supra* in Section IV.A.2, achieve that objective is outside the scope of this Note.

177 *See supra* Section IV.A.2.

178 *See supra* notes 85–92 and accompanying text.

The European Union's approach to pseudonymization is most akin to that proposed by Hartzog and Rubinstein—that the objective of data protection laws be risk management as opposed to perfect anonymization, and that sound data-release policy couples de-identification techniques with other technical, physical, and procedural safeguards.[179] And while Hartzog and Rubinstein recognized that transparency and disclosure *alone* are inadequate to protect consumers,[180] the European Union too recognizes this by strongly encouraging pseudonymization in conjunction with those safeguards.[181]

V. CONCLUSION

Based on the foregoing analysis, this Note proposes that the United States conceive of a second tier of de-identification, akin to pseudonymization, that falls within the scope of data protection laws while receiving relaxed requirements designed to still incentivize de-identification. Concededly, even with the introduction of pseudonymization, U.S. data protection laws would remain subject to the criticism that they continue to exempt data that has been anonymized in compliance with their requirements, and that there is no such thing as truly anonymous data. Perhaps the proper solution is more radical than that embraced by the European Union—perhaps privacy regulations should never exempt de-identified data whatsoever. However, such an analysis falls outside the scope of this Note, and for present purposes, it should suffice to suggest that the United States consider and encourage a second, less demanding tier of de-identification that remains subject to certain statutory privacy requirements. In doing so, its regulatory framework would achieve a more impressive balance between data utility and the risk of privacy harm, while providing protections to data subjects faced with the modern-day reality of re-identification.

179 *See supra* note 94 and accompanying text.

180 *See supra* note 95 and accompanying text.

181 *See supra* notes 162–172 and accompanying text.